An Evaluation of a Pedagogical Reform Designed for College Chemistry Teaching with Large Classes

by

Scott Edwin Lewis

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:
March 22, 2006

Keywords: guided inquiry, cooperative learning, hierarchical linear modeling, equity, higher education

© Copyright 2006 , Scott Edwin Lewis

www.manaraa.com

UMI Number: 3240395

UMI

UMI Microform 3240395

Copyright 2007 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

www.manaraa.com

Dedication

This work is dedicated to long list of dedicated and inspiring teachers who have placed their time and efforts for others, including myself, to learn and grow. At the very beginning of this list are my first and most influential teachers, my parents and brother. It is difficult, perhaps impossible, to give a complete description of the impact they have had on my life, my education and even the work presented here, suffice to say it has been tremendous and positive.

From a more formal environment, I consider the efforts from the many school teachers and mentors, beginning with elementary school, middle school and high school, and consider how lucky I have been to encounter such a group of individuals. In particular the teachers at King High School (1993 – 1997) who pushed me to discover who I was and what I may be capable of. Many instructors and professors at the University of South Florida played a similar role, providing ample opportunity in dealing with the abstract, honing analytical reasoning and the creativity to develop and consider alternatives. This includes the many and varied professors and colleagues I have enjoyed working with while in graduate school at the University.

For all teachers, I have learned a great deal from you, and consider the role you play instrumental in my life and the lives of many others. I hope this work, in some way, pays tribute to your efforts.

www.manaraa.com

Acknowledgements

It is astounding to reflect on the tremendous amount of support present in every stage of the research presented here. First, I thank Jennifer Lewis, my major professor, for having both the vision to begin a chemical education program and the perseverance and patience for me to develop the needed perceptions and skills to address the problems faced. I am also in gratitude to my committee, Dana Zeidler, Robert Potter and Abdul Malik for their welcome advice and their efforts in preparing this manuscript. Similarly I acknowledge the support and recommendations from the members of the Chemistry Education Research Group, past and present. Rosa Walsh provided exemplary access and resources to conduct this work, without her support the data could not have been collected. The Chemistry Department office staff, including Roberto Avergonzado, Adrienne McCain and Venice Osteen assisted in preparing many of the materials used here, and their contribution was essential and appreciated. I take particular note of the instrumental role the chemistry department instructors, peer-leaders, test proctors and students played, and thank them for their sincere efforts. The analysis of the data was guided by a long series of interactions with the students and faculty in the Measurement and Evaluation Department, the service they perform in this regard is phenomenal. In particular, I acknowledge Jeffrey Kromrey, John Ferron, Robert Dedrick, Melinda Hess and Bethany Bell-Ellison. Finally, I would like to acknowledge the steadfast support from many individuals, including my family, Michael Kamprath, Leila Amiri, Donovan Cantu, Cara Breier, Christina Fennell and Karen Taylor.

www.manaraa.com

Table of Contents

www.manaraa.com

www.manaraa.com

www.manaraa.com

List of Tables

www.manaraa.com

www.manaraa.com

www.manaraa.com

List of Figures

www.manaraa.com

An Evaluation of a Pedagogical Reform Designed for College Chemistry Teaching with Large Classes

Scott Edwin Lewis

ABSTRACT

This work presents an evaluation of a reform teaching practice, known as peer-led guided inquiry, that combines guided inquiry and cooperative learning for college chemistry teaching. Integral to implementing the reform in a large class (greater than 100 students) was the role of peer leaders, undergraduate students who have successfully completed the target course. These peer leaders facilitated cooperative learning groups during weekly guided inquiry activities in general chemistry.

The evaluation, using data collected over a 3-year period, had two main foci: effective teaching and promotion of equity in the classroom. Both of these aims were evaluated using hierarchical linear models. The reform was found to be effective, with a progressive increase in the test scores of those students in the reform classes versus the students in the traditional classes. Furthermore, students in the reform outperformed their counterparts on an externally-constructed national exam. Both findings also held true when controlling for student SAT scores.

www.manaraa.com

Effectiveness is not sufficient cause for recommendation amid concerns that distinct groups of students may be disadvantaged by a reform. The evaluation therefore had special concern for students who were at significant risk of low performance in a college chemistry course, such as those with poor high school preparation. No evidence was found that the reform made the situation worse for these students; in fact, the reform was determined to be effective regardless of preparation as measured by SAT scores. In addition, formal thought ability was found to be an important factor in chemistry performance, independent of SAT scores, with low formal thought ability placing students at-risk. The evaluation data indicated that the reform may have allowed students who entered the course with low formal thought ability to overcome this disadvantage, though this effect could be attributed to chance.

Finally, to understand further the students in this setting beyond cognitive factors, an inventory of student study approaches was administered. Three specific approach profiles were prevalent: surface, surface achieving and achieving. Two less prevalent approach profiles, deep and deep achieving, were related to better understanding of chemistry as measured by the national exam.

www.manaraa.com

I. Reform Teaching for Large Classes

This dissertation is an evaluation of a pedagogical reform that was performed over the course of three years within a college general chemistry curriculum. This work is done with a belief that such evaluations are necessary to promote a better understanding of the learning processes students employ at this level, and to promote the dissemination of reform techniques to other institutions and instructors. In short, to hopefully answer their deserving question, 'How do we know it works?' In this chapter and Chapter 2 the reform and data collections procedures are discussed in detail. The following chapters will present the analysis of the data in relation to specific research questions and the results and interpretations that follow from this analysis.

The reform combines two existing reform pedagogy models to create a student centered learning environment with large classes of students (approximately 200 students per class). The context for the reform is the first-semester general chemistry course at a large public university in the southeast United States. Traditionally this course meets for three one-hour lectures a week. To maintain the same amount of student contact time, the reform replaces one of the lecture time-slots, so that students in the reform section meet twice a week for one hour each in a lecture setting and once a week, for one hour, in the reform setting.

www.manaraa.com

Reform Development and Precedents

The first existing reform pedagogy model is guided inquiry. Inquiry as a pedagogical approach was described as early as the 1960's.[1] On a basic level, inquiry employs three phases of a learning cycle in a specific order: an exploration phase, in which students perform an unstructured investigation; an invention phase, in which an integrating concept is introduced to the learner; and an application phase, in which the same concept is applied to a variety of situations.[2] Modern interpretations of inquiry adopt a "guided" approach in which curricular materials pose specific questions to move students through these phases.[3] There is a substantial body of research providing evidence that inquiry is an effective tool for learning science,[4-6] and this has led to the promotion of inquiry within national science education policy statements, most notably the National Science Education Standards.[7] Inquiry materials for college chemistry have been developed by the Process Oriented Guided Inquiry Learning (POGIL) consortium.[8]

In the traditional lecturer model of the instructor, the instructor's primary tasks are to provide answers, explain concepts and serve as a source for confirming answers. This follows a transmission model of learning [9] where students receive the knowledge directly from the instructor, regardless of previous knowledge and experience. This is in contrast with an inquiry setting, as described by Spencer[10]. In an inquiry setting the course instructor is thought to act more as a consultant for students than as a lecturer. For example, an instructor may refer students to a certain piece of data while working on a problem, or encourage students to explain their concepts to other students. Instructors

www.manaraa.com

may also pose questions to help students formulate their own concepts. The inquiry approach is designed to work within the constructivist model of learning, which states that students construct knowledge by combining new information with their existing knowledge. Based on the instructor responsibilities in an inquiry setting, and the amount of student-instructor interaction required, it is nearly impossible for an instructor to implement inquiry in a large classroom setting. This hindrance serves as a cause for incorporating another existing reform pedagogy.

The second reform pedagogy used is the Peer-Led Team Learning (PLTL) approach, where peer leaders enter the classroom to lead small groups of students. A peer leader is an undergraduate student who has successfully completed the course in a previous semester. Traditionally in the PLTL approach peer leaders lead a workshop where a team of six to eight students work on materials that correspond to the course and to the assessments given in the course. Each peer leaders is responsible for facilitating the group work and is "trained to avoid being an information provider." [11] Additionally, peer leaders provide the norms for group work, promote communication among students, resolve conflicts between students and provide motivation to the students.[12]

PLTL is thought to promote student understanding by working under Vygotsky's model of cognitive development. Based on interviews with students working on problems, and the usefulness of prompts provided during the interviews, Vygotsky theorized what is termed the Zone of Proximal Development.[13] This zone describes a beginning point, which is a student's current understanding, and a final point, which describes the student's potential understanding, or what the student could achieve with

www.manaraa.com

minimal instruction. As part of this theory, instruction should be geared at the student's potential understanding, and not beyond that. Returning to the PLTL model, peer leaders and other students are thought to better provide instruction at the level of a student's potential, more than a course instructor who may be too far removed from the students' levels of understanding.

PLTL may also be beneficial to students simply as a derivative of cooperative learning. Cooperative learning can be described as occurring any time a group of students work together toward a common purpose.[14] Empirically the benefits of cooperative learning in improving performance on traditional academic measures have been well documented in a large variety of educational scenarios.[14-16] The theoretical justification for the effects of cooperative learning is still debatable, though, as a 1996 review by Slavin points out.[17] Slavin's work presentes two main theoretical perspectives. The first is motivational, indicating that successful cooperative learning alters the reward structure of the learning experience to provide additional motivation toward obtaining the learning outcome.[18] In the post-secondary environment, though, there are some indications that cooperative learning may actually reduce motivation among some students.[19, 20] The other theoretical perspectives can be classified as cognitive and includes the aforementioned Vygotskian approach where students are working within each other's Zone of Proximal Development. An additional approach that fits the cognitive perspective is that of cognitive elaboration, where the act of explaining material is thought to provide a meaningful reorganization of the knowledge for the explainer, resulting in the learning benefits.[21-23]

www.manaraa.com

Another cognitive perspective that may explain the benefits of cooperative learning is that of social constructivism, where students' knowledge is building upon their own existing knowledge, but facilitated by negotiating concepts with others.[24] Note that the claims based on the cognitive perspective are not mutually exclusive. For example, with Vygotsky's model and social constructivism the former may describe the latter as a useful means for working within a student's Zone of Proximal Development.

Returning to the intent to perform guided inquiry with large class sizes, peer leaders offer a means to do so. By working with small groups of students, peer leaders can facilitate student work on guided inquiry materials. This combination of peer-led guided inquiry (PLGI) is the basis for the reform that is to be studied. In this reform students work in assigned groups of four on assigned POGIL guided inquiry materials. The topic of the POGIL activity is selected to precede the material presented in the traditional portion of the course.

An example of an inquiry-based activity taken from the POGIL materials that is used in the reform investigates the factors affecting atomic radii.[25] The activity begins with an exploration phase where students receive a table of data describing valence shell, core charge and atomic radii for various elements, where each of these terms have been previously defined. Students are prompted to look for patterns among the information presented, leading to the concept invention phase, which is the relationship between core charge and atomic radii, and between valence shell and atomic radii. The culmination of the concept invention phase is when students are required to describe the reason for the relationship between each of the variables. In the application phase students are asked to make predictions of the size of an atomic radius relative to the atoms previously

www.manaraa.com

discussed. A similar process is undertaken for ionic radii, and that would comprise the activities performed in one PLGI session.

Group assignment is performed according to literature recommendations. The primary concern is in formulating groups that are heterogenous in terms of academic ability.[26-29] This was done by student SAT sub-scores. The math and verbal SAT sub-scores were combined in a way that best relates to chemistry performance, as determined by multiple regression on data from past semesters. The combined score for the set of students is then placed in a quartile, where the top quartile represents the students with the highest SAT scores, the second and third quartile represents students who fall in the middle, and the bottom quartile incorporates the students with the lowest SAT scores. The groups are then formed by random selection from one member of each quartile, making roughly 48 groups of four.

The groups are then reviewed for demographic characteristics. Research recommendations indicate that isolating students based on sex, where a person in the group is a sole member of a sex, may put that member at a disadvantage.[30] With similar concerns, groups were also scrutinized for indications where students are isolated based on race. Thus when the random selection produces a group that features these types of isolation, students are switched with another group to eliminate the isolation. Students are only switched within a quartile, that is a student from the second quartile can only be switched the second quartile student from another group, so as to ensure the academic heterogeneity of the group. The switching results in groups that are all female, two female and two male, or all male with regard to sex. A similar distribution happens with student race, regarding minority or non-minority. While it was not always possible

www.manaraa.com

to eliminate all forms of isolation on these constructs, it was often the case where at most one group of the 48 groups featured this type of isolation. Three or four groups (12 to 16 students) are then assigned to a peer leader.

Because of the hybrid nature of the reform, the peer leader is responsible for promoting guided inquiry practices in much the same way an instructor would, by encouraging students to consider data, explain concepts and perform self-evaluations of their own answers. Additionally, the peer leaders are responsible for ensuring the group work is productive, including equal participation among group members. Because of the essential role played by the peer leaders the training practices for peer leaders are an essential part of the reform and will be discussed shortly. In return for their service, peer leaders receive a small stipend.

The course instructor in the reform generally follows the PLTL guidelines for course instructor.[11] Primarily this involves planning the introduction of course material to follow the guided inquiry activities, and making specific mention to the activities done in the PLTL sessions during the course lecture. Thus instructor familiarity with the guided inquiry materials is an essential piece. Additionally the instructor or a learning specialist is responsible for the oversight of the peer leaders in both the peer leader training session and during the PLTL session. In this reform, the oversight of peer leaders was performed by a learning specialist, who is a faculty member with research interest in chemical education. This decision was made primarily to provide consistency in peer leader development and oversight from year to year.

www.manaraa.com

Details of the Peer-Led Guided Inquiry Reform

The reform setting is titled Peer-Led Guided Inquiry after the reform's antecedents. Students begin each PLGI session by turning in their assigned homework. The homework usually involves beginning the exploration phase in the inquiry activity, and is put in place to provide more time during the sessions and to promote students considering the concepts prior to the session. While the peer leader reviews the turned-in homework, students take a short multiple-choice quiz on the previous week's activity. Upon completing the quiz, the student receives the homework back. When all students are done with the quiz, they join the assigned groups of four each, which remain relatively constant through the semester, with adaptations made in cases where students drop the course.

Once in the group, the students receive a group folder from the peer leader that contains assignments for the roles to each student. The roles are taken from the POGIL project and are meant to promote equal participation among the students in the group, by assigning responsibilities to each student. First, the manager is responsible for keeping the group on task and checking for adequate progress on the activity. Additionally the manager represents the group to the peer leader. Second the recorder is responsible for keeping an official record of the group consensus with each problem. The peer leader may call upon the group recorder at any time to provide the group's responses, however to ensure all students understand the peer leader will at times call upon any student in the group to explain the recorder's written answers. The third role is the reflector, who is responsible for assessing the group dynamic and reports on strengths and areas to

www.manaraa.com

improve. At times throughout the session the peer leader will call for a report from the reflector to be given to the other students within the group. The final role is the presenter, who is asked from time to time to explain the group's answers on a specific set of problems, either to all groups on a board in front of the class, or to a specific group. Each week the peer leader changes the role assignments so that each student will have experience with every role. In addition to promoting equal participation among students in the group, the roles also promote communicative skills, self-reflection, and self-regulation, which the roles specifically request of the students.[29]

After the roles are assigned students are asked to compare homework answers and discuss discrepancies or different approaches to the assigned questions. This task provides students an opportunity to negotiate understanding of the material presented in the exploration phase, and to arrive at a consensus prior to proceeding with the inquiry activity. After students reach a consensus on the answers and the underlying reasons for the answers, students are then assigned a series of problems that complete the exploration phase and begin the concept invention phase. Frequently during this time, group members debate the nature of the concepts presented. As part of the debate, the peer leader encourages students to supplement their understanding with the material previously covered. Also during this time, students or groups are often called to present to the rest of the class (where a class now comprises 12 to 16 students) their understanding and why they believe so. The concept invention phase is generally complete when all students in a group appear to reach a satisfactory understanding of the concept. The peer leader is advised to question students to evaluate their understanding

www.manaraa.com

of the new concept. If time remains, application style questions are then assigned, which requires students to apply the concept with a different context than the exploration phase.

Approximately five minutes at the end of each PLGI session is specifically set-aside to promote students self-assessment skills. Students are asked to record any information they weren't clear on from the activity and list areas where the group may improve. Some additional questions are particular to the point in course, for examples, asking students to list effective ways to study for exams or to approach homework assignments.

Peer Leader Training

Peer leader training occurs primarily in a separate course established specifically for peer leader preparation. The course meets for one two hour session once a week, prior to the PLGI session. Attendance to the training course is a mandatory requirement for all peer leaders, if peer leaders cannot attend a training session then they may not meet with students during the PLGI session. The training session is generally split between two hours. The first hour is spent with peer leaders working on the upcoming guided inquiry assignment in a matter very similar to how the students will experience it. In other words, peer leaders work in groups of four on the assignment as the instructor, who leads the training session, models the role of the peer leader in a guided inquiry setting, encouraging the students to consider the data to answer questions they have and to consult with others within their group. As the semester progresses, peer leaders are called upon to lead the training course, with the instructor becoming an in-class observer and providing feedback. The second hour of the training session is usually split between

www.manaraa.com

administrative issues, discussion among peer leaders of challenges, and sharing successful techniques. An additional training technique involves peer leaders keeping a weekly journal that describes the PLGI session and also incorporates the peer leaders' assessments of the strong points and areas of improvement during the PLGI session. The keeping of a journal is meant to promote reflective practice, a key aspect in promoting reform teaching, such as inquiry, among beginning instructors.[31] The weekly journals are sent to the training session instructor, who provides regular feedback concerning the reflective nature of the journal.

Finally, as a means for quality assurance, PLGI sessions were chosen at random for observation by members from a group of graduate students and a faculty researcher interested in chemical education. Typically in a given week multiple peer leaders were observed by different members. The observations were geared toward providing the peer leader direct feedback on the session. Of particular importance were occasions where the peer leader short-circuited the guided inquiry intent of the sessions and became an information provider in the classroom. For example, if a peer leader provides answers when students struggle with a question, then the inquiry activity becomes more like a recitation exercise. Unfortunately this bypassing is relatively easy to do, in particular as traditional teaching practices, whom the peer leaders and students likely have a large amount of experience with, promote the model of the instructor as an answer provider. [32] After the in-class observations of the PLGI sessions, the observer/researcher meets with the peer leader to discuss the strengths and areas of improvement for the session. As mentioned any process seen that may short-circuit the guided inquiry nature of the activities is one area of improvement that is focused on, and this typically results in a

www.manaraa.com

brain-storming session with the peer leader on how to deal with a similar situation in the future. In the course of a typical semester, a peer leader is usually observed two to three times.

Directions for this Work

A principle objective of this work is to evaluate the outcomes of this reform in terms of student understanding. In order to do this, data was collected from the fall and spring semesters over three academic years. Among the immediate patterns is that there is a quantifiable difference between fall and spring semester cohorts (see Chapter 2). The reform, however, was only offered in one class during each fall semester. Because the differences between fall and spring semesters introduce a potential confounding variable, the reform classes were only compared to other classes that took place in the fall semesters. Otherwise any differences found may be attributable to the comparing the reform classes that took place in the fall, to comparison classes which took place in the spring semester. As part of the evaluation, there is a reliance on a series of acronyms, the most common of which for convenience have been collected into a table that is in Appendix A.

Chapter 3 begins this evaluation, looking for indications that the reform is an effective and equitable teaching environment, with the traditional, lecture-based fall semester classes serving as the comparison group. Equity in this case is measured by the effect of the reform on students of differing high school preparation. Chapter 4 examines the role of another measure of students' incoming skills, Piaget's construct of formal thought. Since the role of formal thought in the general chemistry course has not been

www.manaraa.com

satisfactorily established, Chapter 4 seeks to establish this role. As a result, the study presented in Chapter 4 focuses on all classes, fall and spring semesters of general chemistry, which did not take part in the reform. Once this is complete, the effect of the reform on students of varying formal thought is incorporated in Chapter 5. Like Chapter 3, Chapter 5 again looks only at students in the fall semesters.

Chapter 6 seeks to establish the role of students' study processes, in the context of the general chemistry class, and how it relates to students' chemistry understanding. This chapter uses data available from both the fall and spring semesters in the last year of the study (Fall 2004 to Spring 2005), and does not include the one reform class (taking place in Fall 2004) as it is perceived as a change in context. This chapter closes with a preliminary investigation into the effect the reform may have on study processes used in the chemistry classroom.

Students participated in the reform by self-enrolling in the class that was assigned to the reform. Prior to enrollment students had no way of knowing which section would be the reform; however, students were made aware of the reform during the first week of class and were provided an opportunity to switch out of the reform without penalty during the first week. No unusual drop rate for the reform class was witnessed. Students were provided the inquiry materials for the reform free of charge: a curriculum development grant through the National Science Foundation (DUE-0310954) covered the associated costs.

Through the course of the three years of the study, 491 students participated in the PLGI reform, 2558 students enrolled in general chemistry in the fall semester classes that did not use the PLGI reform and 1808 students enrolled in general chemistry in the spring

www.manaraa.com

semesters. Of those, 344 students (70.1%) in the PLGI reform completed the course (took the mandatory final exam), 1807 students (70.6%) in the fall semesters without the reform completed the course and 1261 students (69.7%) in the spring semesters completed the course, or approximately 70% of each set of students completed the course.

Preliminary analyses were conducted on the first semester of data collected (Fall 2002), comparing one reform class to a comparison class taught by the same instructor. The results indicate that the reform class out-performed the comparison class on all measures taken, and that the improvement increased as the semester progressed. The measures used in this work will be discussed in Chapter 2. Additionally, a significant, positive correlation was found between attendance to the reform sessions and performance on each of the measures. This positive correlation remained when controlling for student SAT scores, reducing the likelihood that the relationship between attendance and performance was spurious and attributable to student background. Further reading on the results of this preliminary analysis can be found in the *Journal of Chemical Education*[33], while the promising results from this analysis served as a basis for continuing the reform in following semesters.

www.manaraa.com

II. Instruments and Methods

This study relied principally on surveys and tests, traditional methods of quantitative measurement. A description of each instrument and when it was employed in the course will be discussed in this section. The theory underlying each instrument, and the psychometrics of each instrument, will be described in later sections when each instrument is introduced in the context of a research question. The majority of the data collected in this study was determined exempt from informed consent procedures as determined by the university's Institutional Review Board (see Appendix B), with one exception which is noted in Chapter 5.

Immediately prior to each semester, a request is made of the university registrar office for SAT scores, sex, and race, for all students enrolled in the course. During the first class meeting, students were given a demographic survey that ranged from 9 to 16 items. A copy of the survey is in Appendix C. The survey was multiple choice, and to be filled out on scan-trons. Questions on the survey were focused toward students' intended major, college and high school chemistry and math background as well as sex and race. The survey results agreed with university records over 99.3% of the time on student sex. For race the survey allowed multiple classifications (e.g. a student could choose white-Hispanic) while on the university records they could not. The university records agreed with at least one of the classifications that students chose on the survey at a rate of

www.manaraa.com

96.4%. Because of the multiple classifications in the survey, university records will be used for student sex and ethnicity when discussed.

Comparisons between Fall and Spring Semesters

The survey and university records allow a more in-depth description of those students enrolled in General Chemistry I, as well as an indication of the differences between students taking the course in the fall versus spring semesters. Survey results are available for 2634 students from the fall semester and 1689 students in the spring semester. In the fall semester 49.4% of the students were in their first year in college, compared to 61.8% in the spring semester. The fall semester has a 40.7 to 59.3 male to female ratio, while in the spring the ratio is 34.4 to 65.6. The ethnic breakdown for each semesters has 61.4% white, the fall semester has a higher percentage of students reported as Asian by the registrar's records, 10.8% to 8.8%. The spring semesters have a slightly higher percentage of minorities that are traditionally underrepresented in the sciences; in the spring semester 14.6% of students are reported as Black compared to 12.6% in the fall. Hispanic students comprise 12.0% of the spring semester population, and 11.4% of the fall semester population.

Students were also asked to identify their majors or intended majors out of five choices. In the fall semesters 55.5% identified themselves as pre-med or allied health profession majors and 14.0% as engineering majors. This is the primary contrast to the spring semester where 60.7% chose the pre-med or allied health professions option and 9.6% chose engineering. The other categories were relatively similar, with other science majors (20.3% fall, 20.1% spring), non-science majors (6.5% fall, 6.9% spring) and

www.manaraa.com

chemistry (3.7% fall, 2.7% spring) between the fall and spring semesters. Among the fall semester students, 80.7% intend to take the follow-on course General Chemistry II, in the spring semester this value was 81.3%.

Asking about high school chemistry background seems to indicate that students in the fall semester have taken more high school chemistry classes. In the fall semester, 16.5% have had more than one year of high school chemistry, 59.5% said exactly one year, 18.2% chose one semester and 5.9% have not taken any chemistry in high school. In the spring semester 12.9% had more than one year, 59.0% had exactly one year, 20.8% had one semester and 7.3% had not taken chemistry in high school. High school chemistry background was one of the considerations in advising students to take Chemistry for Today prior to enrolling in General Chemistry I, and this represents one of the largest distinctions between fall and spring semesters. In the fall semester roughly one-quarter of the students had taken Chemistry for Today (24.3%), while in the spring semester roughly one-half of the students enrolled had taken Chemistry for Today (49.9%).

Based on the survey results and university records, a class in the fall semester has more male students, more engineering students, students with more high school chemistry experience and college experience (as determined by years in college) and fewer students with Chemistry for Today experience. The difference between fall and spring semesters is also demonstrated by pre-semester measures, with students in the fall semesters averaging 566.1 on the Math SAT versus 529.6 in the spring semesters. Verbal SAT has a similar trend, with 545.0 in the fall semesters and 519.7 in the spring semesters.

www.manaraa.com

Instruments Used In-class during the Study

During the second week of class, the Test of Logical Thinking (TOLT) was administered in the students' normal exam setting. The normal exam setting is a set fifty minute time-block, where all students in General Chemistry meet regardless of their normal class time. Enrollment in the course is conditional on having this time-block free for tests. During the normal test setting, students are sent to one of eight to twelve lecture-halls, to take the test. This allows all students in the course to take a test at exactly the same time, and prevents students from an earlier class telling students in other classes the content or nature of the test. Each lecture hall has one to three proctors to administer the exam and prevent cheating. The TOLT is the first test administered in this setting. The TOLT has 20 items, and students are told they will be given credit as long as their performance demonstrates meaningful effort. The TOLT will be further discussed in Chapter 4.

Approximately four to five weeks into the semester students are given their first content-based exam, in the normal exam setting. During the semester, students take a total of four content-based exams, also termed midterm exams. Each midterm exam has between 20 and 25 questions. The exams are created by the instructors who currently teach the course, with each instructor submitting four to five questions, and then the entire test is reviewed by all instructors for appropriate length and content. The midterm exams are given in the normal exam setting and students submit their answers via scantrons. Students may dispute the grading of the exam or the appropriateness of a question to the course instructors, and any modifications the instructors choose to make

www.manaraa.com

are carried out on the exam scores of all students. Students are graded on the number of correct responses and approximately 50% of the final grade in the course is determined by performance on the midterm exams. The lowest midterm exam score is not considered when calculating the contribution to the final grade, and for this reason no make-up exams are permitted if a student misses an exam.

At the end of the semester students take a final exam, which is an exam that is constructed by the American Chemical Society (ACS) Exams Institute. The content of the exam is secure and confidential in accordance with ACS Exams Institute requirements. Each semester instructors write ten additional questions which are given after the ACS Exam, meant to represent the most important aspects of the semester. Because the ACS Exam is given independently of the ten additional questions, and by itself serves as an external measure of students' understanding, only the questions on the ACS Exam are used for the end of semester measure. This exam is given in a setting similar to the normal exam setting, however there is a two-hour time block for the final exam. The final exam score accounts for approximately 25% of a student's final grade, and completing the final exam is required to successfully complete the course.

Also used in this work is the Study Processes Questionnaire (SPQ), which is a 42 item survey that was administered in-class. The SPQ is discussed in further detail in Chapter 6. Students had 15 minutes to complete the questionnaire at the beginning of class and were given credit for class attendance for taking the survey. The attendance credit was a very small portion, less than 2%, of the final grade. For students who were not in attendance, or arrived late, the day of the survey, a make-up survey was administered outside of class on a later date, where students could receive the attendance

www.manaraa.com

credit they missed.

All of the instruments aforementioned were completed on scan-trons, and scanned by the University of South Florida scanning office. The resulting files were read into Excel and screened for unusual or missing information. For example if a student filled out a scan-tron using ink instead of pencil lead, the responses were hand-entered into the Excel sheet. After this, the Excel file was converted into an SPSS .sav file, and the information was sorted based on Student ID. Student ID is a unique identifier within the population, and because of this, it serves as an ideal variable for merging multiple data sources. Thus as each semester continued and more information was collected on the set of students, it was continually merged back into one growing base file. For each semester one base file was made for the cohort of students, and during the course of the study six base files were made. At the end of the study, the six base files were merged into one matrix that contained all information available for students in General Chemistry I over the past three years. The file that incorporated all the data was read into both SAS and SPSS. Depending on the analysis at-hand, missing data and outliers were considered for their impact on generalizability, and are discussed in each chapter as they pertain along with the validity and reliability of the instruments used.

www.manaraa.com

III. PLGI Effectiveness and Equity

Based upon data collected from three semesters in three subsequent years, this investigation focuses on two major considerations: efficacy and equity. An effective reform demonstrates improved student achievement, while an equitable reform reduces differences in achievement among groups (e.g. among students of different sex, race/ethnicity, socioeconomic status, or high school preparation). To address these considerations the study employs hierarchical linear models considering students as nested within a class. This approach allows for evaluating the reform as a classroom-level implementation, unlike previous quantitative analyses that model reform pedagogies as a student-level implementation. This chapter begins by describing why even well-studied pedagogical reforms can benefit from the use of hierarchical linear models. Next, the construction and interpretation of hierarchical linear models for both a single outcome measure and a time-series of outcome measures are presented. Finally, these models are used to evaluate a reform and discuss the implications of the results for science teaching and educational research practices.

Consensus on Pedagogical Reforms

Two pedagogical approaches provide the basis for the reform investigated in this study: cooperative learning and inquiry. When each approach is considered separately in

www.manaraa.com

terms of demonstrated effectiveness and national policy statements, it is clear why there has been an impetus in college science teaching to combine the two approaches.[8]

To begin with cooperative learning, it has been extensively studied [14] and has been called "one of the greatest success stories in the history of educational research."[17] Very simply, cooperative learning occurs when students work together in groups to achieve a shared goal.[34] While the focus of research has been on K-12 settings, there has been considerable work examining collegiate settings.[15, 16] This research has shown that cooperative learning is effective (i.e. demonstrates improved student achievement for all students) in a variety of settings. It has also been postulated that cooperative learning may help to promote equity by reducing existing gaps in achievement, [12] but equity has yet to be investigated in a college science setting. Currently, numerous policy statements calling for curricular reforms in post-secondary science education promote cooperative learning as a successful pedagogical strategy, worthy of adoption.[35, 36]

On a basic level, inquiry employs three phases of a learning cycle in a specific order: an exploration phase, in which students perform an unstructured investigation; an invention phase, in which an integrating concept is introduced to the learner; and an application phase, in which the same concept is applied to a variety of situations.[1, 10] Modern interpretations of inquiry adopt a "guided" approach in which curricular materials pose specific questions to move students through these phases.[3] While an investigation of equity has provided mixed results, [37] the substantial body of research providing evidence that inquiry is an effective tool for learning science [5, 6, 38] has led

www.manaraa.com

to the promotion of inquiry within national science education policy statements, most notably the National Science Education Standards.[7]

With such strong support from national policy documents for each approach, a combined pedagogical strategy for college science in which inquiry activities are the shared goal of cooperative learning groups can reasonably be expected to be effective, and possibly to be equitable. However, this support can also divert would-be adopters from looking closely at the research underlying the recommendations.

The Case for Hierarchical Linear Models

While inquiry and cooperative learning have indeed been separately evaluated for efficacy in the college science setting, there are a few caveats. Typically these evaluations have used a quasi-experimental design, comparing a class with reform to another class without reform on a performance-oriented outcome measure (e.g. exams or course grades). Efficacy determinations are based on statistical treatments such as ANOVA, with each student serving as an individual data point.[7, 14-16, 39-42] By design, most statistical tests comparing average scores rely on an independence of observations assumption, which mandates that one observation cannot affect other observations. Violations of this assumption (i.e. dependent observations) lead to error values greater than those assumed for significance testing. The net result of a violation is a statistical test that is more liberal than the set significance level, sometimes severely so, which leads to misinterpretation.[43] This intrinsic problem renders procedures such as ANOVA and regression less than robust to violations of the independence assumption. To satisfy the independence assumption so that these procedures may be used, it is

www.manaraa.com

necessary to provide a theoretical justification that the units of observation are independent of each other. However, students within a classroom have multiple opportunities to impact the learning of other students.[44] For example, in a traditional lecture setting, if one student in a class were to ask a question of the instructor, the entire class would have the opportunity to hear the instructor's explanation and thus share in a common learning experience. The problem is compounded when a group-based reform such as cooperative learning is introduced. Inherent to group learning is the idea of interdependence: students within a group can and should affect each other's learning. Progress in individual student learning is therefore expected to relate to the progress of other students within a group. With a typical cooperative learning implementation, students also have frequent opportunities to hear presentations from other groups or instructor responses to questions from other groups. In this situation, it is very hard to argue theoretically that the independence of observations assumption is satisfied.

A common recommendation to alleviate a lack of independence of observations at the individual student level is to change the unit of analysis. It may be more feasible to provide theoretical support for the belief that classrooms are independent of each other. If so, one could treat classrooms rather than students as individual cases, but sample size considerations would require resources sufficient to support reform in a large number of classrooms. In addition, the unit of analysis change implies a change in the level of analysis as well, so that implications stemming from research findings would now properly apply only to classrooms, not to students.[45] When possible, statistical methods of pooling data from separate research studies, also known as meta-analysis, can also allow for the selection of a unit of analysis that would satisfy the independence of

www.manaraa.com

observations criterion. A major drawback of meta-analysis is that applicable research on the reform of interest must already be in existence and provide substantial detail. A third approach, adroitly requiring neither additional resources nor an existing body of information-rich literature, is hierarchical linear modeling (HLM). HLM incorporates related linear models for each desired level of analysis. The error arising from dependent observations can be directly addressed by specifically defining an error term associated with that dependence. This ability to incorporate data for multiple levels of analysis makes HLM ideal for evaluating pedagogical reforms. For comparisons between intact classes, at least two levels are desirable: the individual student level and the whole class level. The variability resulting from student interdependence is part of the class level model, while the student level model contains only the independent student variability. Additionally, because it is able to model dependent data sources, HLM is particularly well suited for studies based on cases containing multiple data points, such as looking for indications of an effect over time.[46]

Addressing Efficacy and Equity

Despite its many advantages, HLM is relatively new and has not yet been used extensively for research in science education. A search of the abstracts of *The Journal of Research in Science Teaching*, *Science Education* and *International Journal of Science Education* from January 1995 to May 2005 produced four articles that have employed HLM, [37, 46-49] but only one of these investigates pedagogical reform. Von Secker used HLM and the National Center for Educational Statistics' High School Effectiveness Study to look for evidence of efficacy and equity associated with inquiry in high school

www.manaraa.com

science.[37] The study treated students as nested within schools and investigated three key features of inquiry practices: a reduction in teacher-centered instruction, the promotion of critical thinking skills, and the incorporation of inquiry-based laboratory investigations. Results were mixed, with only the laboratories associated with both higher overall achievement and reduction in achievement gaps. One limitation of the study is the use of data from teacher questionnaires to represent a school-level characterization of inquiry. In support of this decision, the authors report that there were significant differences between schools on the relevant questionnaire responses, but not between teachers, and that there were too few teachers in the sample to add the classroom level to the model. However, the accuracy of using teacher self-reports to measure the implementation of desired reform activities in the classroom has been called into question.[50]

Von Secker's HLM work provides some empirical justification for the idea that the use of inquiry can, in some cases, be associated with both efficacy and equity. While the reasons for the efficacy of cooperative learning are still debatable, [17] it has long been held by constructivists that actively working with information is the only way students can construct their own knowledge, [24] and cooperative learning does provide opportunities for students to explore the status of their conceptions actively, in particular while crafting explanations for other students.[51] Although proponents of cooperative learning have claimed that cooperative learning also has the potential to promote a more equitable classroom, [14] Elizabeth Cohen has explicitly demonstrated that the use of cooperative learning does not automatically produce equity. She has written extensively on the behavior of middle school students during group work, using expectation states

www.manaraa.com

theory to guide her investigations.[52] She reports that perceived status characteristics play a large role in shaping student interactions, with low status students at a disadvantage. The most important student status marker was found to be academic ability, but sex and race/ethnicity also play a role.[53] Her work is an authoritative warning that cooperative learning, rather than promoting equity, is likely to perpetuate the status positions with which students enter a class, and she recommends careful attention to group composition and specific interventions designed to change perceptions of low status students as a potential remedy.[54] More recent work has explored the effectiveness of these techniques designed to modify status roles, with mixed results.[29, 55] Additionally, Noreen Webb's examination of student interactions and achievement in K-12 mathematics classes indicates that cooperative learning may preferentially benefit high ability students. Webb has found that high ability students engaged in cooperative learning give verbal explanations, a behavior that is linked to achievement gains, more frequently than low ability students.[23] In short, a consideration of Cohen's and Webb's work leads to the disturbing possibility that cooperative learning, rather than automatically producing an equitable classroom, can actually make matters worse.

Concerns about the effects of status roles on achievement in the context of group work resonate with the literature investigating diversity and achievement in science. Sex, race/ethnicity, socioeconomic status (SES), and high school preparation (often associated with "tracking" practices) have consistently been linked to undesirable differential achievement patterns. Causal explanations for patterns of inequity have been constructed in terms of separate factors. For student sex and race it has been postulated that sexism [35] and dysconscious racism [56] from both teachers and peers leads to the under-

www.manaraa.com

representation of women and minorities in the sciences. Students of low SES have been said to experience limited access to both individual and school resources.[57, 58] "Tracking" has been criticized as further preventing success in college by virtually ensuring poor high school preparation for significant numbers of students.[58] At the college level, these prior inequities can affect performance in entry-level science courses. These courses have been held responsible for diminishing diversity in science fields, as those students held back are disproportionately women, racial or ethnic minorities, and those with poor high school preparation.[59] In particular, general chemistry, a required entry-level science course for a variety of science majors, is a critical juncture for many students, with poor performance preventing the pursuit of science-oriented careers.[60, 61] Choosing an appropriate pedagogical reform for this course therefore requires attention to equity as well as to effectiveness.

Research Goals

Will a pedagogical reform incorporating both cooperative learning and inquiry in a large entry-level college chemistry lecture course be effective and equitable? As discussed, prior research on effectiveness has often glossed over potential problems associated with dependent observations, and prior research on equity has mixed results at best. In fact, there is definite risk of a rather spectacular failure to achieve equity when implementing cooperative learning and inquiry – actually amplifying existing achievement gaps.

In light of these concerns, the present study had several goals. The first was to demonstrate that HLM is a viable strategy for handling observations theoretically

www.manaraa.com

expected to be dependent and to describe the appropriate approach for students nested in classes. The second was to use HLM to determine whether a reform combining cooperative learning and inquiry produces achievement gains in General Chemistry. An analysis of data collected at a single institution over a three-year period from lecture sections both with and without reform was used to achieve this aim. The third goal was to use HLM to investigate the equity of the reform. Von Secker's mixed results and Cohen's work on status indicate that equity cannot be assumed, despite national policy documents recommending inquiry and cooperative learning to remedy a range of inequities.[62]

Implementation of Pedagogical Reform: Peer-Led Guided Inquiry

This chapter uses HLM to investigate a pedagogical reform undertaken in the first-semester general chemistry course at a large urban public research university in the southeastern United States. The study design, while comparable to Von Secker's national database HLM study, [37] maintains a greater degree of control over the reform variable by utilizing two distinct advantages: the reform can be explicitly described in terms of setting, curricular materials, and training, and it can be directly contrasted with non-reform classes occurring in the same setting.

The reform, called Peer-Led Guided Inquiry (PLGI), is a form of cooperative learning utilizing inquiry but suitable for large classes. At the institution in which the reform took place, first semester general chemistry is taught in classrooms capable of seating up to 190 students, and frequently the enrollment cap is met. Traditionally, these large classes meet three times a week for fifty-minute lectures, and, each semester, four to eight classes (or "sections") are taught concurrently. The reform is implemented in

www.manaraa.com

only one section, but all students enrolled in the course, regardless of section, are guided by a common syllabus and take common exams. Implementation is designed so that a PLGI session simply supplants one of the lectures in the reform section each week, keeping the amount of instructional time the same as for the traditional sections.

Peer leader training is also designed to utilize suggestions from Cohen's work to promote a more equitable classroom. Specifically, the interventions are geared toward "equalizing rates of student-student interaction."[53] Toward this end, peer leaders are reminded to have students adhere to assigned group roles, such as manager, recorder, reflector and presenter, which rotate weekly. Each role has a set of prescribed responsibilities to encourage the development of specific process skills and a high degree of student-student interaction among all students in a group, not just the high status students. To assist them with enforcing group roles, the peer leaders practice intervention strategies focused on group dynamics and process skills rather than on content. Finally, the peer leaders are coached to solicit feedback and explanations from all students, including those who appear to be struggling or disengaged. To promote these strategies, each peer leader was observed at least twice per semester, with feedback provided regarding student involvement. Additional information on PLGI peer leading training practices is presented elsewhere.[63] Since Peer-Led Guided Inquiry (PLGI) has been explicitly designed for working with large classes (greater than 100 students) in first-semester general chemistry, the reform is ideal for a university setting.

www.manaraa.com

Hierarchical Linear Model Construction

In this section, two different hierarchical linear models that will be used to investigate both class-level and student-level effects of the reform are presented. Both of the models are suitable for a quasi-experimental design in which data is available from classes both with and without implementation of the reform. The first model relies on student final exam scores as a single measure of chemistry achievement, while the second model uses student midterm exam scores to investigate change over time. Both models use SAT scores to characterize pre-existing achievement gaps. The features of each model that allow interpretations of results in terms of both efficacy and equity will be highlighted.

Final Exam Scores: The External Exam Model

To introduce HLM the notation of Raudenbush and Bryk is used.[64] For an education setting, consider a multiple regression model:

$$Y_{ij} = \beta_{0j} + \beta_{1j} MSAT + \beta_{2j} VSAT + r_{ij} \qquad (1)$$

where Y_{ij} is an outcome measure for student i in classroom j, MSAT is a student's math SAT score, VSAT is a student's verbal SAT score and r_{ij} is an error term to describe the unique effect of each student. This serves as the Level 1 model, or the model that examines student effects, in an HLM. It is identical to a multiple regression model, except for the subscript j, which indicates that the values for the coefficients will change

www.manaraa.com

depending on the classroom variables. Accordingly, the β coefficients in the Level 1 model emerge as outcome variables from the Level 2 model:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}SATavg + \gamma_{02}REFORM + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}SATavg + \gamma_{12}REFORM + u_{1j} \\ \beta_{2j} &= \gamma_{20} + \gamma_{21}SATavg + \gamma_{22}REFORM + u_{2j}\end{aligned} \qquad (2)$$

that describes the classroom effects on performance. In this Level 2 model two variables characterize the classroom: SATavg is a class's average SAT score, and REFORM is a dichotomous variable describing if a class experienced the reform (REFORM = 1) or did not (REFORM = 0). Note that, similarly to the variable r_{ij} in the Level 1 model, the variable u_{0j} in this Level 2 model describes the unique effect of class j on β_{0j}. Combined SAT scores (Math SAT plus Verbal SAT) were used as classroom level predictors to avoid interpretation difficulties associated with multicollinearity, since the correlation between the sub-scores (MSAT and VSAT) was high ($r > 0.8$) at the classroom level.[65] A high correlation between sub-scores is not present at the student level, and individual sub-scores have frequently been related to performance in chemistry, [66-68] so the sub-scores are used in the Level 1 model.

Consider a full HLM model incorporating both Level 1 and Level 2, with the outcome measure, Y_{ij}, performance on an American Chemical Society Exam given at the end of the semester. This model will be referred to as the External Exam Model. Statements regarding the efficacy and equity of a reform to which all students in a particular classroom were exposed arise from an examination of the impact of classroom

www.manaraa.com

effects (Level 2) on each of the three Level 1 coefficients: β_{0j}, the intercept; β_{1j}, the relation between math SAT and performance; and β_{2j}, the relation between verbal SAT and performance. If the reform has a significant positive impact (via γ_{02}) on β_{0j}, this would indicate a rise in overall performance, *i.e*. effectiveness. Another important consideration is whether the reform reduces the relation between math SAT and performance (β_{1j}): a significant and negative γ_{12} coefficient would indicate the reform's association with reduced dependence of performance on Math SAT sub-scores. This reduction in dependence of the outcome measure on incoming ability would be a sign of a more equitable classroom.[64] On the other hand, a significant and positive γ_{12} would indicate that the reform is increasing β_{1j} (the coefficient relating math SAT and performance) and actually enlarging the performance gap between students with high math SAT sub-scores and students with low math SAT sub-scores. An undesirable effect such as this would provide grounds to reconsider the reform implementation. Similar information can of course be obtained concerning the relation between the reform and verbal SAT sub-scores (β_{2j}) via γ_{22}.

Performance on Midterm Exams: The Time Series Model

Because HLM can model dependent data, it can also be used to examine multiple measures over time within an individual to look for changes in performance. The External Exam Model discussed above is able to demonstrate whether, at the end of a reform experience, students perform better or worse on the selected achievement measure relative to those who did not have the reform experience, but it cannot provide any information about whether performance differences changed over time. With regard to

www.manaraa.com

time, the ideal reform would have two key features: (1) greater effectiveness over time, with students who experience reform not only doing better than those who do not experience reform, but with the gap between the two groups growing the longer they experience the reform; and (2) greater equity over time, with existing achievement gaps among the students experiencing reform decreasing as the reform continues. The following model is designed to use midterm exam data to investigate whether these ideal features were realized within the context of a semester-long reform implementation.

All students in the study took four midterm exams at approximately equal intervals throughout the semester they were enrolled in the course. Each semester the instructors teaching the course created each midterm exam collaboratively, and the exams were administered to all students (in both comparison and reform sections) at the same time. An HLM model to describe change over time can be depicted in the following equations.

Level 1 – Within Student

$$P_{ijk} = \pi_{0\,jk} + \pi_{1\,jk} Time + e_{ijk} \tag{3}$$

Level 2 – Between Student

$$\begin{aligned} \pi_{0\,jk} &= \beta_{00k} + \beta_{01k} MSAT + \beta_{02k} VSAT + r_{0\,jk} \\ \pi_{1\,jk} &= \beta_{10k} + \beta_{11k} MSAT + \beta_{12k} VSAT + r_{1\,jk} \end{aligned} \tag{4}$$

Level 3 – Between Class

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + \gamma_{001} SATavg + \gamma_{002} REFORM + u_{00k} \\ \beta_{01k} &= \gamma_{010} + \gamma_{011} SATavg + \gamma_{012} REFORM + u_{01k} \\ \beta_{02k} &= \gamma_{020} + \gamma_{021} SATavg + \gamma_{022} REFORM + u_{02k} \\ \beta_{10k} &= \gamma_{100} + \gamma_{101} SATavg + \gamma_{102} REFORM + u_{10k} \\ \beta_{11k} &= \gamma_{110} + \gamma_{111} SATavg + \gamma_{112} REFORM + u_{11k} \\ \beta_{12k} &= \gamma_{120} + \gamma_{121} SATavg + \gamma_{122} REFORM + u_{12k} \end{aligned} \tag{5}$$

www.manaraa.com

Here P_{ijk} represents the percent correct on exam i by student j in classroom k. Time is an ordinal variable that spans from 0 (representing the first exam) to 3 (representing the fourth exam). The Level 1 equation depicts a linear relation between the time in the class and the performance on the series of exams. Level 2 models both the intercept and slope coefficients of Level 1 based on individual student characteristics, and Level 3 models all six coefficients in the Level 2 equations based on the two classroom variables, SAT average and presence or absence of reform. This three-level model will be referred to as the Time Series Model. Earlier research showed that the students in a reform section out-performed those in a comparison section with the difference between the two groups qualitatively increasing as the semester progressed.[33] The Time Series Model allows for the inclusion of additional students and has the potential to supply a better description of the evolving difference between reform and comparison sections: γ_{102} relates reform directly, controlling for class SAT and student SAT sub-scores, to the dependence of exam scores on time. If γ_{102} is positive and significant, the reform can be linked directly to improvements in exam performance over time as compared to the non-reform sections.

In terms of equity, positive and significant β_{01k} or β_{02k} coefficients would indicate achievement gaps present with the first exam. If achievement gaps are present, β_{11k} and β_{12k} describe how these achievement gaps change over time: if they are negative, the gap is shrinking as the semester progresses; if they are positive, the gap is growing. The effects of the reform enter similarly at Level 3: γ_{012} and γ_{022} describe the effect of the reform on equity for the first exam, with positive values indicating a wider gap under the reform. A significant and negative value for γ_{012} would mean that, under the reform, student test scores were less dependent on Math SAT. In a similar manner, γ_{112} and γ_{122}

www.manaraa.com

describe the effects of the reform on equity as the semester progresses: a significant and negative γ_{112} or γ_{122} would indicate that, under the reform, classrooms are becoming more equitable, and less dependent on student SAT sub-scores, as the semester progresses.

The two models provide separate appraisals, with robust significance testing, of effectiveness and equity that may result from the implementation of a reform. The first model looks at an end-of-the-semester measure, while the second model investigates the progress in the reform over time. For both models, the traditional lecture-based teaching method serves as the ground for comparison to allow for claims of effectiveness and/or equity. The combination of the two models creates greater understanding of the effects of the reform than either model alone could provide.

Research Methods

To consider whether cooperative learning as it was implemented reduces achievement gaps between students, it is necessary to identify a measure of prior achievement as well as an outcome measure. The relation between high school preparation as measured by SAT sub-scores and college chemistry performance has been well documented in a number of studies [66-69] indicating that this measure is an excellent choice for prior achievement. Both mid-term exams and an ACS exam, as described under instruments, serve as outcome measures.

Data from all 16 fall sections were considered in this analysis. This resulted in data for 2838 students enrolled in the course. SAT sub-scores were available for 2255 students (79.5%) with the most likely causes of missing data students having taken the ACT rather than the SAT or students enrolling in the course after the records were pulled.

www.manaraa.com

There was a correlation of –0.10 between ACS Exam score and missing SAT data, indicating a near negligible relation between the two. For that reason, there is no evidence that this source of missing data affects the generalizablility regarding students who enter the course without SAT scores on record. This sample of 2255 was used for the Time Series Model analysis.

Among the 2255, ACS Exam scores were available for 1747 students (77.5%). This sample was used for the External Exam Model. Missing data in this case very likely represents students who did not finish the course (the ACS Exam, taken as the final, was mandatory for course completion). Students who did not take the ACS Exam tended to have lower SAT sub-scores with a correlation of –0.17 between Math SAT and missing the ACS Exam. This result approaches a small effect size as proposed by Cohen, [70] so it is safest to conclude that missing data resulting from missing the ACS Exam is non-random. The External Exam Model results presented in this study can therefore only be generalized to those students who completed the course. Examining missing ACS Exam data by class showed that the reform classes had a similar percentage of students not finishing the course as the other classes, making a fair comparison for evaluating the reform. Class sizes ranged from 60 to 133 students, providing a sufficient sample size to evaluate each class.

HLM analyses were conducted using the PROC MIXED procedure in SAS version 8.2. The analysis and interpretation were guided by Singer, [71] who, addressing her own warning that the PROC MIXED procedure is flexible enough to cause errors in model specification that can lead to misinterpretation, has systematically described the necessary procedures for using HLM to consider an education setting.

www.manaraa.com

Instruments: The ACS Exam, Midterm Exams and SAT Sub-scores

There were two different kinds of course exams providing data for this study: an externally-constructed final exam given at the end of each semester to all students in the course at the same time, and several internally-constructed midterm exams written by a panel of course instructors and given during a semester, again to all students at the same time. The final exam was a product of the American Chemical Society Examinations Institute [72] and features 19 algorithmic questions and 21 conceptual questions. Each question has five multiple-choice answers. Construct validity was established by the ACS Exams Institute, which designed the exam for first semester general chemistry students, and by a panel of course instructors, who reviewed the exam and determined the material was appropriate for the students and the course. The ACS Exam has a high relation ($r > 0.5$) with the internally-constructed exams, indicating convergent validity, and a Cronbach's alpha of 0.83, indicating adequate internal consistency. The internally-constructed exams had slightly less internal consistency, with Cronbach's alpha ranging from 0.70 to 0.80, and had moderate correlations among the set, ranging from 0.481 to 0.645. These exams had between 20 and 25 multiple choice questions, each with five multiple choice answers, and covered specific topic areas within the course. All instructors submitted questions for inclusion on each of the internally-constructed exams, and each exam was reviewed by the instructors for content and level of difficulty before being administered to students.

SAT sub-scores were obtained from the university's registrar as they were reported from the Educational Testing Service. SAT sub-scores have been found to have

www.manaraa.com

reliability coefficients exceeding 0.9 and a large body of research has demonstrated predictive validity towards college grades, convergent validity with other predictors used in admissions, and construct validity by panel reviews and item analysis.[73] As such, they are a suitable measure of pre-existing achievement gaps.

Analysis and Discussion

External Exam Model Results

The outcome variable, labeled Y_{ij} in equation 1, is the number of correct responses on the ACS Exam, ranging from 0 to 40. The sample is composed of the 1747 students previously described. An analysis for outliers (discussed below) recommended the removal of 7 students, resulting in a sample size of 1740 students. Descriptive statistics of the variables used in this model are presented in Tables 3.1 and 3.2.

Table 3.1 – Descriptive Statistics on Student Level

	Student Level		
	PLGI N = 287	**Non-PLGI** N = 1453	**All Students** N = 1740
Math SAT			
Mean	576.2	576.0	576.0
Std. Dev.	85.0	86.3	86.1
Maximum	780.0	800.0	800.0
Minimum	270.0	290.0	270.0
Verbal SAT			
Mean	547.9	552.5	551.8
Std. Dev.	87.2	85.9	86.1
Maximum	800.0	800.0	800.0
Minimum	240.0	200.0	200.0
ACS Exam			
Mean	22.8	21.9	22.0
Std. Dev.	6.9	6.8	6.8
Maximum	38.0	38.0	38.0
Minimum	8.0	6.0	6.0

www.manaraa.com

Table 3.2 – Descriptive Statistics on Class Level

	Class Level		
	PLGI N = 3	**Non-PLGI** N = 13	**All Classes** N = 16
SAT			
Mean	1120.0	1127.0	1126.0
Std. Dev.	15.2	20.2	19.1
Maximum	1132.4	1166.4	1166.4
Minimum	1103.0	1077.0	1076.9
ACS Exam			
Mean	22.8	21.9	22.1
Std. Dev.	0.9	1.2	1.2
Maximum	23.8	23.9	23.9
Minimum	22.1	19.1	19.1

From Tables 3.1 and 3.2 it can be seen that the class level parameters are similar to the student level parameters, except they are noticeably more stable, with deviations and ranges markedly reduced. This is an initial indication that the class average scores do not vary a great deal from one class to another as compared to student scores. The students in the sample tend to span almost the entire range of possible SAT and ACS scores. Also note in Table 3.1 that the variance for the PLGI and Non-PLGI classroom are similar on the outcome measure, which satisfies the homogeneity of variance assumption that underlies the models that will be introduced. Finally, the students in the PLGI class are equivalent to the students in the non-PLGI classes on all measures except the outcome measure, a promising finding for the effectiveness of the reform.[74]

Covariance parameter estimates indicate that the between class variance accounts for relatively little of the overall variance (between class variance = 0.900, within class variance = 46.391, intraclass correlation coefficient = 1.9%) in the outcome measure as compared to the between student variance. This observation makes sense, as it indicates that class values for the ACS Exam are more stable than the student values. This may

www.manaraa.com

also serve as an indication that the class itself provides very little influence on student, and that treating students within a class as independent (as in a multiple regression) may be a valid approximation.[61] However, since the reform was implemented at the class level, HLM is still the more appropriate method for examining these effects.

The External Exam Model was initially run as prescribed in Equations 1 and 2. Level 1 predictors, SAT sub-scores, were centered on the class means and the Level 2 predictor, class average SAT scores, were centered on the grand means to improve model convergence and ease in interpretation.[75] The other Level 2 predictor, reform, describes the peer-led guided inquiry intervention, and consequently appears as PLGI in equation 6. The model converged using the minimum variance quadratic unbiased estimates method, which is appropriate for large data sets, [76] and the "between/within" method for estimating the denominator degrees of freedom for the significance tests of the coefficients was used. The initial model was evaluated for outliers by examining residuals. The residuals were found to follow a normal distribution, with a mean of approximately zero and a standard deviation of 5.27. Students with residuals of greater than 3 standard deviations from the mean (N = 7) were removed as outliers, and the analysis was run again with similar results as the initial model. For space reasons, only the model with the outliers removed is presented here. Tables 3.3 through 3.5 provide the estimates for the coefficients in this model.

Table 3.3 - Estimating the Intercept Coefficient (β_{0j})

Symbol	Description	Estimate	Std. error	Sig.
γ_{00}	Intercept	21.86	0.21	<0.001
γ_{01}	Class SAT average	0.0435	0.0104	0.0011
γ_{02}	PLGI	1.19	0.51	0.0351

www.manaraa.com

Table 3.4 – Estimating the Slope Coefficient (β_{1j}) Relating Student Math SAT to ACS Exam

Symbol	Description	Estimate	Std. error	Sig.
γ_{10}	Intercept	0.0365	0.0030	<0.001
γ_{11}	Class SAT average	0.00033	0.00016	0.0349
γ_{12}	PLGI	0.00466	0.00750	n.s.

n.s. = non significant ($p > 0.050$)

Table 3.5 – Estimating the Slope Coefficient (β_{2j}) Relating Student Verbal SAT to ACS Exam

Symbol	Description	Estimate	Std. Error	Sig.
γ_{20}	Intercept	0.0188	0.0030	<0.001
γ_{21}	Class SAT average	-0.00017	0.00015	n.s.
γ_{22}	PLGI	0.00381	0.00724	n.s.

n.s. = non significant ($p > 0.050$)

The coefficient estimates can also be represented in equation form:

$$\begin{aligned} \beta_{0j} &= 21.86 + 0.0435 * SATavg + 1.19 * PLGI \\ \beta_{1j} &= 0.0365 + 0.00033 * SATavg + 0.00466 * PLGI \\ \beta_{2j} &= 0.0188 - 0.00017 * SATavg + 0.00381 * PLGI \end{aligned} \qquad (6)$$

the outcome variables from these equations provide a set of coefficients for equation 1 when the class parameters are specified. In other words, this model can be used to predict the performance of an individual student on the basis of class level information (average SAT subscores, presence or absence of the PLGI reform) and student level information (student SAT subscores). The appropriate method for calculating the amount of explained variance for HLM is debatable, but the application of Snijders' and Bosker's [77] interpretation indicates 42.9% of the student level variance and 55.0% of the classroom level variance is accounted for by this model.

Interpretation begins with the significant parameters. First, note the impact of the class variables on the Level 1 intercept, β_{0j}, as presented in Table 3.3. The Level 2

www.manaraa.com

intercept, γ_{00}, gives the expected student score on the ACS Exam for a student whose SAT sub-scores match the class average. Without PLGI, and in a hypothetical class with an SAT average of the grand mean (1127.7), the expected score is 21.86. This expected score increases with an increase in class SAT average, as the positive γ_{01} describes. Finally, the positive γ_{02} indicates that, after controlling for class SAT average and individual student SAT sub-scores, the reform is expected to result, on average, in scores 1.19 points higher on the ACS Exam. PLGI can therefore be termed an effective reform: it improves performance on the ACS Exam.

The intercepts in Table 3.4 and Table 3.5 reflect the relation between the outcome measure and a student's SAT sub-scores. As has been seen elsewhere, students with higher SAT sub-scores tend to score higher on chemistry measures. As discussed, classroom conditions that lower these values would produce a more equitable classroom, in which a student's final exam score is not as dependent on the SAT sub-scores with which he or she enters the class. From Tables 3.4 and 3.5 (in particular, γ_{12} and γ_{22}) it appears that there is no evidence of a significant effect of the PLGI reform on the dependence on student SAT sub-scores. That is, the PLGI reform tends to have no effect on the equity condition already present in the classroom. Although this analysis indicates that PLGI cannot be termed an equitable reform, neutrality is still preferable to the potential for negative impact as discussed earlier.

Even though the model results demonstrate that PLGI is neutral with respect to equity, they do indicate a potentially negative impact on equity associated with a class-level variable. Only one class-level variable significantly affects a slope: γ_{11} shows that the higher the class average SAT score the stronger the relation between Math SAT and

www.manaraa.com

the outcome measure. In other words, classes with high average SAT scores tend to separate students based on Math SAT more, while classes with low average SAT scores tend to be more equitable regarding entering math ability. This relation is relatively sizeable. As one goes from a class with a mid-range average SAT score to a class with the maximum average SAT score (see Table 3.2) the relation of student Math SAT to the outcome measure increases by 36% (going from 0.0365 to 0.0497). The overall advantage (described by γ_{01}) of having a high class average SAT is therefore offset somewhat by a disadvantage for students in that class with low Math SAT scores, and there is a tipping point below which the disadvantage prevails. In concrete terms, the model would predict that a student with a Math SAT of 400 would decrease his score by 0.57 points on the outcome measure by joining a class with a high class average SAT. This negative impact would be even more pronounced for students with slightly lower Math SAT scores and less pronounced for students with slightly higher Math SAT scores, but students with Math SAT of 450 or above in this class would not experience any disadvantage. This situation is undesirable, since it puts students with low Math SAT scores at even further disadvantage, but it is not associated with the PLGI reform.

Time Series Model Results

The analysis decisions for the Time Series Model followed a similar pattern to the External Exam Model. For example, individual student SAT sub-scores (Level 2 variables) were centered on class means and class average scores (Level 3 variables) were centered on the grand mean. The model converged using the minimum variance quadratic unbiased estimates method. An analysis of residuals greater than 3 standard

www.manaraa.com

deviations revealed 11 data points that disagreed with the general trend. Analysis was run both with and without these data points with similar results; the results presented have these data points omitted. One additional issue that was not relevant for the External Exam Model did need to be addressed for the Time Series Model. Since midterm exam questions (though not content) differed from semester to semester, the appropriateness of combining data across the three semesters could be questioned. To address the issue, exam scores for each semester were standardized (so each exam had a mean of 0 and standard deviation of 1) prior to combining the data. The model was run with both standardized and unstandardized scores, both resulting in similar interpretations. This similarity between standardized and unstandardized results was central to the decision to combine data across semesters. Because both methods led to the same conclusions, only the results for the unstandardized model, for space reasons and for ease of interpretation, are presented here. The standardized model results are present in Appendix D.

In the External Exam Model analysis, missing ACS Exam data meant that those who did not finish the course (i.e. did not take the final) had to be omitted from the analysis, which limited the generalizability of the results. For the Time Series Model analysis, midterm exam scores can be used to measure the performance of students who did not finish the course; however, missing data is still an issue. As part of the course guidelines, students took four midterm exams during the semester and could drop their lowest exam score for calculation of their grade. If a student missed an exam, the missed exam simply counted as the student's dropped exam score. This policy resulted in a large amount of missing data, as presented in Table 3.6, which explores the exam-taking

www.manaraa.com

patterns of the 2255 students with SAT sub-scores in both reform and non-reform sections.

Table 3.6 – Missing Data on Midterm Exams

Description	Total Frequency	Total Percent	PLGI Percent	Non-PLGI Percent
Completed all exams	1737	77.0%	76.8%	77.1%
Miss 1 of the 4 exams	137	6.1%	5.5%	6.2%
Miss 2 of the 4 exams	203	9.0%	10.4%	8.7%
Miss 3 of the 4 exams	178	7.9%	7.4%	8.0%

First, from Table 3.6, note there is no consistent relation between the PLGI sections and the non-reform sections in terms of class size retention. Roughly 17% of the sample missed more than one exam, leaving only one or two data points for these participants. HLM can estimate coefficients for these students, but there is some concern over how legitimate these coefficients can be with only one or two data points.[64] The HLM Time Series Model was therefore run twice: once with all data points and again omitting those who had missed more than one exam. Both models had similar results and identical interpretations, which agrees with research recommendations that missing data on Level 1 does not strongly affect HLM results.[78] In light of this, the model presented includes all data points available from the sample of 2255 students and is thus representative of students who begin the course. Descriptive statistics for the midterm exams between the PLGI and non-PLGI groups is shown in Table 3.7.

www.manaraa.com

Table 3.7 – Descriptive Statistics for Midterm Exams

	Student Level		
	PLGI N = 365	**Non-PLGI** N = 1873	**All Students** N = 2238
Exam 1			
Mean	57.80	58.91	58.73
Std. Dev.	17.43	16.73	16.85
Exam 2			
Mean	53.90	52.79	52.97
Std. Dev.	18.90	18.71	18.74
Exam 3			
Mean	54.00	49.85	50.52
Std. Dev.	18.57	19.08	19.05
Exam 4			
Mean	56.52	53.87	54.30
Std. Dev.	20.10	20.51	20.46

Intraclass correlation coefficient (ICC) values were also calculated for midterm exam scores in each semester. In general ICC values increased as the semester progressed, as shown in Table 3.8, so that by the fourth midterm exam, class accounted for approximately 5% of the variance in student exam scores. This finding lends support to the decision that HLM, with its ability to handle class effects as well as student effects, is the appropriate approach to take for this data.[79]

Table 3.8 –Intraclass Correlation Coefficients for Midterm Exams

	Semester 1	**Semester 2**	**Semester 3**
Exam 1	2.6%	0.6%	1.5%
Exam 2	1.5%	0.8%	3.1%
Exam 3	4.7%	0.8%	2.5%
Exam 4	6.7%	2.6%	4.2%

The Time Series Model was run as prescribed in Equations 3, 4 and 5. The results of the full Time Series Model showed that none of the Level 3 variables related significantly to the slopes in Level 2, only to the intercepts. Since PLGI is a Level 3

www.manaraa.com

variable, the model results indicate no evidence of an effect of the reform on the equity within the classroom. This finding is congruent with the External Exam Model analysis. For the sake of parsimony, the Time Series Model was re-run with Level 3 variables relating to the Level 2 intercepts only, with similar overall results and greater clarity of interpretation. The results of this abbreviated Time Series Model are presented in Tables 3.9 and 3.10.

Table 3.9 - Estimating the Intercept Coefficient (π_{0jk})

Symbol	Description	Estimate	Std. Error	Sig.
γ_{000}	Intercept	56.93	0.36	<0.001
γ_{001}	Class SAT	0.0812	0.0176	<0.001
γ_{002}	PLGI	-0.728	0.879	n.s.
β_{01k}	Student Math SAT	0.0850	0.0049	<0.001
β_{02k}	Student Verbal SAT	0.0209	0.0049	<0.001

n.s. = non significant ($p > 0.050$)

Table 3.10 - Estimating the Slope Coefficient (π_{1jk})

Symbol	Description	Estimate	Std. Error	Sig.
γ_{100}	Intercept	-3.08	0.23	<0.001
γ_{101}	Class SAT	-0.00717	0.01169	n.s.
γ_{102}	PLGI	1.55	0.58	0.007
β_{11k}	Student Math SAT	-0.00421	0.00323	n.s.
β_{12k}	Student Verbal SAT	0.00183	0.00321	n.s.

n.s. = non significant ($p > 0.050$)

Interpretation rests on the significant parameters. First, consider the estimates of the intercept, π_{0jk}, in Table 3.9, which is showing the variables' influence on Exam 1 scores. The estimated value of the intercept γ_{000} (56.9) is the average Exam 1 score for a student who meets the following conditions: SAT sub-scores equal to the class average, class SAT scores equal to the grand mean, and not in the PLGI reform class. As expected, the positive and significant β_{01k} and β_{02k} indicate that student SAT sub-scores

www.manaraa.com

affect performance on the first exam: those with higher SAT scores tended to score higher on the exam. Interestingly, the γ_{001} result shows that the class SAT score also has an impact on the Exam 1 score. While class SAT score has an estimated value similar in magnitude to student's Math SAT sub-score, considering the range of each variable (see Table 3.1) indicates the impact of the class effect is somewhat smaller. Students' Math SAT sub-scores may vary by 100 points, for example, yielding a difference of over 8 percentage points on an exam, but Class SAT varies at most by 10 points, for a change of just over 0.8 percentage points. Also note from this table that the PLGI reform (see γ_{002}) does not have a significant impact on Exam 1 scores.

The slope coefficient, π_{1jk}, relating time and exam performance, is addressed in Table 3.10 and has fewer significant parameters to be interpreted. First, the intercept γ_{100} (-3.08) shows that, as students progress through the course, the tendency is for their exam scores to decrease. This tendency, though regrettable, corresponds with common experience. It may be that students are less familiar with concepts presented later in the course, or perhaps the exams themselves become more difficult. Second, the positive and significant γ_{102} (1.55) indicates that PLGI mitigates this tendency. In other words, for PLGI classes the overall slope relating time and exam performance (π_{1jk}) becomes –1.53 (calculated by –3.08 + 1.55), which is noticeably higher than for classes without the reform. To place this result in context, consider a calculation based on average Exam 4 scores (time = 3 in the model). For the PLGI group the average Exam 4 score would be 52.4, but for the group without reform the average would be 47.7. This difference of 4.7 percentage points translates to roughly 0.25 of a standard deviation, which is a fairly substantial effect. Also notable from Table 3.10 are the non-significant parameters:

www.manaraa.com

student SAT sub-scores and the class SAT score do not relate to the overall slope. Apparently, while these factors have a strong influence on how students start the course, they do not impact the change in performance over the course of the semester.

Conclusions and Implications

The first aim of this study was to demonstrate the use of HLM for a quasi-experimental study investigating a pedagogical reform. To achieve this aim, distinct models were developed for both single and time-series measures of academic achievement, and a detailed discussion of variable considerations and missing data was provided. Additionally, model interpretation investigating both effectiveness and equity was discussed. The implicit expectation underlying this detailed description is that researchers in similar settings will be able to use similar models for their own work. Given the national impetus to engage in educational research that can ultimately produce "improvement in student academic achievement [and] reduction in the achievement gap between high-performing and low-performing students," [80] it is imperative to showcase research tools that allow high quality investigation of these two goals. As discussed, the rationale for the use of HLM involves three major considerations: 1) its freedom from reliance on an independence of observations assumption, 2) its ability to combine data for multiple levels of analysis, and 3) the ease with which it can be used to speak to issues of equity as well as to efficacy.

The question of independent observations should be a primary concern for any quantitative analysis of educational reform. During the design phase of an investigation, the extent to which the independence of observations assumption is likely to be violated

www.manaraa.com

in any given setting must be decided on a theoretical basis, but, as shown here, there is relatively little variability among large classes in general chemistry on the externally-constructed ACS Final Exam. For this situation, the assumption may be valid, and designs relying on regression models or ANOVA are reasonable. However, as observed for the midterm exams in this study, more caution is warranted for internally-constructed exams, particularly as the semester progresses. As one would expect, the common experience of students who are in class together seems to grow as the semester progresses. Even for the large classes in the study, by the end of the semester intraclass correlation coefficient values were as high as 6.7% for the midterm exams, over 3 times what was found with the externally constructed final exam.

Ideally students in a particular class interact with each other and with the instructor, developing shared understandings. For example, students can talk with each other outside of class to interpret class notes and follow these conversations with in-class questions that are responded to by the instructor for the benefit of all students. As a result, students in a class have a reasonable chance of developing common interpretations of terminology, particular ways of expressing and interrogating important concepts, and, as they experience exams, collective wisdom about how to handle exam questions. This scenario of shared interpretations is typically desirable from a pedagogical point of view, but it can create problems for research. The findings indicate that researchers investigating academic performance among students in college classrooms may more safely rely on the independence of observations assumption if the data is collected near the beginning of a course or if the measures used are externally constructed. This assumption is less tenable as students progress within a class, for example with measures

www.manaraa.com

administered late in the semester or for studies of classes that are intact for longer than a semester. From a teaching perspective, for situations where multiple sections of a course take common exams, the findings suggest the use of external final exams (i.e. exams obtained from an outside source) is warranted in cases where it is desirable to reduce classroom effects on the final grade.

The second aim of the study was to use the ability of HLM to harness student-level and classroom-level data to investigate the efficacy of a reform applied at the level of the classroom. The results from the HLM Time Series Model and from the HLM External Exam Model are markedly congruent. PLGI was associated with improved performance on the ACS Exam given at the end of the semester, regardless of student SAT sub-scores or class SAT average. This result matches the trend in the Time Series Model in which students in classes with the PLGI reform, regardless of student SAT sub-scores or class SAT average, experience a much less severe drop in performance as the semester progresses than do their counterparts in classes without the PLGI reform. The observations of improved performance are in agreement with other research promoting the effectiveness of cooperative learning and inquiry as pedagogical tools, and the analysis spans three years of implementation in a college general chemistry course. Based on these results, the use of inquiry activities as the shared goal of cooperative learning groups, with PLGI as an option for a large class, is recommended for other entry-level college science courses. Large classes are commonplace at the college level but do not lend themselves readily to pedagogical reform, [81] so PLGI may be particularly helpful in providing a feasible model for implementation of reform in this setting.

www.manaraa.com

In both models, student exam scores are influenced by the class average SAT score, even when controlling for individual student SAT sub-scores. This holds true for both PLGI and non-PLGI sections: in general, the higher the class average SAT score, the better a student's performance, after controlling for the student's individual SAT sub-scores. It may be that informal study groups containing a greater number of students with high SAT scores are more beneficial, and that students in a class with a higher average SAT score have a higher probability of forming such study groups, but this is merely speculative and would require further investigation. In the External Exam Model, class average SAT score also featured an interaction with student Math SAT sub-scores, but this trend was not observed for the Time Series Model. This interaction represents the only occasion in the data where an increase in the class average SAT score can reduce a student's performance. For students with a Math SAT sub-score 150 points or more below their class average (e.g. a Math SAT sub-score of 450 or below for this data), the model results indicate that student performance decreases as the class SAT score goes from average to the highest found in the sample. Since this demonstrated inequity was not present in both models, future research into interactions between class average SAT and student SAT to determine their prevalence and associated achievement consequences is recommended. Decisions that affect class average SAT, for example the introduction of course pre-requisites, would benefit from this additional knowledge.

The third aim of the study was to exploit HLM's ability to examine the equity of a pedagogical reform. While the effectiveness of PLGI is an important outcome, it would have been desirable also for it to reduce the dependence of exam performance on SAT sub-scores, thus promoting equity by reducing the outcome gap between students with

www.manaraa.com

different SAT sub-scores. Preparation for college can be thought of as a mediating variable for several other characteristics that are associated with achievement gaps in college science. For example, controlling for the number of high school science classes taken weakens the relationship between SES and college science performance.[82] Similar trends were also found regarding student sex [61] and student racial/ethnic identities.[83] In this study, SAT sub-scores served as an indicator of preparation for college. Unfortunately, both models feature non-significant interaction terms associated with the potential relationship between PLGI and student SAT sub-scores. In other words, no evidence for any effect on the current equity situation in the classroom was found, and, under PLGI, the dependence of student performance on incoming preparation remains similar to what is seen in a traditional lecture environment. From an examination of the results of both models, PLGI tends to help all students; however, it does not improve on the present state of equity in the classroom.

Past research led us to the expectation that the reform could either increase or decrease the existing achievement gaps in the classroom. Given the work of Cohen and Webb, the possibility of an increase in achievement gaps resulting from perceptions of low status was of particular concern, because the reform pedagogy relies on student-student interaction and reveals student status roles more clearly than a lecture situation. In this light, the fact that we found the reform had no effect on equity is a positive result, but certainly a decrease in achievement gaps would be most desirable. The failure to realize a more equitable situation for students with low SAT scores is cause to focus a future investigation explicitly on this group in the reform setting. An in-depth study focusing on student-student interactions, similar to Cohen's work with middle-school

www.manaraa.com

students, would help uncover any detrimental effects arising from status perceptions. Another possibility is that the reform is working for students with low SAT scores, and in time would reduce achievement gaps, but the one semester, once-per-week intervention is too limited for results to be seen. If so, equity must be a harder goal to reach than effectiveness, since the results do indicate that even this limited application of reform is effective. Additionally it could be an indication of low statistical power for the interaction of PLGI on the effect of SAT sub-scores, in particular as interaction terms have reduced power from main effects.[84] Although this is a possibility, the estimate of the impact the reform has on SAT sub-scores is less than 13% of the original effect SAT sub-scores Math SAT relation is 0.0365, the PLGI reform impact would increase it by 0.00466. Thus even though power may not be sufficient to find significance, by the data available it is unlikely that the reform constitutes a meaningful change in the relation between SAT sub-scores and in-course performance.

Further investigations that examine both effectiveness and equity of more intensive reforms would be an important first step in discerning whether it is the limited nature of the intervention that prevents equity. In the final analysis, however, available evidence indicates that PLGI does not put students with low SAT sub-scores at any more of a disadvantage than is present with traditional lecture-style teaching. Considering that the results demonstrate that PLGI is an effective reform, it is better for students with low SAT sub-scores to experience a lecture course with PLGI than to experience one without PLGI.

www.manaraa.com

IV. Formal Thought as an Independent Measure

The PLGI reform was found to be effective, but had a negligible impact on the relationship of student SAT sub-scores to performance. That is, students' performance in the collegiate level course was still dependent, to some degree, on their prior high school preparation. Recommendations to improve the achievement of students entering college chemistry with low SAT sub-scores may include math or linguistic skill (analogies) tutorials.[85, 86] This chapter looks at another construct, formal thought, that relates an incoming ability to achievement in chemistry. Among specific interest is whether this construct is independent of SAT sub-scores as they relate to chemistry achievement. If so, than this would be another dimension on which to evaluate the equity created by the reform. As mentioned before, to establish the role of formal thought without the impact of the reform, this study examines both fall and spring semesters but does not include any students who participated in the reform.

Justification for the Study

All too often, substantial numbers of students in college fail to demonstrate sufficient understanding of chemistry to proceed beyond the introductory course, general chemistry. This circumstance hinders not only the individual student but also the field of chemistry. While the costs to the individual are immediate and obvious (not only the regrettable lack of knowledge of chemistry but also a closed door to any major field of

www.manaraa.com

study requiring that knowledge) the costs to chemistry are also significant. With each year this trend continues, chemistry loses numerous individuals who now will not contribute to the growth of the discipline. Indeed the ramifications stretch beyond chemistry, as other science curricula require general chemistry prior to course work within their program.[61] Students who cannot muster an acceptable understanding of general chemistry are prevented from contributing to many science fields. On a more systemic level, the inability of students to continue in science-oriented courses because of low performance in general chemistry represents a major setback in efforts to create a scientifically-informed populace and a technically-proficient workforce. For these reasons, unsatisfactory student performance in college-level general chemistry remains a critical area of concern.

Since basic constructivism indicates that the prior knowledge and skills with which students enter a course play a role in success (or its absence), it is both possible and valuable to identify students who are at-risk of not succeeding in a course at the point when they first enter the course. To do so provides the opportunity for assisting these students early on, while success is still possible. Further, knowledge about the factors contributing to low (at-risk) performance can inform the design of interventions aimed toward reducing the challenges faced by these students. The first task is identifying the at-risk population with reasonable accuracy, and the second is suggesting potential interventions. Ideally the measure used for identification contains within itself implications for a potential remedy. This chapter compares the accuracy, degree of overlap, and implications for potential interventions of two measures that can be used to identify students at-risk of not succeeding in general chemistry. It therefore joins a long

www.manaraa.com

history of 'predictor papers' but is unique in its combination of generalizability, a focus on at-risk students, and consideration for the implications of choosing a particular predictor.

Formal Thought and Science Achievement

With the intent of identifying at-risk students, the predictor selection had to be focused on a measure that has the potential to describe a large hindrance for students. The work of Piaget provides a reasonable approach to this problem. Piagetian theory details four successive stages of cognitive development, the latter two: concrete operational and formal operational, providing the most interest for this study. Students operating at the concrete operational stage tend to focus on those concepts which can be directly perceived, [87] and are unable to abstract to generalizations beyond their senses. As Shayer and Adey describe them, these students often improperly equate associations with causal relationships and have difficulty in dealing with situations that feature more than 2 variables.[88] The ability of concrete thinkers to model is also limited, with classifications and ordering always centered on two aspects at a time, with one of the variables treated implicitly as dependent.[88]

Formal operational thought is the last stage of cognitive development as described by Piaget, in which "deduction no longer refers directly to perceived reality but to hypothetical statements."[87] In formal thought then, possibilities are regarded as hypothetical at first, and then verified by empirical evidence. Contextually, this ability leads to the meaningful manipulation of empirical results, as well as a familiarity with the abstract. Also taken from Piaget's work is a series of reasoning patterns that would

www.manaraa.com

describe formal thought operations. Adey and Shayer [89] grouped the reasoning patterns into three main categories. The first category, the handling of variables, includes the control and exclusion of variables, the recognition of multiple classification schemes, and the description of combinatorial possibilities. The second category, relationships between variables, includes the use of ratios, and proportion (comparing of two ratios), as well as compensation (use of inverse relationships), correlation and probability. The final group, formal models, describes the creation of an abstract representation of complex behaviors. Also included in this last group is the use of logical reasoning. Keeping with Piagetian theory, the onset of formal thought would be characterized by the development of all the cognitive operations at about the same time, a postulate that has been supported by empirical evidence.[90-92]

Alternative theories of cognitive development that may also prove fruitful in identifying at-risk students in college chemistry exist. The principle difference between the alternative theories is global restructuring versus domain-specific restructuring.[93] Piagetian theory postulates that developmental change occurs via global restructuring, which the concrete and formal stages are examples of. The progression to a new stage is thought to affect an individual's approach on a variety of diverse subjects.[89] As an alternative, domain-specific restructuring postulates that an individual can incorporate new concepts or theories within a specific domain independent of the individual's general logical capabilities.[93-95] Within a domain specific restructuring framework, identification of at risk students may more appropriately be done with a chemistry diagnostic test [96] and remediation then would be explicitly designed around an individual's existing chemistry knowledge.[97] However, given the substantial amount

www.manaraa.com

of research that indicates the utility of Piagetian theory in science success, this study utilizes Piagetian theory in the investigation to identify at-risk students.

One noteworthy example is the work of Lawson and Renner [92] who showed that students at the concrete operational stage are unable to develop an understanding of formal concepts, and that students at the formal operational stage demonstrate an understanding of both formal and concrete concepts. Later on Lawson [98, 99] points out that such results could be interpreted largely as a spurious correlation, describing what might be a more general intelligence measure underlying the success seen on both measures. In the 1982 study [99] a partial correlation between formal thought and biology achievement while controlling for fluid intelligence was conducted, finding a significant relation present, illustrating that it is the formal thought measure that better corresponds to this achievement measure.

It is also the aim to continue this investigation by examining whether formal thought features a unique relationship to achievement in college chemistry. In addition, while it has been demonstrated that controlling for one general intelligence measure did not effect the relation between formal thought and achievement, it has not been investigated whether a general achievement measure, such as SAT, may also underlie such a relationship. For example, would a student with high SAT scores, but low formal thought measure, still be expected to perform poorly? Or are students with low SAT scores and high formal thought scores expected to perform poorly? This investigation will be focused on those students who are at-risk of performing poorly in general chemistry. Since formal thought has been described as one of a series of factors necessary for a successful performance, [100-102] the presence of formal thought would

www.manaraa.com

not be expected to predict a successful performance, but rather the absence of formal thought is expected to envisage a poor performance. This expectation leads to a series of research questions:

RQ1. Which predictor, SAT or a formal thought measure, is better able to identify at-risk students?

RQ2. Are the at-risk students identified by each predictor distinct groups, which may lead to more specific interventions geared toward each group of students?

RQ3. Can a combination of SAT and formal thought measures provide a distinct advantage in identifying at-risk students?

RQ4. And, to what extent are all at-risk students identified by this set of predictors?

Past Work with Predictors

Extensive work has been done on the ability to predict students' college GPA. The predictors that have been used span a wide range including academic orientation [103, 104], self-efficacy [105], student personality traits,[106, 107] student approaches to learning [108] along with SAT scores [103, 104, 106, 107] or high school rank [103]. The use of college GPA as an outcome variable may certainly lead to the ability to distinguish a student who is in danger of not completing college, and may lead to certain interventions aimed at preventing this problem. However, it can provide no course-specific information that may play a role in student's lack of success, such as the demands of a course, or the assumptions a course may make of the skills or knowledge of incoming students.

www.manaraa.com

Narrowing the focus to success in college chemistry, considerable work has also been performed with this aim. Past studies have examined the ability of SAT [66, 68, 109-113], ACT [113-115], high school GPA [114], high school chemistry grade [112, 113, 116], personality characteristics [115] and Piagetian tasks [68, 111] to predict final chemistry course grade. The use of chemistry grades as an outcome variable, however, relies on the ability of chemistry grades to approximate chemistry understanding. The extent to which this approximation is valid depends on several decisions peculiar to the course, the instructor, and the institution. Decisions such as grading on a curve or on an absolute scale, grading based completely on exam performance versus consideration of student homework, the allowance of extra credit, and even the method by which the exam were created, can all alter the extent chemistry grades reflect true student understanding of chemistry. As a result, the generalizability of the above studies depends on the extent that the grading used is applicable to other institutions. However, of the studies presented above, the work by Bender and Milakofsky [111] is the only one to provide detailed evidence of the grading procedures employed.

Another option for an outcome variable is the use of a single exam, as a measure of students' chemistry understanding. In addition to providing a clear picture of what constitutes success in chemistry, the scoring of a single exam lends itself to the statistical procedures commonly used with predictors. One example of such a procedure is present in Yager's 1988 article [117], for an examination of the effects of taking high school chemistry. In this study, students were measured on a normed exam, a course final exam, and by final course grade to provide multiple measures of success in chemistry. In particular, the use of a normed exam allows for a ready assessment of generalizability.

www.manaraa.com

In the current study, the ability of a formal thought measure to identify at-risk students is examined. Formal thought was chosen as the predictor because of the theoretical basis for its inclusion, but also due to the extensive prior work in the literature aimed at improving formal thought.[118-122] In addition, the outcome variable for the current study is an available-to-public normed exam designed to measure student understanding of chemistry, making the results generalizable to the extent that the content of this exam matches the desired outcomes at other institutions.

Instruments: Test of Logical Thinking and ACS Exam

Several measures of formal thought have been developed, validated and utilized in the research literature. What these measures share in common is an attempt to approximate the original Piagetian interviews. To begin, emulating the Piagetian interview is problematic, especially with large groups of students, owing to the time-intensive nature of the interview procedure. As a result written exams, in particular, have been constructed to take the place of these interviews. Perhaps the closest approximation to the interview procedure is Shayer & Adey's Science Reasoning Tasks, [88] where written predictions are elicited from students, followed by an instructor performing a demonstration, and then students are asked to explain the phenomena witnessed. Depending on the task, questions may be free-responses or students select among the responses available.

However, for the present study, with class sizes approaching 200 students, there was concern about the timing for student responses and doubt that all students would be able to adequately witness a demonstration. As a result, a completely written exam was

www.manaraa.com

chosen. Among the possibilities present are the Inventory of Piagetian Developmental Tasks IPDT [111], the Group Assessment of Logical Thinking GALT [123], the Test of Logical Thinking TOLT [124] and Piagetian Logical Operations Test PLOT.[125] Of these choices, the TOLT was selected owing to its: ease of administration, two-tiered question design, which reduces the possibility of students guessing the correct answer [126], published validity [124], and use in the research literature.[127-129]

The TOLT was developed and validated by Tobin and Capie to measure what they termed formal reasoning ability. In order to do so items previously used by Lawson [101, 130] were selected so that the test comprised of two items for each of the five modes of formal reasoning: controlling variables, proportional reasoning, probabilistic reasoning, correlational reasoning and combinatorial reasoning. These modes are based on Shayer and Adey's second category of reasoning patterns that are evidence of formal operations. To receive a correct score for each item students need to select the correct answer from up to 5 choices, and select the correct reason for the answer from 5 possible reasons. The only exceptions are the combinatorial reasoning questions where students are required to list all the correct combinatorial possibilities without any replication. Construct validation of TOLT was done by relating student scores on the TOLT with student performance via interviews, for students ranging from grade 6 to college. By relating this instrument to college chemistry performance, predictive validity will be investigated.

As noted, in the past a large variety of predictor papers rely on student grades as an outcome variable. As a research base, the results of these studies are generalizable only to the extent one can assume that student grades at the research institution match the

www.manaraa.com

desired student outcomes of other locales. More importantly, without extensive details of the grading scheme used, this assumption becomes impossible to assess. With the desire to produce a generalizable model for identifying at-risk students, a normed exam produced by the American Chemical Society (ACS) was selected. "The American Chemical Society is a self-governed individual membership organization that consists of more than 163,000 members at all degree levels and in all fields of chemistry."[131] As part of the ACS the Division of Chemical Education features an Exams Institute where exams are available to chemistry teachers and administrators in high schools, colleges, and universities.

The exam institute offers more than fifty exams covering general chemistry, organic chemistry, analytical chemistry, physical chemistry, inorganic chemistry, biochemistry, polymer chemistry, and high school chemistry.[72] The first semester general chemistry exams include various lengths of a conventional exam and a special examination (SP97A) meant to combine conceptual knowledge questions with the conventional (algorithmic) type questions. Given the recent push towards conceptual understanding of chemistry in the research literature [132-135] both conceptual and conventional assessment methods play an important role in the objectives of most general chemistry courses. As a result the ACS special examination was selected as the outcome variable for this study. Though this exam played a large role in determining student grades (it served as the final exam for the course, which was 25% of student grades), there were other factors that also contributed to student grades. Thus, it is possible for a student to successfully complete the course despite a poor performance on the exam. Given this acknowledgement, this study may be viewed more appropriately as identifying

www.manaraa.com

students at-risk in terms of the successful demonstration of chemistry knowledge on an exam at the end of the course, rather than at-risk of passing the course per se, though these two are highly related. Convergent validity and reliability for the ACS Exam was established by relating to internally-constructed exams and by Cronbach's alpha respectively, and is discussed in detail in the previous chapter.

Research Methods

The TOLT was administered during the first week of classes in 22 sections of the first semester of general chemistry at a large southeastern public urban research university over the course of three academic years. This resulted in TOLT scores for 3798 students, out of an estimated 4180 students enrolled in the 22 sections. At the end of the course, students took a final exam that was a normed American Chemical Society exam to measure student academic achievement. Of the 3798 students, ACS exam scores was available for 2871 students (75.6%). Since completing the ACS exam was a course requirement, the likely reason for not obtaining ACS exam scores were students not completing the course. Student SAT scores were obtained from institutional records. Among the 2871 students that took the ACS exam, SAT scores were available for 2284 students. The most likely cause for missing SAT scores was the student taking the ACT in place of the SAT or the student enrolling in the course after SAT records were pulled. The focus of the analysis was the 2284 students for whom complete data was available. The decision to omit missing data will be revisited in a later section, in particular since the missing data may disproportionately represent at risk students.

www.manaraa.com

The main focus of the analysis will be the 2284 students for where there is complete data. First steps were taken to determine if there were any outliers in the data, so that no single data point would have an unusually large effect on the results of the analysis. Outliers were determined by evaluation of the standardized residuals for a multiple regression model that included both SAT sub-scores and TOLT. By examining for any standardized residuals greater than three, [65] there were nine students found to be inconsistent with the general pattern. The data analysis then proceeded with 2275 students.

Prior to examining the trends between variables, descriptive statistics were evaluated, and are presented in Table 4.1.

Table 4.1 – Descriptive Statistics for Measures Used

	TOLT (0-10)	**Math SAT (200-800)**	**Verbal SAT (200-800)**	**ACS Exam (0-100)**
Mean	6.80	559.14	540.58	52.02
St. Dev.	2.613	83.505	82.648	16.638
Skewness (Std. error = 0.048)	-0.664	-0.048	0.070	0.240
Kurtosis (Std. error = 0.097)	-0.452	-0.166	-0.115	-0.690
25^{th} percentile	5	500	480	40
50^{th} percentile	7	560	540	50
75^{th} percentile	9	620	600	65

The normality tests indicate that the TOLT scores feature a significant negative skew, indicating the scores were more heavily distributed at the higher values. This phenomenon may be a result of the setting of the study, since the TOLT was designed for grades 6 through college, while the sample consists entirely of college students. While most statistical tests rely on a normality assumption, the tests employed are very robust to violations of normality.[136]

www.manaraa.com

For generalizability reasons, it is also necessary that the distributions span the entire range of possible values for each measure. For the TOLT and Verbal SAT, the 2275 students spanned the entire range, but for Math SAT the scores ranged from 290 to 800, out of possible scores that could span from 200 to 800. For the ACS exam scores ranged from 10.0 to 95.0, and the exam was graded as a percent score making the possible range of 0 to 100. Thus the results cannot be generalized to the extreme lower scores on these measures, which may serve as a caution for employing these results in alternative settings, specifically settings which may feature an alternative range of scores, such as early high school.

As described previously, several research questions will guide the nature of this investigation. To investigate each question inferential and descriptive statistics will be used. Inferential statistics will be used to establish the utility of the predictors by relating the predictors to performance and assist in interpretation of the descriptive statistics. Where possible, effect sizes will be reported as a standardized measure of the differences seen, and operationalized using Cohen's qualitative terms: small, medium and large effects. As Cohen describes them, small effects are where the effect is small relative to the effect of uncontrollable extraneous variables (noise), medium effects are thought to be large enough to be visible to the naked eye and large effects are described as grossly perceptible.[70] Descriptive and inferential statistics will be used in relating the ability of the models to identify at-risk students. Finally to further describe the role of formal thought in chemistry performance, a semi-structured interview was used with students of varying formal thought backgrounds employing a think-aloud approach to solving a set of chemistry problems.

www.manaraa.com

The first step in identifying at-risk students is to classify what would constitute an at-risk student. By consideration of the conventional assessments used in class, where C is meant to denote an average performance, it appears that at-risk students would be those described as below average. The decision was made to consider students who score on the bottom 30% of the sample on the ACS exam to be considered at-risk. In order to do this, the ACS exam was assumed to approximate a normal distribution making the bottom 30% those that scored less than 0.525 standard deviations below the mean, which is equal to 43.3% correct on the exam. In the sample described, 802 out of the 2275 students (35.3%) scored below this cut-off.

Results and Discussions

RQ1. Which predictor, SAT or a formal thought measure, is better able to identify at-risk students?

First the extent TOLT and SAT sub-scores have a linear relationship with academic performance, was determined by running correlations with the academic performance. The results from these correlations are presented in Table 4.2.

Table 4.2 – Comparison of Correlation Coefficients

	TOLT	VSAT	MSAT	ACS Exam
TOLT	---			
VSAT	0.492	---		
MSAT	0.654	0.625	---	
ACS Exam	0.510	0.527	0.608	---

* all coefficients $p < 0.001$

The presence of significant positive correlation coefficients is indicative of a relationship with academic performance among all the predictors. Correlation coefficients also provide an indication of the strength of relationship between the predictors and the

www.manaraa.com

outcome variables. Using Cohen's effect size operation, each of the predictors features a medium effect size with the outcome variable, and a medium effect size between each predictor. Thus each predictor is believed to be a reasonable construct in explaining ACS exam score, and consideration will need to be given to the possibility of the predictors overlapping.

In order to determine the ability of the predictors to identify at-risk students two linear regression models were used. The first model relates TOLT to students' scores on the ACS exam, and the second relates the SAT sub-scores to the ACS exam. The combination of both SAT sub-scores in one model was chosen to represent the practical option for data available to instructors. The results from the two regression models lead to the following equations:

TOLT model

$$\text{ACS Exam} = 29.936 + 3.245 * \text{TOLT}$$

$$R^2 = 0.260$$

SAT model

$$\text{ACS Exam} = -25.196 + 0.04864 * \text{VSAT} + 0.09107 * \text{MSAT}$$

$$R^2 = 0.405$$

By each model, it is possible to depict which students would be identified as at-risk by each set of predictors. For the TOLT model, TOLT scores of 4 or less are predicted to be below the cut-off. By this criterion, the TOLT model identifies 471 students in the sample to be at-risk. Of those 471 students, 332 students had an actual ACS exam score below the cut-off, indicating 70.5% of those predicted were correctly classified. Of the 139 incorrectly classified, 75 scored below average on the ACS exam.

www.manaraa.com

The SAT model could predict students to be below the cut-off via a variety of SAT score combinations, so no single set of SAT cut-offs can be established. However, in general, scoring below 500 on both the math and verbal portion would qualify as an at-risk student in this context. (Note: scoring above 500 on one sub-score could be off-set by a lower score on the other sub-score, a scenario that would still lead to the at-risk classification). This criteria led to a classification of 451 students as at-risk based on the combination of SAT sub-scores, slightly lower than the number of students TOLT predicted. 327 of the 451 students were correctly classified, a 72.5% success rate, a rate slightly higher than the TOLT model. Of the 124 incorrectly classified in this group, 69 of them scored below average. The performance of each model is summarized in Table 4.3.

Table 4.3 – The Model Predictions: At-risk Students

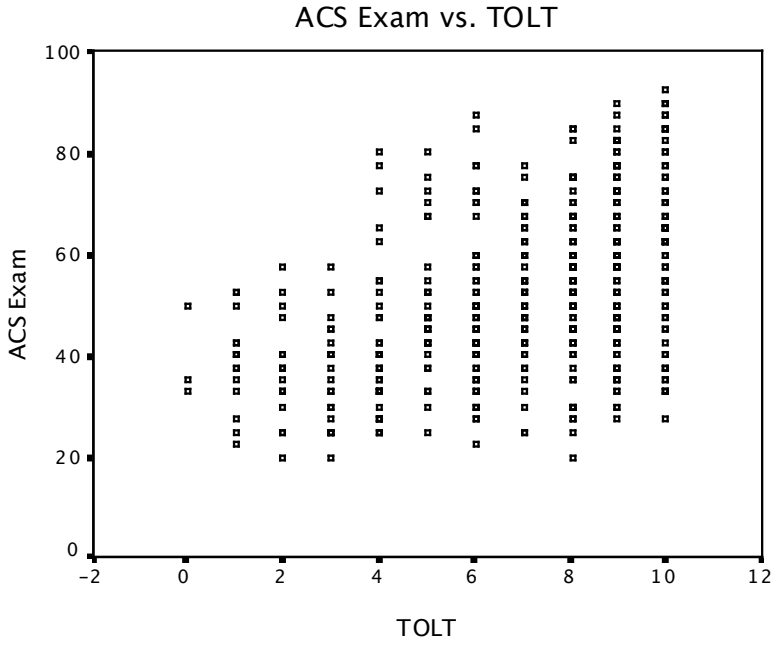
	Predicted At-risk	Actually At-risk	% correct predictions
TOLT model	471	332	70.5%
SAT model	451	327	72.5%

The similar success rates in identifying at-risk students is curious, given the lower R^2 of the TOLT model compared to the SAT model. As a measure of goodness of fit, it is expected that the higher the R^2 value, the better success would be available at predicting scores. This expectation would hold true if predictions for the entire sample were considered. However, by looking at only the at-risk students, a subset of the sample is being examined. How TOLT is able to better identify at-risk students may be better understood by the following scatter-plots, where ACS exam is shown based on TOLT and on Math SAT score (Figures 4.1 and 4.2). Because of the large number of data

www.manaraa.com

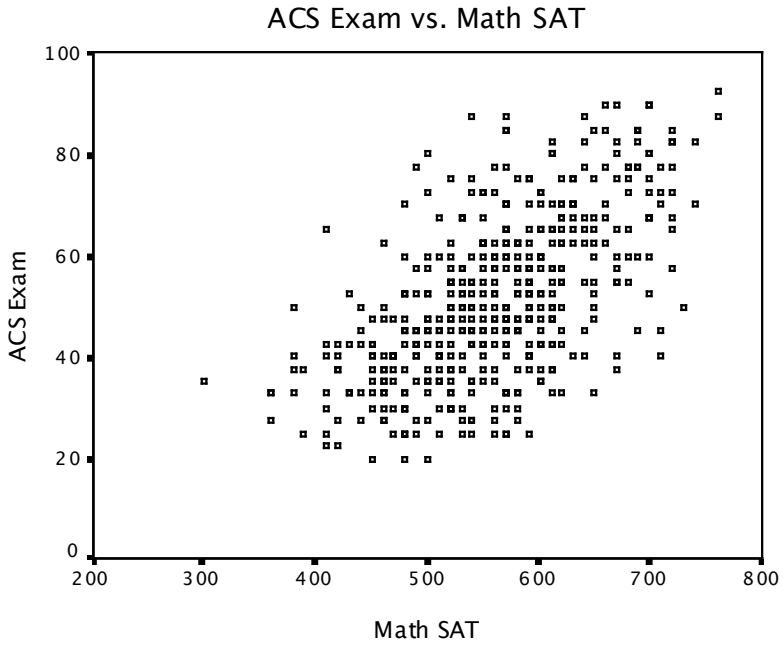
points, a 20% random sample of the data is used in the following plots. Additionally, the R^2 for the TOLT model is somewhat under valued compared to the SAT sub-scores due to the restriction of range, as the TOLT has only 10 possible values but the combination of SAT sub-scores has a distribution which is approaching continuous by comparison.[137]

Figure 4.1 – Relation between TOLT and ACS Exam Scores



www.manaraa.com

Figure 4.2 – Relation between Math SAT and ACS Exam Scores



Among the things to take note is the distribution in each plot. In the TOLT plot, the variability of scores is low for the low TOLT scores, but the scores span almost the entire range for the high TOLT scores. This broad distribution at the high end is the likely cause for the lower R^2 with TOLT, as compared to the SAT model. In the math SAT distribution, a more linear trend is observed: high math SAT scores correspond to higher ACS exam scores, while lower math SAT scores correspond to lower ACS exam scores, which would lead to a higher R^2. (Verbal SAT has a similar distribution, as does a combination of the two SAT sub-scores using the weighting found in the regression model). For this reason, SAT would be better suited for identifying successful students than TOLT, while the at-risk students in this sample are comparably identified by each model. This is consistent with theory, as previously discussed formal thought is one of a series of factors necessary for a successful performance. Other factors such as familiarity

www.manaraa.com

with course content and motivational and affective issues would also play a role in student success. And since SAT scores better identify success, it is suggested that the more general measure, SAT scores, tends to correlate better than TOLT with such factors, in particular those related to familiarity with content.

RQ2. Are the at-risk students identified by each predictor a distinct group, which may lead to more specific interventions geared toward each group of students?

Since all three predictors feature strong correlations, it becomes necessary to examine to what extent all three share in measuring similar qualities, or if the overlaps between each pair are distinct. With at-risk students it may be hypothesized that poor performance on any of these measures is indicative of poor performance on all measures. This is supported by the above measures indicating that the predictors are strongly related. However, this turns out to hold true for approximately half of the cases predicted at-risk: of the 471 students predicted by TOLT and the 451 students predicted by the SAT model, only 266 of the students were classified by both models.

By narrowing the focus to the number of correct predictions, there are three exclusive categories of at risk which students can be classified as: at-risk by only the TOLT model, at-risk by only the SAT model and at-risk by both models. Table 3.4 shows the number of students that fall into each category, and the resulting performance on the ACS exam for each category.

www.manaraa.com

Table 4.4 – The Overlap between Models

Model Predictions of At-risk Status	Only TOLT At-risk (n=205)	Only SAT At-risk (n=185)	Both models At-risk (n=266)
Correct (scored Below Cut-off)	113	108	219
Incorrect (scored Above Cut-off)	92	77	47
Percent of Correct Predictions	55.1%	58.4%	82.3%

From Table 4.4 it appears that each model, TOLT and SAT, describes a unique trait that hinders success in chemistry. There is a distinct group of 113 students that performed poorly on the TOLT and on the ACS final exam while performing satisfactorily on the SAT measure. A similar situation occurs for 108 students who performed poorly on the SAT and on the ACS final exam while performing reasonably well on the TOLT. These two cases demonstrate that the two models identify different groups of students as being at-risk, even though 219 students were correctly predicted by both models to be at-risk. It should also be noted that neither of the models identifies all students who are at-risk: out of the 1619 students not predicted to be at-risk by either model, only 1257 (77.6%) in fact performed above the cut-off.

Statistical comparisons between percent correct predictions employed an arcsine transformation to stabilize variances.[70] The highest percent correct is for those who would be classified as at-risk by both the TOLT model and by the SAT model, demonstrating that a combination of low scores on both measures leads to a greater chance of students performing poorly on the ACS final exam. The differences in correct prediction rate between this 'both' category and each of the two 'only' categories were significant with a medium effect size. No evidence supporting a significant difference in

www.manaraa.com

percent correct between the only TOLT category and the only SAT category was found, indicating that neither model isolates a distinct group of at-risk students better than the other.

RQ3. Can a combination of SAT and formal thought measures provide an advantage in identifying at-risk students?

The previous discussion has shown that, if both the TOLT model and the SAT model predict a student to be at-risk, that is very likely to be the case. However, this post-hoc combination of the predictions of two different models may be too conservative, identifying only a relatively small number of at-risk students. It may be possible to construct a single model using both sets of predictors that will retain a high success rate and identify a larger number of at-risk students. To investigate this possibility, a model was constructed to use both SAT sub-scores and TOLT scores:

SAT & TOLT model

$$\text{ACS Exam} = -19.477 + 0.04410 * \text{VSAT} + 0.07253 * \text{MSAT} + 1.044 * \text{TOLT}$$

$$R^2 = 0.420$$

It was found that each predictor entered the model significantly indicating that even when controlling for the variability that comes from the other predictors, both SAT sub-scores and TOLT still feature a significant relation with the ACS exam, which is consistent with the interpretation that TOLT and SAT map onto performance in distinct ways. Similar to the SAT model, the model that combines both SAT sub-scores and TOLT scores, has many combinations of predictor scores that would result in an at-risk prediction. This model predicted 489 students to be at-risk, higher than the 266 found by

www.manaraa.com

the overlap of the two individual models. Of those 489 students, 354 scored below the ACS cut-off and were correctly classified. This leads to a success rate of 72.4%, which is only slightly higher than the 70.5% seen with the TOLT model, and nearly equivalent to the 72.5% rate of the SAT model. Of the 354 students correctly classified by this model, 351 had been identified by one of the two previous models. Thus the combination of both predictors in a regression model fails to provide an improved model for identifying at-risk students, since only three additional students were correctly found by combining the two sets of predictors. Of the 135 misclassified, 71 scored below average on the exam.

RQ4. To what extent are all at-risk students identified by this set of predictors?

Of the 2275 students, 802 students finished the course below the ACS cut-off. Of these 802 students, 443 students (55.2%) were identifiable based on scores from either TOLT, SAT or a combination of the two. Thus a sizable portion of the students that performed poorly on the ACS exam was not identifiable by these models. This finding may be representative of a need to include other types of predictors, such as affective measures like motivation or confidence, if the goal is to predict all at-risk students.

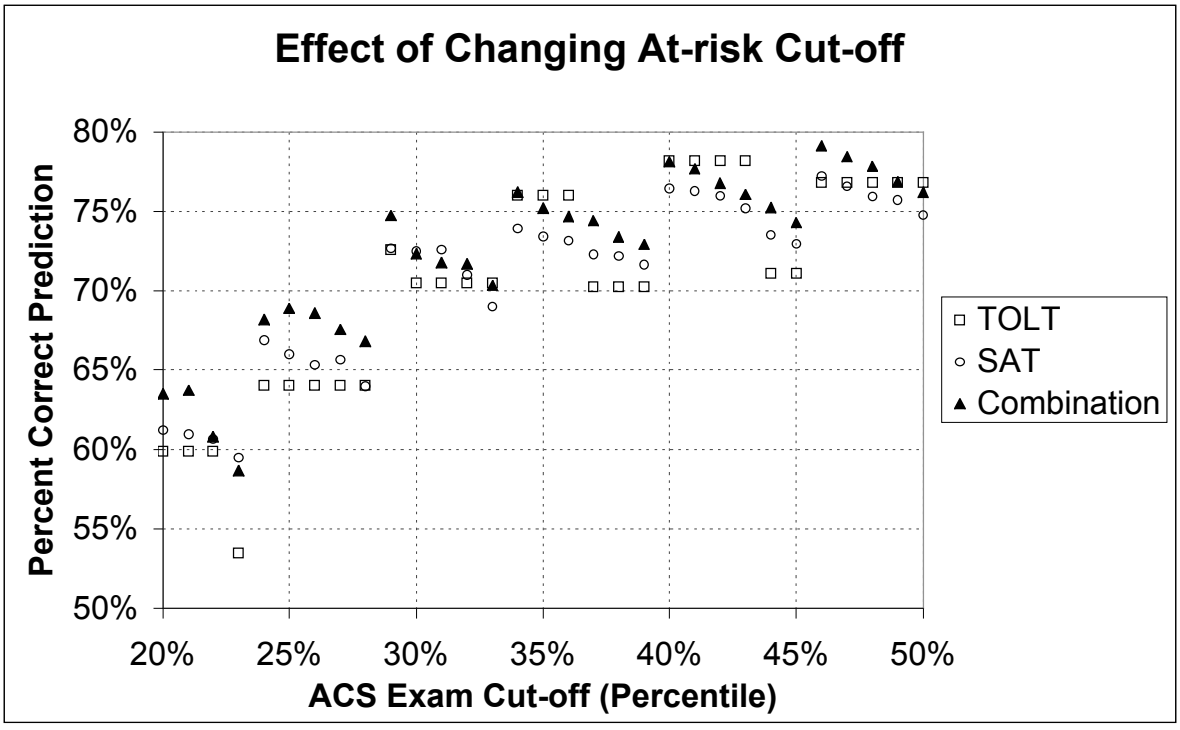
The At-risk Cut-off Decision

It is recognized that the decision to employ the cut-off at the bottom 30% of the sample is somewhat arbitrary, as other values such as the bottom 25% or bottom 33% could reasonably suffice. To address these concerns and to understand the impact of this decision on the conclusions reached, a SAS program was developed to calculate the

www.manaraa.com

percent correct predictions for each model as the cut-off point is changed. The results have been plotted in Figure 4.3. Among the things to note from the plot, first the models switch places depending on the cut-off decision, but all of them remain relatively close together, so that no model offers a distinct advantage over the others in terms of accuracy of identify at-risk students. Also note the general upward trend of percent correct predictions as the cut-off decision increases. This can be attributed to chance guessing. For example, if the cut-off is placed at 20%, randomly selecting students would get 20% correct prediction in identifying at-risk students. However, if the cut-off was 40%, there would be a 40% chance of identifying at-risk students by random selection. In general, each model stays approximately 30 - 40% above the random selection method of identifying students.

Figure 4.3 – Effect of Changing At-risk Cut-off



www.manaraa.com

Missing Data: Students without SAT scores

As mentioned, cases for which SAT scores were unavailable were omitted for comparative purposes with the formal reasoning measure. This presents some interesting implications for the study. A chief concern with missing data is the presence of a trend in those students who have missing data, because the presence of any such trend represents a limitation in the generalizability. By omitting students without available SAT scores, it is necessary to check if the group omitted differs from the group studied. If so, then the applicability of the analysis to those omitted may be questionable. Table 4.5 presents the results from this comparison

Table 4.5 – Comparison of Those with SAT Scores to Those Without

	Avg. score for those with SAT (n, st dev)	Avg. score for those without SAT (n, st dev)	t-test	p-value	d-value
TOLT	6.65 (2957, 2.656)	6.24 (841, 2.645)	3.934	0.000	0.155
ACS Exam	52.11 (2284, 16.724)	49.92 (587, 16.178)	2.839	0.005	0.133

The students without SAT scores scored significantly lower on both the TOLT measure and the ACS exam measure than students with SAT scores. The d-value is the effect size for comparing two means, with both values representing a small effect. Because of the differences between students with SAT scores and those without, the students without SAT scores likely represent a non-randomly selected population. For this reason these students are examined separately in terms of the conclusions presented so far.

There were 841 students in the original sample without SAT scores, and 587 of those took the ACS exam. While no claim can be made regarding SAT for these

www.manaraa.com

students, the role of TOLT in identifying at-risk students can still be investigated. To do this, a new regression equation was fitted for just these 587 students.

ACS Exam = 31.910 + 2.799 * TOLT

$R^2 = 0.201$

As the previous TOLT model, this model also indicates a positive linear relationship between TOLT and ACS exam scores. In addition, the error associated with the TOLT coefficient (0.230) and the intercept (1.599) provide a large enough range to include the original model relating TOLT to ACS Exam. It appears the conclusions reached regarding the previous TOLT model also apply to the students without SAT scores. Of the 587 students, 150 scored at or below a 4 on the TOLT, which was the cut-off for both this model and the previous one. Of the 150 students, 101 scored below the cut-off, for a 67.3% success rate in classification. Of the 49 incorrectly classified, 20 scored below the average score. Nothing discongruent with the previous findings regarding the utility of TOLT was found, and the conclusion that TOLT as a formal thought measure represents a hindrance to the success of chemistry students holds true for those in the sample without SAT scores.

Missing Data: Students Not Completing the Course

As mentioned earlier, those who did not finish the course represent a significant portion of the at-risk student population. However, while leaving the course may be a function of academic performance during the course, there are also a variety of other reasons for such a departure, ranging from personal health to financial trouble. For this reason, classifying all students who did not finish the course as at-risk students is

www.manaraa.com

unsatisfactory. However, given the nature of the conclusions reached regarding the ability of TOLT to predict performance, it will only be necessary to examine those whose TOLT scores fell at or below 4, to determine if those students tended to leave the course for academic reasons. This will be approximated by reviewing students' scores on four instructor generated, multiple-choice in-course tests, in comparison to the class performance on the same test.

Of the 3798 students in this study, 927 students (24.4%) did not take the ACS exam. Of those 927 students, 263 students (28.4%) scored at or below a 4 on the TOLT, which was the criterion previously used for at-risk classification. Forty-two of the 263 students did not take any of the tests, so there decision to drop the course came relatively early, and unfortunately, little else can be said of them. However, of the remaining 221 students, 194 students scored in the bottom 30% within their class on every test they took. Of the remaining 27 students, 22 scored above this mark only once. While it is not possible to extrapolate an exact reason for leaving a course from the data available, and indeed the decision is likely attributable to a number of factors, the data does indicate that academic performance probably played a role in the decision.

Implications

The findings in this chapter are threefold:

1) formal thought (as measured by TOLT) and general achievement (as measured by SAT) represent separate and distinct factors, each of which can be used to identify at-risk students

www.manaraa.com

2) neither the formal thought measure (TOLT) nor the general achievement measure (SAT) is clearly superior in terms of percent correct identification of at-risk students
3) the ease of administration of the TOLT makes it possible to use this measure to identify additional at-risk students for whom SAT scores are not available

The fact that both general achievement and formal thought represent distinct factors in this study is important. It seems there are at least three groups of at-risk students: those who do not have an appropriate knowledge of mathematics and language for success in chemistry, those who do not have the requisite reasoning skills for success in chemistry, and those who lack both. Since this result shows that mathematics achievement and reasoning ability represent different barriers to success, effective remediation will incorporate a review of relevant mathematical and verbal skills as well as the opportunity to work on developing formal thought ability. A remedial course aimed solely at reviewing fundamental mathematical rules in the abstract (e.g. how to isolate variables in an equation, how to manipulate logarithms) for example, is unlikely to be successful for all students. This approach would not help those who need to work on the development of formal reasoning skills – typically by connecting mathematical manipulations with concrete objects first, before abstracting these manipulations into general rules.

Indeed, identifying formal thought as a unique factor for success in college chemistry has several implications for teaching and research. First, the notion that certain chemistry concepts require formal thought is supported.[138] However, it has also been

www.manaraa.com

pointed out that a majority of even the concrete concepts in sciences are presented in a way that requires formal thought.[92] For example, chemistry lectures with few graphics, animations, or demonstrations require students to create their own mental models of concepts in chemistry, a skill associated with formal rather than concrete thinking. Certainly the ability to create mental models is essential for the practice of science, but lectures assuming this ability do little to help students develop it, and the conceptual underpinnings of relatively abstract lecture presentations remain inaccessible to students with low formal thought ability. Taking advantage of the wide array of animations available for presenting major concepts in chemistry is perhaps the simplest way to scaffold these learners in the college chemistry lecture setting.[139, 140] Researchers have also recommended the use of active learning practices to avoid over-dependence on lectures.[102, 141] Tien *et al* [142] and the PLGI reform [33] provide examples of effective active learning reforms that de-emphasize lectures without moving completely from the lecture format. In each of these cases, both quantitative and qualitative investigations of the effects of the reform on different groups of students are needed to provide insight into whether and how these reforms assist those who need to develop formal thinking skills as well as chemistry knowledge.

How should these studies be conducted? The relationship between formal thought as measured by TOLT and chemistry performance (displayed graphically in Figure 4.1) is important: TOLT is a better predictor of at-risk students than of successful students. Therefore, studies that investigate a relationship between TOLT and academic performance through the use of linear regression or correlation [143] may be underestimating the importance of formal thought. A large amount of variation in

www.manaraa.com

academic performance for students with high TOLT scores, while congruent with cognitive development theory, would lead to a reduced proportion of variance explained by TOLT as compared with other predictors of performance. In other words, researchers may be misled into thinking formal thought is not relevant for a given situation, when, in fact, the association of low formal thought ability with poor performance is masked by the large variability in performance for those at the higher end of the TOLT. A suggestion for researchers who are considering such models is to dichotomize TOLT scores, creating a low TOLT score and high TOLT score classification, an approach that is better aligned with the theory of developmental stages. Another option would be to consider students with low TOLT scores as a unique subset of students. This latter option should be of particular use when evaluating whether pedagogical reforms are able to help different groups of at-risk students.

Even though this study focuses on college-level general chemistry, it is also worthwhile to consider broader teaching implications. Longitudinal work from Novak has shown that complex science instruction among elementary age students can show improved understanding at the high school level on similar concepts, far removed from the intervention.[144] Therefore initiatives to improve formal thought ability could also be instituted earlier in the educational stream, with the strong possibility for improving the trends witnessed here at the college level. One such initiative, Shayer and Adey's Cognitive Acceleration program, [89] has shown promising results in promoting cognitive development among middle school students.

www.manaraa.com

Conclusions

In this study, formal thought has been found to have a unique relationship to chemistry achievement apart from SAT sub-scores, even though the two constructs share a medium-sized correlation. Low formal thought ability impedes success in chemistry as much as low SAT sub-scores, and formal thought has been shown to represent a necessary factor for success in college-level general chemistry for a distinct group of students. Recommendations for remediation and for future research were discussed in light of these findings. It is important to note that, while both measures used in the study had reasonable success at identifying students at-risk of performing poorly in college-level general chemistry, there was an additional group of students whose poor performance was not predicted by either measure. Therefore, factors that are unaddressed in this paper are also likely to play a role in success in chemistry. Research into affective aspects of chemistry learning with specific emphasis on at-risk students would complement the cognitive approach taken in this paper. In particular, identifying those affective components which prevent students from achieving success despite high cognitive abilities may help identify other distinct groups of at-risk students and lead to the development of targeted remedies for these groups. As a follow-up study, student interviews were conducted to describe specifically the barriers students with low formal thought scores may experience in chemistry problem solving.

Follow Up Study: Role of Formal Thought in Chemistry Problem Solving

To investigate what specific roles formal thought may have in chemistry problem solving, student interviews were conducted. Thirty students, stratified to represent

www.manaraa.com

different formal thought abilities, were invited to participate. Nine students agreed to participate in an interview. Because the interview falls outside of the normal classroom procedure informed consent was obtained from each participant. See Appendix B for a copy of the informed consent form and the IRB authorization for the student interviews.

Each student was interviewed individually. During the interview a student was given a series of chemistry problems, one at a time, and asked to solve each of them using a 'think aloud' approach.[145] The chemistry problems were chosen among four problems available, which were adapted from the chemistry textbook that was used in the course, [146] with an intent to cover a wide range of topics, stoichiometry, thermodynamics and properties of gases; and focus on both conceptual and algorithmic understandings. Each interviewee was given at least three of the four problems available, but often stopped at three due to time constraints. The interviews took place during the last two weeks of class in Spring 2005, lasted less than 30 minutes, and were audio-recorded and transcribed for analysis.

Three of the interviewees, Lucy, Marcie and Charlie with TOLT scores of 3 or less, were considered low formal thought ability. Two interviewees, Sally and Patty had TOLT scores of 4, which though identified as at-risk in the sample, were considered borderline in formal thought ability and will be evaluated separately. This coincides with the discussion about ACS cut-off, if the cut-off had been changed, there would be similar results but a different TOLT cut-off. The final four interviewees, Margaret, Gina, Martha and Dennis had TOLT scores of 6 or above and were considered high formal thought. To mirror the ACS exam findings, student responses from one algorithmic question and one conceptual question are presented.

www.manaraa.com

Stoichiometry Problem

The algorithmic question was a stoichiometry problem.

For the reaction

$Fe_2O_3\ (s)\ +\ \ CO\ (g)\ \rightarrow\ \ Fe\ (s)\ +\ \ CO_2\ (g)$

How many grams of CO are needed to react with 3.02 g of Fe_2O_3?

Students were given a calculator and the periodic table. Two students in the low formal thought group, Lucy and Marcie, tried this problem. Neither could correctly balance the chemical equation. Marcie employed a guess-and-check method, and did not reach the correct coefficients:

Interviewer: Okay. Let me ask you another question, is this equation balanced?

Marcie: Oh. No.

Interviewer: Okay, would that change your answer?

Marcie: Yea. Okay, so we need 2 Fe, 3, 4 oxygen… it's gonna be a 2 here, carbon will have 2, that's 3, 4, 5. O 1, 2 and 4. 2 here, give this 4, and 2 carbon, and… 5 over here.

Marcie: So it seems to be.

Marcie: It seems to be 5 in here. I don't know.

She was aware that her solution was incorrect, but could not correct it. Lucy could not simultaneously balance the carbon and oxygen in the equation, a process that requires the manipulation of two independent variables:

Lucy: Alright, first you have to balance the equation. [inaudible] 2 carbons on that side, two. Two, three, four so that would be four oxygen, two carbons, two. Alright, so two, four, carbon to carbon we have four… maybe it'll work out. Two carbon, two carbon, two oxygens, four oxygens, that's five though.

Lucy: I don't know. I'm supposed to balance the equation.

Interviewer: Okay, which part isn't balanced?

www.manaraa.com

Lucy: The carbons. I mean oxygens. Because you have four on this side and then five on this side.

Interviewer: Okay, which numbers can you change with oxygen?

Lucy: Which numbers can I change? What do you mean?

Interviewer: What can you do to affect the amount of oxygen on each side?

Lucy: Multiply by two.

Interviewer: Multiply what by two?

Lucy: I know when you have a higher number you're supposed to divide by.

Both students could do the necessary gram-to-mole and mole-to-mole conversions, and with a balanced reaction would likely find the correct answer.

In the borderline group, both Sally and Patty were given this problem. Sally was not sure about the gram to mole conversion, or where to begin, but was able to balance the equation. While there is no direct evidence, her initial attempt placed a 2 in front of both CO and CO_2, and she changed both coefficients to the correct 3, indicating an iterative approach toward balancing the equation. This is in contrast to Marcie's guess-and-check approach where values for the coefficients were changed independent of each other and evaluated, a process that is less likely to reach a correct answer. Patty reached an incorrect set of coefficients, only balancing the Fe_2O_3 and Fe and indicated it was correct. Like Sally, Patty wasn't sure of the gram to mole conversion, and neither had the correct gram to mole conversions in place to solve the problem.

Among the group with the higher formal though scores, Margaret, Gina and Martha were given this problem. Both Margaret and Gina had difficulty with the gram to mole conversions.

Interviewer: Okay, what're you thinking?

www.manaraa.com

Gina: Always need the chapter for this. That's the molar mass, 28.01 grams per mole. And, I always forget how to do this problem.

Interviewer: Okay, how do you think you'd start?

Gina: Well, do they, umm, find out what the molar masses of each compound, element, are. Get that. I know how to find out how many moles, like, are in that, get the molar mass, divide that by 3.02 is how many moles of them there are. Then I forget how you go from there.

Interviewer: Do you want to go ahead and start, and find out how many moles of Fe2O3 there are?

Gina: Yea.

Gina: Okay, you get that, there's two Fe and three oxygens, so you just times molar mass of it by the values and you get 159. Take the 3.02 grams divide it by total which is 159, you get .019 moles of Fe2O3.

Gina: How many grams? I always forget, easy problem but…

Interviewer: Why do you think it's easy?

Gina: Well it's just one of the simple problems we learn first. In the beginning of the, course.

Interviewer: Do you have any guess on what to do next?

Gina: No, I'm usually better if I have a guide in front of me. About what the steps are.

Errors include using the coefficients in calculating molecular weight and multiplying instead of dividing terms. Both also had difficulty with balancing the equation. Margaret stated her balance equation was incorrect, and that it was the C that was not balanced, but had no path to proceed. Gina, like Patty, reached a set of incorrect coefficients and believed they were correct. In contrast, Martha in this group solved the problem correctly, using the same iterative approach as Sally to balance the equation.

The results from the interviews seem to indicate that the low formal thought students expressly struggle with the balancing equation component. As mentioned, the balancing of this equation requires a control for variable acknowledgement, in that the

www.manaraa.com

coefficient of CO and CO_2 must be the same for the equation to be balanced. For the most part students with higher formal thought also struggled with balancing the equation, sometimes in a similar fashion to the low formal thought students. However, two of the five students in the borderline and high groups showed evidence of the more robust iterative approach, manipulating both coefficients in sync, and reaching the balanced equation. Also note this was the only approach demonstrated that was successful in balancing the equation. The low formal thought students seemed more adept at setting up the gram to mole conversion part, and this may be a process that can be more readily learned algorithmically, or by drill. The wide range of ability with this concept among the other groups may be part of the reason for the large variability seen among ACS scores of students with high formal thought (see Figure 4.1). The balancing reaction exercise however is more difficult to drill, with guess and check the primary method demonstrated, though it is often insufficient for more complex chemical reactions.

Gas Molecules Problem

All of the interviewees were given a conceptual question concerning gas molecules. In this problem students were asked:

> Which sample contains more molecules: 1.00 L of O_2 at STP, 1.00 L of air at STP, or 1.00 L of H_2 at STP? Why?

A periodic table and calculator were available and if the student inquired, the average molecular weight of the species in air was given. As before students will be discussed in terms of their formal thought group.

www.manaraa.com

Among the students in first group, with low TOLT scores, there were a variety of responses. Beginning with Marcie:

> Interviewer: Okay, what's the first thing that came to mind?
>
> Marcie: I don't know, I guess, umm, density. To see which sample contains more molecules. It would depend on how dense it was. Density will tell you grams per liter. If you had one gram of O2, that'd be 16, and air, that's more, cause it's a bunch of things, or H2. I guess it would be air, that'd be my guess.
>
> Interviewer: Okay, and why?
>
> Marcie: Because. It would just have tons of molecules because it's so, big.
>
> Interviewer: Just describe to me what you mean by big.
>
> Marcie: Like. It's, there's just so much space I guess. And… cause like air contains a bunch of stuff, so therefore it has the most number, the greatest number of grams per liter. It would have a greater density, which means it would have more molecules.
>
> Interviewer: Okay, you say a bunch of stuff, you mean different gases?
>
> Marcie: Yea. Like Nitrogen and Oxygen that stuff.

Her first indication was that the density of the gases is necessary to solve the problem. Without knowing the gas density, she stated that air would be the most dense, because it "contains a bunch of stuff", when asked to clarify, "Nitrogen and Oxygen that stuff". Because she believed it had the greater density, she also believed it would have the most molecules. Charlie used the ideal gas law for the oxygen sample to solve for the number of moles, and then the number of molecules. Upon finding the number of molecules, he stated "if I do the rest of them the same way, I'm going to get the same number, so I don't think I did this right." When asked why, he indicated that molecular weight must play a role in it somewhere. Despite indications from the ideal gas law, he chose Oxygen as having the most molecules, with the reason being that "it has air in it and Oxygen and air." In contrast, Lucy initially answered that she "just did this like four hours ago" and

www.manaraa.com

used the ideal gas law and reached the conclusion that "they all contain the same amount because they are all at standard pressure." When asked to explain her answer, she created a general rule that all gases at standard temperature and pressure would contain the same number of molecules. After a prompt, she added volume as an important factor, and indicated that gases at standard temperature and pressure with the same volume would all have the same number of molecules. When asked to apply this rule to situations of gases with differing volumes, she correctly deduced that the gas with the largest volume would have the most molecules.

In the borderline group, Sally began with Oxygen having the most, because it has the highest molecular weight. When asked how she could test this idea, she cited an incorrect form of the ideal gas law: volume, temperature and pressure multiplied together, and found that they would all be equal in that circumstance. When asked which is correct, she went with the latter conception, that they would all be the same. Patty's initial thought was that formula mass would play a role, and that the lighter ones would have more molecules, because they would take up less space. The interviewer prompted her to consider the ideal gas law, but this did not lead to a definitive answer to the question.

The final group, those with high formal thought scores, had a wide range of answers. Margaret, like Marcie, stated air would have the most molecules because it has H_2O in it, a combination of hydrogen and oxygen. This seems to confuse the terms molecules with atoms, a previously identified misconception that has been seen at the college level.[147] Martha had a similar conception to Marcie as well, indicating that density was the determining factor and "that they all have one liter that the one that's the

www.manaraa.com

heaviest will contain more molecules, and that would be O_2." Gina had the initial answer that the three gases would have the same number of molecules. When asked to explain why she thought this, she discussed diffusion rates, that Hydrogen would diffuse faster, and because it would spread out faster it would have more molecules than the other two within one liter. Dennis indicated that they would all be the same, because they are all STP, with his reason for thinking so, because he remembered his instructor discussing it. He was asked to solve for the number of molecules in one of the samples, but stated that he could not do it without knowing the density of the gas:

> Dennis: If, see um, if we're on a [inaudible] talk about Avogadro's number, we have to change umm, how much the molecular weight divided by it's molar mass, it's molecular mass and times by Avogadro's number. To find out how many molecule are in the substance. Right now, it's telling us STP, so I think it's same.
>
> Interviewer: You think what's the same?
>
> Dennis: All of them. They all contain the same molecules.
>
> Interviewer: And why do you think they are the same?
>
> Dennis: I remember doing this question in homework.
>
> Interviewer: If molecule mass is different, and you said that would matter, right?
>
> Dennis: Yea. Molecular mass. I don't know, if it makes a difference if it's in STP or not. That's what I'm thinking.
>
> Interviewer: Well, they are all at STP, so
>
> Dennis: Yea.
>
> Interviewer: Does that tell you anything?
>
> Interviewer: Let me ask you this, could you solve for how many molecules are in one liter of oxygen?
>
> Dennis: Yea, I forgot the conversion, liters to grams
>
> Interviewer: Okay, that's density,
>
> Dennis: That's density

www.manaraa.com

Interviewer: But we don't know that in this problem.

Dennis: I think that if it's not the same it would be this one. Because H2 the weight is only 1, so if you do the math this one's 16, this one's 32, and this one's about the same,

Interviewer: Air is more nitrogen so it's 29

Dennis: Okay. But this is if you do diffusion and you have to use the rate divided by the molecular weight, the H2 is 2, but the higher you divide by the smaller you get.

Interviewer: That makes sense, let me ask you about, just wondering, given this information, 1 liter of oxygen at STP, can you calculate how many molecules are in that?

Dennis: You said it had something to do with the density so

Interviewer: You do mention density, and I believe you can get it from density but not knowing the density… Is there anyway you can do it without the density?

Dennis: Isn't it density to convert it from liters to grams?

Interviewer: With density, but you don't know that number.

Dennis: Can't think of it off my head. I don't think so.

Table 4.6 summarizes the interview responses for quick reference. Some of the misconceptions discussed are prevalent throughout the groups. For example, the idea that density would determine the number of molecules seems to suggest that since the volume is the same, the higher the mass determines the number of molecules. This implies a 1:1 relationship between mass and number of molecules, disregarding the role of molecular weight in this relationship. This disregard for an underlying factor may be an indication of low formal ability, but both Marcie in the low group and Martha in the high group indicated this misconception. Similarly that one gas could be a combination of other gases and would therefore have more was indicated by Marcie, Charlie and Margaret, also spanning the low and high group. Reliance on factors that play no role in the question such as the combination of gases or the rate of diffusion seem to indicate an

www.manaraa.com

incomplete understanding of the nature of gases but this does not seem to map onto any specific formal thought trait.

The lack of relationship between formal thought groups and problem solving for this problem may indicate one of several possibilities. First, it is possible that the question does not rely on formal thought ability, since it could be solved by the ideal gas law and algebraic manipulation. Lucy, in the low formal thought group was the only student to rely completely on this fact in constructing her answer. Similar to the gram-mole conversions discussed previously, the extent this problem depends on formal operations could be drilled. As described Lucy did mention having seen this particular problem before.

Second it may be possible that the students being interviewed did not have the necessary background knowledge to solve this problem regardless of their formal thought ability, though when the interviews were conducted the ideal gas law and properties of gases had been covered in the course. In this case students may construct alternative answers such as the misconceptions previously discussed, or they may mimic another source. Note that the two students who settled on the correct answer, Lucy and Dennis, both mentioned another source as providing the answer. Concern about students mimicking facts to answer formal thought questions has been discussed in the literature, [87, 148, 149] and this may be the case for some students with the formal thought measure. In discerning the role of formal thought in applied chemistry questions, in particular with conceptual questions such as this, it seems difficult, yet imperative, to first arrive at a clear understanding of the student's prior chemistry knowledge, as it applies to this question.

www.manaraa.com

Table 4.6 – Interview Responses Organized by Formal Thought Group

Group	Pseudonym	Balancing	Mass-to-mole	Initial Gas Molecules	Final Gas Molecules
Low Formal thought	**Lucy**	Not balance carbon and oxygen at same time	Correct gram to mole; incorrect coefficient	*Use Ideal Gas Law → all same*	*Created rule relating volume to number of molecules*
	Marcie	Guess-and check; did not solve	Correct gram to mole; incorrect coefficient	Density is necessary	Air most molecules, multiple components
	Charlie			Ideal Gas Law → all same; expressed doubts	Oxygen most molecules, multiple components
Border-line	**Sally**	Correct; changed CO and CO_2 simultaneously	Incorrect gram to mole conversion	Oxygen most molecules; highest MW	Incorrect Ideal Gas Law → all same
	Patty	Incorrect coefficients, believed correct	Incorrect gram to mole conversion	Least MW would have the most	No change in response
High Formal thought	**Margaret**	Stated C was not balanced, but could not proceed	Incorrect gram to mole conversion	Air most molecules; multiple components	No change in response
	Gina	Incorrect coefficients, believed correct	Incorrect gram to mole conversion	All same	Hydrogen most molecules, faster diffusion rate
	Martha	*Correct; changed CO and CO_2 simultaneously*	*Correct gram to mole and coefficient*	Density is necessary	Oxygen most molecules; highest MW
	Dennis			All same; instructor mentioned it	Need density to solve for number of molecules

MW = molecular weight *Italics* = fully correct answer

www.manaraa.com

A third possibility is that the abstract modeling required in such a problem is available among all three formal thought groups. Among the low formal thought group, the molecular composition of the gases and the density of the gases were perceived to play a role, each of which require abstract modeling. Moreover, these perceptions match the explanations of the high formal thought group. This may be support for indications in the literature that students with low formal thought may succeed with conceptual questions more than traditional algorithmic questions.[134, 135, 150] Each of the three possibilities presented are tentative, and would require further investigation to support.

Finally, the inability of this question to distinguish between the formal thought groups seems to represent an overall poor performance on this question. As previously stated, we do not necessarily expect students with high formal thought to succeed in the course, so that this still seems to correspond to the current beliefs in the class. In fact, the high formal thought group's average on the ACS Exam was 50% correct, just below the overall average for the entire sample, and the students in this group also showed some difficulties with the stoichiometry problem.

Implications for Future Research Projects

The findings from the interview did not uniformly agree with the expectation that a high score on the formal thought measure is necessary to perform the operations needed to solve chemistry problems.[138] For example, the low formal thought students actually had the most success with the mass to mole calculations in the stoichiometry problem, an operation thought to require proportional reasoning. This unexpected result can guide future research projects. The success of the low formal thought students on the mass to

www.manaraa.com

mole calculations begs the questions, than what operations require formal thought? Though matching nicely to proportional reasoning, the mass to mole conversions may be seen as a starting point in calculation for the general chemistry course, as the interviewee Gina pointed out, and further calculations become more complex. Introducing molarity, for example, adds one step to the calculation. Limiting reagent and thermodynamic calculations, which come later in the course, rely on many more steps. Taking low formal thought students through a series of more complicated calculations to determine areas where they struggle would be an important step in further detailing the role formal thought plays in general chemistry success. One classification scheme for the complexity of calculations may be found in Niaz's work on the role of M-space. Understanding at which point the algebraic operations in chemistry require high formal thought scores may indicate an appropriate target in designing remediation teaching exercises for students who enter the course with low formal thought scores.

The gas molecules problem offers suggestions for future research into the role of formal operations in conceptual understanding. As mentioned, an understanding of the concepts students bring in prior to approaching the problem will be a necessary condition. Some possibilities for doing this may be to have students first relate concepts of gases by drawing chemical species in a gaseous form, similar to Ebenezer's work with ions in solution.[151] Delineating these responses on formal thought ability may provide more evidence of the role of formal thought in conceptual understanding. Another avenue for approaching this problem would be to consider alternative chemical concepts, as the property of gases investigated here faced consistent misconceptions even with the students who scored high on the formal thought measure. A final option may be

www.manaraa.com

providing students with physical samples of gaseous substances that can be manipulated, with a series of tasks. The results from manipulating the materials may help students resolve the cognitive dissonance between the ideal gas law expectations and their existing conceptions. Students of differing formal abilities may explore this cognitive dissonance in alternative ways. To the extent this is successful, the use of real world manipulative may also prove a useful tactic as a remediation teaching exercise.

www.manaraa.com

V: PLGI Impact on Formal Thought Groups

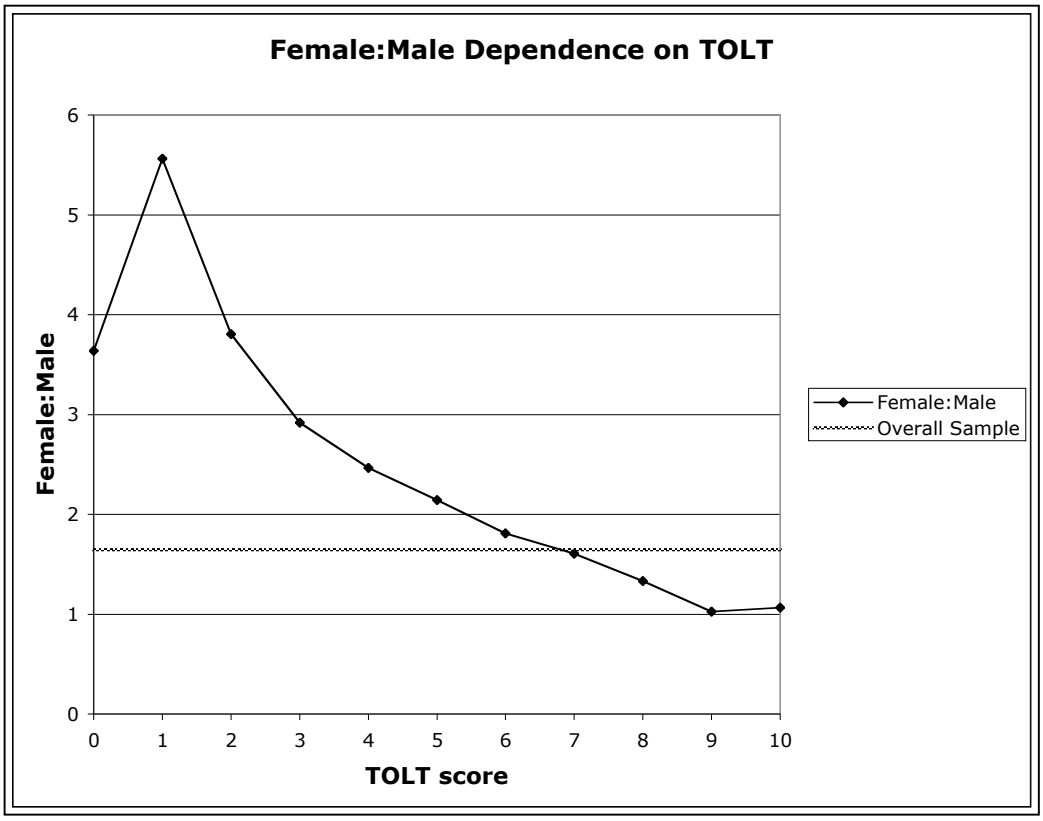
As previously shown, students who score low on a formal thought measure have a tendency to perform poorly in the course, as determined by performance on an external exam. This is believed to be an indication that poor performance on a formal thought measure is an indication of low formal thought ability, and this low ability represents a hindrance to success in the course. Furthermore, this group of students shows a noteworthy departure from those who exhibit low scores on a general achievement measure, SAT sub-scores, indicating that formal thought may be describing a specific set of traits that map onto chemistry performance and these traits are independent of the more general questions and application required in the SAT test.

Because of this trend, efforts made to improve chemistry teaching should explicitly consider the effects of the reform on students with low formal thought ability. Improving the performance of this group of students would have several beneficial implications. First, making chemistry accessible to this group is consistent with educational goals of providing opportunities for all students and as a result reduce the attrition and failure rates currently seen in the course. Second, any improvements witnessed would suggest a means for improving formal thought ability across multiple settings. The reform could be studied, as a follow-up, for the means by which it offers this improvement. When this is well understood, the specific traits could be used in other college classrooms or even in the middle and high school curriculums where efforts are

www.manaraa.com

currently made to improve formal thought development. Finally, improving the performance of low formal thought students in a general chemistry course may offer a specific mechanism for improving the diversity of those who continue in the science or science-related professions. Demographic information obtained from the registrar for a limited portion (N = 3258 students) of the sample described in the previous chapter suggests that students with low formal thought ability in the sample disproportionately occurs with female students and under-represented minorities, as described in Figures 5.1 and 5.2.

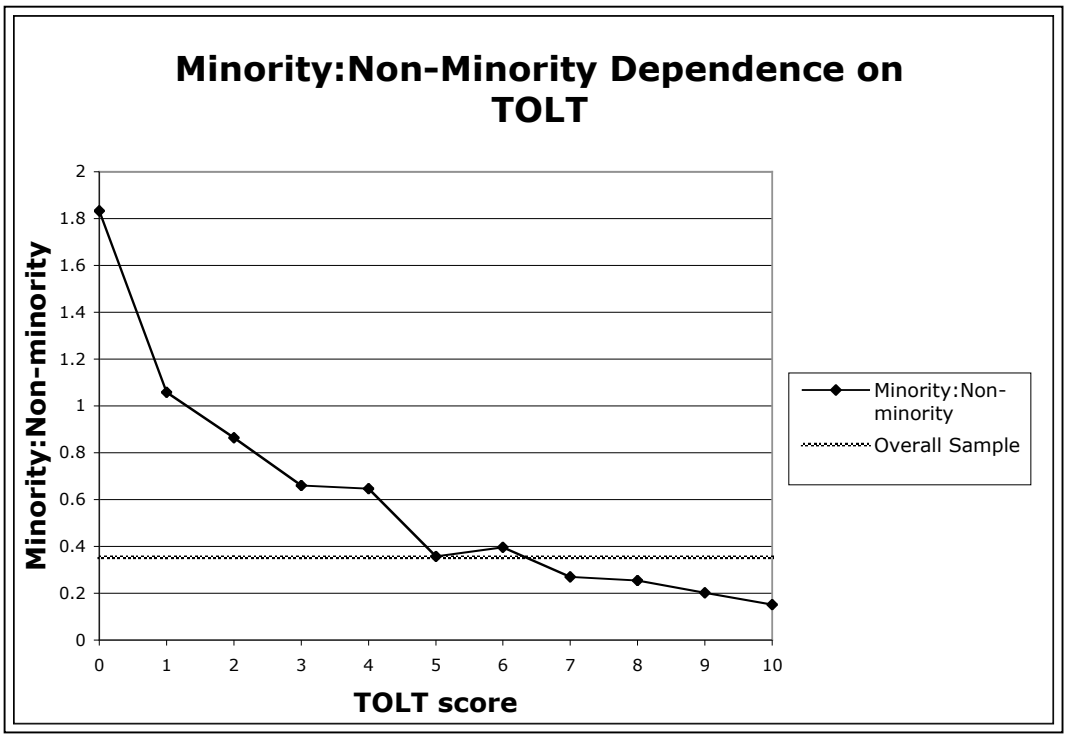
Figure 5.1 – Female:Male Dependence on TOLT Scores



www.manaraa.com

In Figure 5.1, by plotting the ratio of female to male students for each TOLT scores, there is evidence that students with low TOLT scores are disproportionately female. In particular of TOLT scores of 5 or less, where female students are in a 2:1 or greater ratio compared to male students. Overall for the sample a 1.64:1 ratio of female to male students is expected, so that assisting students with low TOLT scores may also improve the retention and course performance of female students in the sample.

Figure 5.2 – Minority:Non-minority Dependence on TOLT



In Figure 5.2, minority refers to the races that have traditionally been classified as under-represented in the sciences by the National Science Foundation, that is African-American and Hispanic students.[152] Non-minority refers to the Asian and White students in the sample. Overall the ratio of minority to non-minority is 0.35:1 in the sample. However, for students with TOLT scores of 4 or less, this ratio is consistently over 0.6:1 and despite

www.manaraa.com

the overall ratio, for very low TOLT scores minority students out-number the non-minority students. In a similar fashion, differences are found between the minority and non-minority students in areas like high school chemistry background, where 31.0% of minority students reported less than one full year of high school chemistry compared to 24.7% of their non-minority counterparts. However assistance for students with low-formal thought in the chemistry curriculum would very likely improve the diversity of those who successfully complete the course and works toward the National Science Education Standards overall objective of "Science for everyone."[7] In particular as research has found that formal thought plays a role in student success in science content knowledge even after controlling for English language proficiency.[153]

Steps to improve the performance of students with low formal thought ability may borrow from a large amount of prior research. Some of this research has been discussed previously, for example Lawson and Renner suggests essentially a discrepant event model on certain formal operations and found some benefit.[92] Other suggestions include making abstract concepts more concrete and therefore more accessible to these students, but in the course described this can be exceedingly difficult with some concepts and near impossible with others (e.g. kinetic molecular theory, molecular orbital theory).

The PLGI reform, as previously discussed, borrows on a variant of cognitive development first proposed by Vygotsky, termed the Zone of Proximal Development. Vygotsky believed that a person's formal thought ability may be at a certain point, but the person's formal thought potential was some place immediately beyond their current ability.[13] Teaching to this potential, he suggested, was the best way to promote formal thought development. It has been argued that traditional course instructors, such as

www.manaraa.com

college professors, may be too far removed from their students to teach at this potential, however peer students may be more apt to do so. In this framework, the students working in cooperative learning groups in the PLGI reform may be privy to teaching at this potential and subsequently benefit. Additionally the use of peer leaders may also contribute to this phenomenon. The intent of this study, then is to investigate if PLGI assists students with low formal thought, and if so to what extent.

Hierarchical Linear Model Construction with Formal Thought Measure

To determine the effects of PLGI on low formal thought students, an HLM analysis was conducted in a similar manner as described in Chapter 3, while incorporating the formal thought measure introduced in Chapter 4. The initial equation for the Level 1 model is:

$$Y_{ij} = \beta_{0j} + \beta_{1j}VSAT + \beta_{2j}MSAT + \beta_{3j}TOLT4 + r_{ij} \qquad (1)$$

where Y_{ij} is an outcome measure for student i in classroom j, VSAT is a student's verbal SAT score, MSAT is a student's math SAT score, TOLT4 is a dichotomous variable where students scoring above 4 on the TOLT are given a 1 and equal to or less then 4 are coded 0. r_{ij} is an error term to describe the unique effect of each student. For the Level 2 model:

www.manaraa.com

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}SATavg + \gamma_{02}TOLT4avg + \gamma_{03}PLGI + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}SATavg + \gamma_{12}TOLT4avg + \gamma_{13}PLGI + u_{1j} \\
\beta_{2j} &= \gamma_{20} + \gamma_{21}SATavg + \gamma_{22}TOLT4avg + \gamma_{23}PLGI + u_{2j} \\
\beta_{3j} &= \gamma_{30} + \gamma_{31}SATavg + \gamma_{32}TOLT4avg + \gamma_{33}PLGI + u_{3j}
\end{aligned} \qquad (2)$$

that describes the classroom effects on performance. In this Level 2 model three variables characterize the classroom: SATavg is a class's average SAT score, TOLT4avg is a class's average score on the TOLT4 variable, and PLGI is a dichotomous variable describing if a class experienced PLGI (PLGI = 1) or did not (PLGI = 0).

Results and Discussions

Initial run of this model using the ACS exam as the outcome variable, suggested that SATavg and TOLT4avg, the classroom variables, do not have a significant relation to the outcome variable. The lack of relation with SATavg is in contrast to the ACS Exam model from Chapter 3, where SATavg had a significant relation to both the intercept coefficient and the Math SAT slope coefficient. This discrepancy is likely a result of the class level TOLT4avg variable overlapping with the SATavg variable to reduce the effect witnessed. It is also possible that the student level TOLT variable is responsible for the overlap, but unlikely given the dichotomous nature of this variable. Removing the class level variables SATavg and TOLT4avg and re-running the model produced the results listed in Tables 5.1 through 5.4.

Table 5.1 - Estimating the Intercept Coefficient (β_{0j})

Symbol	Description	Estimate	Std. error	Sig.
γ_{00}	Intercept	20.0805	0.4577	<0.001
γ_{02}	PLGI	2.7104	1.2021	0.042

www.manaraa.com

Table 5.2 – Estimating the Slope Coefficient (β_{1j}) Relating Student Verbal SAT to ACS Exam

Symbol	Description	Estimate	Std. error	Sig.
γ_{10}	Intercept	0.01590	0.1835	n.s.
γ_{12}	PLGI	0.008001	0.4103	n.s.

n.s. = non significant ($p > 0.050$)

Table 5.3 – Estimating the Slope Coefficient (β_{2j}) Relating Student Math SAT to ACS Exam

Symbol	Description	Estimate	Std. Error	Sig.
γ_{20}	Intercept	0.03347	0.1835	n.s.
γ_{22}	PLGI	0.005317	0.4104	n.s.

n.s. = non significant ($p > 0.050$)

Table 5.4 – Estimating the Slope Coefficient (β_{3j}) Relating Student TOLT to ACS Exam

Symbol	Description	Estimate	Std. Error	Sig.
γ_{20}	Intercept	1.9082	0.4958	<0.001
γ_{22}	PLGI	-2.0916	1.2855	n.s.

n.s. = non significant ($p > 0.050$)

Beginning with the significant parameters, the intercept of 20.0805 indicates the average number of correct responses for a student who is not in PLGI, has both SAT sub-scores equal to the class average, and a TOLT score of less than 4. The coefficient γ_{02} indicates the effect of PLGI on this particular scenario, where an increase of 2.7104 questions correct is similar to the improvement witnessed in Chapter 3. Also significant is the coefficient γ_{20} which indicates that , comparing students of above 4 on the formal thought measure to those at 4 or below there is a1.9082 difference in the outcome measure. Finally, there are a couple other things to note about the model. First, the coefficient γ_{20} indicates that PLGI effectively removes the difference in TOLT scores, or

www.manaraa.com

that in PLGI there is no difference between the two TOLT groups. However, this value is not significant and could be attributed to chance. This will be discussed in further detail in a subsequent section. Second, note that the intercept for the Math and Verbal SAT scores is non-significant. This is the first instance where the SAT sub-scores do not relate to the ACS Exam, regardless of what may be controlled for. This may be a result of lack of statistical power to detect these effects (discussed later) or the possibility that the PLGI variable's inclusion maps closely to SAT sub-scores and thus disguises the impact of SAT sub-scores. To pursue the latter possibility, the model was re-run with the PLGI variable dropped from the SAT sub-scores. The results are present in Table 5.5 through 5.8:

Table 5.5 - Estimating the Intercept Coefficient (β_{0j})

Symbol	Description	Estimate	Std. error	Sig.
γ_{00}	Intercept	20.0831	0.4354	<0.001
γ_{02}	PLGI	2.5547	1.1463	0.044

Table 5.6 – Estimating the Slope Coefficient (β_{1j}) Relating Student Verbal SAT to ACS Exam

Symbol	Description	Estimate	Std. error	Sig.
γ_{10}	Intercept	0.01781	0.005479	0.001

Table 5.7 – Estimating the Slope Coefficient (β_{2j}) Relating Student Math SAT to ACS Exam

Symbol	Description	Estimate	Std. Error	Sig.
γ_{20}	Intercept	0.03415	0.005412	<0.001

www.manaraa.com

Table 5.8 – Estimating the Slope Coefficient (β_{3j}) Relating Student TOLT to ACS Exam

Symbol	Description	Estimate	Std. Error	Sig.
γ_{20}	Intercept	1.9036	0.4789	<0.001
γ_{22}	PLGI	-2.3832	1.2374	0.054

With the removal of the PLGI variables, the estimate of the coefficients remains relatively stable, but the standard error for the SAT sub-score intercepts is markedly reduced. This change supports the possibility that the PLGI variable may hide the impact of the SAT sub-scores, and the results of this model seem to better correspond with the models formed in Chapter 3, which showed a strong relation between SAT sub-scores and the ACS Exam.

The effect of PLGI on TOLT is the primary focus of this study, to determine if PLGI assists students with low formal thought. While the model suggests that it does, and in fact eliminates the gap between the two groups, it also suggests that this can be attributed to chance. Underlying this is the large standard error associated with this term. This large standard error is a function of the degrees of freedom associated with the term. The initial recommendation is the need for a larger sample size. In this sample there were 45 students in the PLGI reform that had a low TOLT score, and of those 45 students, 34 students took the ACS Exam. This small number is a likely cause for the large standard error term. However there are a series of other factors inherent in HLM which may also affect these results:

www.manaraa.com

Additional Considerations in Hierarchical Linear Models

Dichotomization Point

Based on the findings from the previous chapter, the decision was made to dichotomize TOLT in this analysis. This decision certainly impacts the findings presented. However, given the large variability in chemistry performance at the high-end of the TOLT score, some dichotomization of the variable is necessary. However, as noted in the previous chapter, different points at which to dichotomize could be justified. To examine the effect of this decision on this model, the above model was run with varying points of dichotomization and the results are presented in Table 5.9.

Table 5.9 – The Effect of Dichotomization on the TOLT Main Effect and Interaction

Dichotomization point	Sample Size Non-PLGI, PLGI	Main effect of TOLT	Interaction of TOLT*PLGI
2 or less	82, 14	0.2705	-1.1461
3 or less	136, 24	1.2558*	-3.9845*
4 or less	220, 34	1.9036*	-2.3832
5 or less	316, 47	1.6654*	-1.4448

* Effect is significant at $p < 0.05$

From Table 5.9, there is a consistent trend of a positive main effect of TOLT, representing students with high TOLT scores perform better, regardless of the dichotomization point, though when the dichotomization is 2 or less, the effect becomes negligible. Also from Table 5.9, the effect of the PLGI reform on the TOLT difference is consistently negative, indicating the reform is moderating the difference based on TOLT.

Note for the dichotomization point of 3 or less, the interaction term is much larger then the TOLT difference. This does not mean that the reform hindered students with

www.manaraa.com

high TOLT scores, for this model the reform main effect (listed as γ_{02} in the above model) is 4.1127. Students with high TOLT, when experiencing the reform, perform 0.1282 more questions correct (4.1127 – 3.9845) than students with high TOLT that do not experience the reform. Students with low TOLT, though, are expected to score 4.1127 more questions correct than low TOLT students without the reform. This difference is compared to the 2.5547 difference when the dichotomization is at 4 or above listed in Table 4.5. By considering alternative dichotomization points, the interpretation of the fixed effects seems constant, but there still is not compelling evidence to indicate the PLGI*TOLT interaction term cannot be attributed to chance.

Degrees of Freedom Method

The PROC MIXED procedure in SAS uses a specialized algorithm to solve for the degrees of freedom for each coefficient. The degrees of freedom estimate is combined with the standard error and estimate of the value in order to determine if the effect is statistically significant. SAS offers the following methods: contain, residual, between-within, Kenward-Roger and Satterthwaite. Per Singer's suggestion, the between-within method was used throughout all the models presented.[71]

Running the model presented above with each of the alternative degrees of freedom methods shows that the contain and residual methods are more liberal than the between-within method used, in that the likelihood the effect can be attributed to chance (by convention, the p-value) is reduced by these methods, to the point where the PLGI*TOLT interaction effect approaches the 0.05 convention used. Conversely, the

www.manaraa.com

Kenward-Roger and Satterthwaite provides more conservative results, the likelihood the effects can be attributed to chance is increased.

Iteration Method

Hierarchical linear models are often solved using an iteration method to estimate the coefficients and the underlying covariance structure. In SAS Proc Mixed the two iteration methods available are Restricted Maximum Likelihood (REML) and Maximum Likelihood (ML). For the data presented here, neither iteration method provided convergence on the result. Per the recommendation of the SAS user manual, [76] the Minimum Variance Quadratic Unbiased Estimate (MIVQUE0) was used instead, given that the sample size was large and the alternatives failed to converge. The choice of iterative method could affect the resulting estimates and standard errors. And as convergence methods improve, the two iterative approaches may provide estimates with reduced standard error associated with them, improving the power of the significance tests associated with them.

Random Effects

In HLM variables must be specified as random or fixed. Typically, the variables in every level but the last level may be considered random. The decision is based on whether the effect of this variable on the outcome could be expected to vary from one group to another. So, for the instance of Math SAT, would the role of Math SAT on the ACS Exam differ if a student were in a PLGI section or not? If so, then Math SAT should be considered a random effect, if not, Math SAT should be a fixed effect. This is

www.manaraa.com

a decision to be made on a theoretical basis. In this model Math SAT, Verbal SAT and TOLT4 were all set as random effects, since the focus was on PLGI's effects on equity regarding these variables. However, since PLGI failed to relate to either SAT sub-score, this may be an argument for making the SAT sub-scores listed as fixed effects. The decision to keep or remove these variables as random effects would effect the estimates and the standard errors reported.

Power

Finding a p-value that fails to indicate significance, in this case values greater than 0.05 and designated n.s., can indicate one of two possibilities. Either there is no relation between the variable considered and the outcome measure, or there is insufficient evidence to demonstrate this relationship to the extent where it cannot be attributed to chance. With the results presented this is an important distinction, either the non-significant PLGI * TOLT interaction indicates that PLGI has no effect on equity as it relates to TOLT, or there is an insufficient sample size to demonstrate the effect of the PLGI * TOLT interaction. As an interaction term, the power for finding a significant fixed effect can be markedly reduced from the main effects in the same model.[84] Furthermore, this term is more dependent on the level 2 sample size than the level 1 sample size, such that increasing the number of classrooms in the study would provide the most benefit in improving power.

All of the above considerations play a role in the HLM model described and the subsequent interpretation of the results. While an effort has been made to follow authoritative recommendations, this also contrasts with the relative newness of such

www.manaraa.com

models, meaning that little may be known about some decisions in certain scenarios. Additionally, the considerations described above interact with each other. For example, reducing the number of random effects in a model, generally improve the ability of REML and ML to converge which would make these options available. Similarly, the degrees of freedom method may impact convergence. As research into HLM progresses, many of these issues should become clear, but for now it looks promising that the PLGI reform may reduce the achievement gap between low-TOLT and high-TOLT students within this sample.

Time-Series Model with Formal Thought Measure

To further consider the impact of the reform on the TOLT differential, a time-series model was developed in a matter similar to the time-series model introduced in Chapter 3. The changes are the addition of the TOLT4 dichotomous variable (1 = TOLT score above 4, 0 = TOLT score of 4 or less), and the addition of TOLT4AVG to the class level variable. The TOLT4avg is the percent of the class that scored high on the TOLT4 measure. All other variables are identical to the Chapter 4 model, including the centering procedure for SAT sub-scores.

Level 1 – Within Student

$$P_{ijk} = \pi_{0jk} + \pi_{1jk} Time + e_{ijk} \tag{3}$$

Level 2 – Between Student

$$\begin{aligned} \pi_{0jk} &= \beta_{00k} + \beta_{01k} MSAT + \beta_{02k} VSAT + \beta_{03k} TOLT4 + r_{0jk} \\ \pi_{1jk} &= \beta_{10k} + \beta_{11k} MSAT + \beta_{12k} VSAT + \beta_{13k} TOLT4 + r_{1jk} \end{aligned} \tag{4}$$

www.manaraa.com

Level 3 – Between Class

$$\begin{aligned}
\beta_{00k} &= \gamma_{000} + \gamma_{001}SATavg + \gamma_{002}REFORM + \gamma_{003}TOLT4AVG + u_{00k} \\
\beta_{01k} &= \gamma_{010} + \gamma_{011}SATavg + \gamma_{012}REFORM + \gamma_{013}TOLT4AVG + u_{01k} \\
\beta_{02k} &= \gamma_{020} + \gamma_{021}SATavg + \gamma_{022}REFORM + \gamma_{023}TOLT4AVG + u_{02k} \\
\beta_{03k} &= \gamma_{030} + \gamma_{031}SATavg + \gamma_{032}REFORM + \gamma_{033}TOLT4AVG + u_{03k}
\end{aligned} \tag{5}$$

$$\begin{aligned}
\beta_{10k} &= \gamma_{100} + \gamma_{101}SATavg + \gamma_{102}REFORM + \gamma_{103}TOLT4AVG + u_{10k} \\
\beta_{11k} &= \gamma_{110} + \gamma_{111}SATavg + \gamma_{112}REFORM + \gamma_{113}TOLT4AVG + u_{11k} \\
\beta_{12k} &= \gamma_{120} + \gamma_{121}SATavg + \gamma_{122}REFORM + \gamma_{123}TOLT4AVG + u_{12k} \\
\beta_{13k} &= \gamma_{130} + \gamma_{131}SATavg + \gamma_{132}REFORM + \gamma_{133}TOLT4AVG + u_{13k}
\end{aligned}$$

None of the between class variables impacted the effect of SAT sub-scores (β_{01k}, β_{02k}, β_{11k}, β_{12k}) significantly so these variables were considered fixed effects, where the effect of the SAT sub-scores does not depend on the classroom conditions. The between class variables SATavg and TOLT4avg also did not enter the model appreciably, and these were removed to create a more parsimonious model. The results from the simplified model are presented in Tables 5.10 through 5.11:

Table 5.10 – Estimating the Intercept Coefficient (π_{0jk})

Symbol	Description	Estimate	Std. Error	Sig.
γ_{000}	Intercept	58.5281	1.1831	<0.001
γ_{002}	PLGI	-1.9551	2.8588	n.s.
β_{01k}	Math SAT	0.08370	0.007001	<0.001
β_{02k}	Verbal SAT	0.01862	0.006596	0.005
γ_{030}	TOLT4	1.1346	1.3702	n.s.
γ_{032}	PLGI*TOLT4	0.72563	3.1244	n.s.

www.manaraa.com

Table 5.11 – Estimating the Slope Coefficient (π_{1jk})

Symbol	Description	Estimate	Std. Error	Sig.
γ_{100}	Intercept	-3.6189	0.8172	<0.001
γ_{102}	PLGI	2.7730	2.0134	n.s.
β_{11k}	Math SAT	-0.00605	0.004490	n.s.
β_{12k}	Verbal SAT	0.000237	0.004221	n.s.
γ_{130}	TOLT4	1.0652	0.9265	n.s.
γ_{132}	PLGI*TOLT4	-1.5916	2.1635	n.s.

Table 5.10 indicates the effect of each variable on Test 1, or the starting point in the semester. As in the previous time-series model, students' SAT sub-scores have a significant impact on Test 1 scores and the reform has no evidence of a significant impact on Test 1 performance. The TOLT4 estimate of 1.1346 is the difference between high TOLT and low TOLT students on the Test 1 score (in percent correct), after controlling for SAT sub-scores, and this difference can be attributed to chance.

Table 5.11 reflects the changes in course performance over time, and as can be seen in the last column, none of the variables have sufficient evidence to claim statistical significance. However, by comparison with Table 3.10, the estimate of the PLGI reform for example is larger, but the standard error has increased to the point that the estimated value could also be attributed to chance. While not significant, indicating that the effects may be attribute to chance or a lack of statistical power, interpretation of the estimates indicate that in general students score decrease as the semester progresses by the negative intercept, and that this decline is moderated by the PLGI reform. Furthermore, students with high TOLT would experience less of a decline in general. The PLGI reform, like with the ACS exam model, moderates the difference between high TOLT and low TOLT students as well. Consider a student in the reform with high TOLT would expect a decline of 1.3723 for each progressive test (-3.6189 + 2.7730 + 1.0652 – 1.5916 = -

www.manaraa.com

1.3723. And a student with low TOLT in the reform would expect a decline of 0.8459 with each progressive test (-3.6189 + 2.7730 = -0.8459). Thus students with low TOLT would decline less with the reform than a student with high TOLT in the reform.

Compare this to students without the reform. In general, a student with high TOLT in the reform would be expected to have a decline of 2.5537 with each progressive test (-3.6189 + 1.0652 = -2.5537). And a student with low TOLT without the reform would be feature a decline of 3.6189 throughout the semester, thus declining at a rate steeper than the high TOLT students. Again, while indications that the reform moderates the gap between students with low TOLT and high TOLT, the lack of statistical significance means one must also consider that these differences could be attributed to chance. And while this aspect of the reform seems promising, future investigation will be needed to support or disprove this interpretation.

Conclusions

Both models indicate that the PLGI reform may moderate the impact of formal thought on achievement in chemistry, which is promising, especially considering the role assisting low formal thought students would have on the diversity in science. However, the results found here are tentative, as both effects witnessed cannot be reasonably distinguished from a chance occurrence. Continued investigation would improve the power of the statistical tests and allow a more substantial determination on the effect of the reform on low formal thought students. Additionally, ongoing research in the nature of HLM models in this applied setting may provide insight to guide future research design, and the ensuing statistical analysis and interpretation.

www.manaraa.com

VI. Study Approaches in College Chemistry

Previous chapters have indicated the effectiveness of the PLGI reform and offer several learning theories which may explain the improvement. This chapter begins the examination of the possibility that the reform promotes more effective study approaches among the students. Study approaches may be affected by the classroom environment and are thought to impact students' academic performance.[154, 155] In this chapter, the study approaches that arise from the traditional lecture-only (non-reform) sections are investigated, and related to course performance. This may serve as a baseline for future studies that investigate the study approaches students employ in the PLGI reform. The chapter closes by investigating preliminary data on the study approaches used in the PLGI reform setting.

The study approaches model employed was originally developed from psychology and cognitive science was used to understand the processes students employ in a chemistry classroom, and how these processes relate to chemistry understanding.[156] This model provides the basis for a classification scheme of study processes that students employ in chemistry, making it possible to describe the approaches, the frequency of the approaches, and their usefulness in chemistry understanding.

www.manaraa.com

Study Approach Theory and Past Work

This study employs Biggs' Presage-Process-Product (3P) model of study approaches.[157] In this model presage refers to factors that exist prior to the course and include student factors and teaching context. Student factors describe qualities that individual students bring to the course, such as prior knowledge and abilities, but this could also reasonably include psychological and social traits. Teaching context describes the classroom environment, including the teaching procedures, classroom climate, course objectives and assessment practices. Process describes the study approaches students employ in the course to learn the prescribed material. Product refers to the learning outcomes or the set of facts and skills that students obtain as a result of the classroom experience. In this model, the presage, process and product constructs all interact, each affecting each other and ultimately decide the study approach or 'process' a student employs in a course. Biggs stresses that both the presage and product constructs change between, and even within, courses and as a result one should expect the study approach assigned to students to be only a temporary description of students' current activities and not a stable description of the student. In this framework, an exploration of the study approaches students employ describes primarily the kinds of study approaches the classroom learning environment promotes.

As a result of this model, Biggs describes three approaches students can employ toward a learning task. First, the surface approach is characterized as an emphasis on memorization and reproducing, sometimes called rote learning. In contrast, a deep approach describes an intrinsic interest in the topics covered, as well as in their

www.manaraa.com

underlying constructs. As opposed to the memorizing/reproducing nature of the surface approach, a deep approach will attempt to inter-relate new information with previously learned material. While the first two approaches, surface and deep, have been seen as opposing approaches, the third approach, achieving, is complimentary to either of the first two.[156] The achieving approach describes the desire to earn high grades, regardless of the interest in the material. Because the achieving approach is complementary to the surface and deep approaches, it can be combined with the previous two approaches to create a surface achieving or deep achieving approach. Depending on the desired outcomes for a course, any of these approaches could be considered productive. But given the recent emphasis in the research literature on conceptual understanding in chemistry, there seems to be a desire among chemical education researchers to create a classroom environment that encourages the deep approach to studying. A deep approach to learning, and its emphasis on relating new material to past material and experience, may be expected to promote conceptual understanding in chemistry and it's focus on the relation between chemical concepts. Conversely, the surface approach's emphasis on rote memorization may promote algorithmic learning, which describes a set of procedures that can be memorized.[150, 158]

Research studies have supported the contention that study approaches are dependent on the contexts discussed. Trigwell & Sleet, [159] with a small sample of first year university chemistry students, showed that different assessment techniques tended to reward different study approaches. In particular, they reached the tentative conclusions that open-ended questions tended to encourage the deep approach and that the achieving approach tended to score higher on traditional closed-ended assessment. The authors

www.manaraa.com

postulate that students' employing the achieving approach tend to succeed on assessments they are familiar with, like the traditional closed-ended assessment. Another study with university-level students found that the teachers' approach to teaching impacted the students study approach, with transmission style teaching promoting a surface approach and more student-oriented approaches promoting the deep approach.[160]

Given the role of contextual factors in student study approaches, this study aims to describe the variety of study approaches employed by students in a first year general chemistry setting. Developing such an understanding may explain student performance as a result of the classroom setting to be described, and offer alternatives that may improve student study approaches and ultimately student understanding in chemistry. The next aim of this study is to relate the study approaches employed to course retention and a measure of chemistry knowledge at the end of the course. Understanding this relation can provide future recommendations for which study approaches are successful and should be encouraged within this setting. Finally, given the hypothetical ideal that students' study approach describes how they organize their understanding, this study will investigate if study approaches relate to differences in algorithmic versus conceptual understanding. Past research has indicated a strong tendency for chemistry students to perform better on algorithmic questions than conceptual questions.[134, 135] One suggested reason for this difference is the nature of traditional teaching and assessment methods do not promote conceptual understanding.[158] Student study approach may provide a more specific mechanism for this difference. For example, hypothesizing that students employing a deep approach, where an effort is made to relate concepts to prior

www.manaraa.com

understanding, should perform higher on conceptual problems but not necessarily algorithmic problems. If this holds true, then efforts to promote the deep approach in the setting may alleviate the discrepancy in conceptual understanding that is evident.

Setting: The Teaching Context

As mentioned, the presage part of the model includes teaching context and effects the process part or the study approaches students employ. The primary teaching context for the General Chemistry class is a lecture hall that seats 206 students, with stadium-typed fixed seats that are directed toward the front of the room. During the class time the instructor relies primarily on lecture, with occasional demonstrations or student activities such as discussing a topic with the adjacent students. The reliance on lecture corresponds with a transmission type teaching approach, which is thought to promote the surface approach to studying.[160] Approximately 75% of student grades are determined by five tests (including a final exam) all of which are multiple-choice and timed. The first four tests are created by a panel of instructors, the final exam is the ACS Special Exam meant to combine conceptual and algorithmic question.[72] Each test is multiple choice and given in the normal testing environment described in Chapter 2. The remainder of student grades is determined by online homework and in-class assignments. Students are advised to perform the online homework independently, for their maximum benefit, but it is likely evident to the students that there is no check on whether they work in groups as no online homework problems require free responses. Free on-campus tutoring is available to students during most weekday hours.

www.manaraa.com

Research Methods

Student study approaches were measured by the Study Process Questionnaire (SPQ) survey developed by Biggs. The questionnaire was administered in class immediately following the first test but before the drop date, the last day where students can drop the course without penalty. This was done so that students would have an opportunity to become familiar with the study approaches they employ in the course, but prior to the time when students who were struggling with the course would typically leave. A small amount of points was offered for completing the survey and multiple make-up opportunities were offered for students who were not in class the day of the survey. This was done in seven classes over two semesters, resulting in surveys from 1057 students. Of those, 136 surveys (12.9%) were incomplete. Since the missing responses tended toward the end of the survey and by those who took the survey in class, the amount of time allotted for the survey in class is the most likely explanation for the missing data. To determine if those students who completed the survey were different from the rest of the sample, a comparison between groups on the average score on the ACS Exam was run as shown in Table 6.1.

Table 6.1 – Survey Completion Comparison

	N	Average score	St. Dev.
Completed survey	761	52.07%	18.94%
Took survey, not completed	106	53.18%	16.91%
Did not take survey	261	51.21%	16.52%

www.manaraa.com

The two one-sided t-tests constructed for an 80% confidence interval show that the completed survey group is within plus or minus 0.2 standard deviations from those not completing the survey and those who did not take the survey.[74]

The survey consists of 42 items which the respondent rates on a 1-5 Likert style scale. Fourteen items correspond to each of the three approaches. Traditionally researchers add up the score on the fourteen items to assign three approach scores for a student and uses these approach scores in correlations or factor analysis.[159] This approach leads to some undesirable cases for interpretation. Consider a sample where the average score for each approach is 40 points, and within the sample one student has a surface approach score of 65, a deep score of 50 and an achieving score of 50. This student would be considered above average on all three approaches if a correlation or factor analysis were run on the data set, even though the approach scores indicate this student leans more toward the surface designation. To avoid this misinterpretation a transformation of scores will be employed, whereby first an average approach score is calculated for each student. For the example given, the average approach score would be 55. Then the student's approach score is found by subtracting the original approach score from the average score. The student in the example would then have a surface score of 10, a deep score of -5 and an achieving score of -5, indicating the students is likely to employ the surface approach. The transformation accounts for student's tendency to rate 'all items highly' or 'all items low' by looking at approach scores only in the context of how other items were rated.

The transformed approach scores were screened for outliers by looking for approach scores over 3 standard deviations from the mean. Twelve of the 921 students

www.manaraa.com

had outlier scores and were removed from the analysis; the data analysis presented focuses on the remaining 909 students. The survey had a Cronbach's alpha of 0.87 for the 42 items.

The other instrument used in this study is the aforementioned ACS exam.[72] This exam was constructed by the American Chemical Society and is meant to measure both conceptual and algorithmic understanding. The exam has 40 multiple choice questions, with 19 algorithmic questions and 21 conceptual questions. Internal consistency is demonstrated by a Cronbach's alpha of 0.82 for this sample, and convergent validity was shown by moderate correlations with the set of instructor-created tests.

Results and Discussions

The first research goal was to identify the types of study approaches students employ in the general chemistry setting described. Initial examination began with descriptive statistics of the transformed approach scores as presented in Table 6.2.

Table 6.2 – Study Approach Descriptive Statistics

	Surface	Deep	Achieving
Mean	3.178	-4.316	1.138
Std. Dev.	5.64	5.68	3.99
Skewness	-0.066	0.090	-0.170
Kurtosis	-0.182	0.128	-0.041

As the average scores indicate, the students in the sample have a tendency toward the surface approach, and away from the deep approach. The normality measures indicate a close to normal distribution for each of the approach scores. Combined, these facts

www.manaraa.com

suggest that the majority of students in the sample chose the surface approach to describe their study approach. In a similar fashion, the majority of students also chose the achieving approach, which can be combined with either the surface or deep approach. To determine if the students with the achieving approach tended to use the deep or surface approach, correlations were run among the 3 constructs with results shown in Table 6.3.

Table 6.3 – Correlations between the Approach Scores

	Surface	**Deep**	**Achieving**
Surface	———		
Deep	-0.751*	———	
Achieving	-0.344*	-0.361*	———

* correlation significant at $p < 0.05$

From Table 6.3, there is a strong negative correlation between the surface and deep approach, as expected from the theory. Also, there is a moderate negative correlation between the achieving approach and both the surface and deep approaches. This suggests that students in the sample tended to score the achieving approach highly primarily in occasions where neither the surface nor deep approach was scored strongly. As students score either the deep or surface approach strongly, the achieving approach score tends not to be as strong. Examination of scatter-plots, see Figures 6.1 and 6.2, support this interpretation. However, still unknown is the number of students who employ each approach within the sample.

www.manaraa.com

Figure 6.1 - Relation between Achieve Score and Surface Score

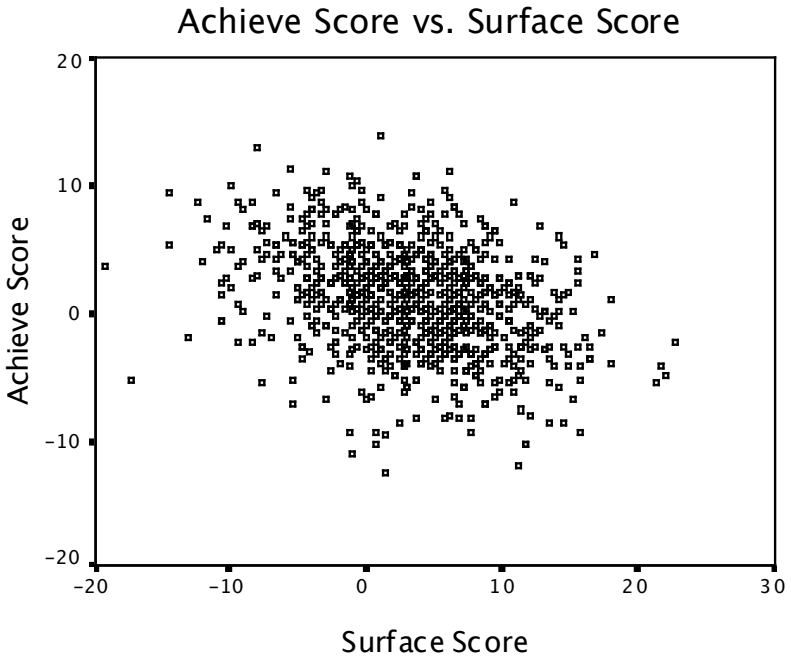
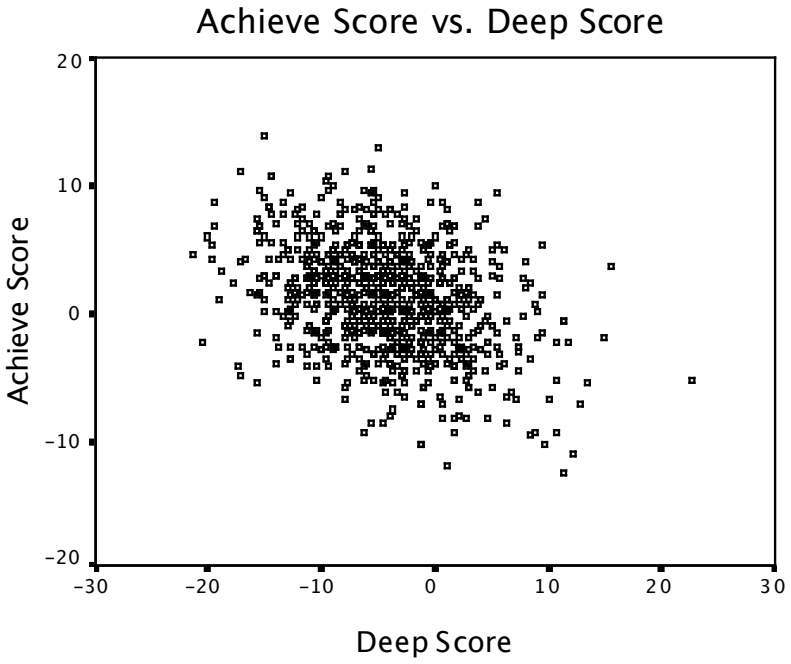


Figure 6.2 - Relation between Achieve Score and Deep Score



www.manaraa.com

To determine the number of students who employ each approach, a cluster analysis on the three approach scores for the entire sample was performed. Cluster analysis is an algorithmic classification scheme that calculates cluster centers in a manner to produce the smallest distance between each data point and the closest cluster center. The algorithmic approach used was Ward's method, with squared Euclidean distance as the measure to cluster center, both decisions were oriented toward producing the most independent groups with minimal over-lap between groups.[161] The remaining parameter to specify is the number of cluster centers; this must be specified prior to the analysis. The number of clusters can be made based on theory or empirical results. In terms of theory, Biggs suggest six study approaches, each of the three approaches independently, a combination of achieving with the surface or deep approach, and a group with a low achieving orientation. However, these suggestions were designed for extreme cases to guide counseling or intervention decisions.[154] Because this suggestion is not suitable for the task at hand, the number of clusters in the sample was evaluated empirically.

To do this, a cluster analysis was run for each number of clusters from two through eight by increments of one. The cluster analysis with eight clusters created eight distinct groups and assigned each person in the sample to one of the eight groups. For each group, the average approach scores were calculated to provide a description for each group. If group 1 had a high surface approach and low deep and achieving approaches, then this group was described as students who employ the surface approach. This description was done for each group provided in the analysis. Evaluation began with the cluster analysis that produced 8 groups, looking for indications of a duplicate group, for

www.manaraa.com

example, two groups that each employed primarily a deep achieving approach. Where duplicate groups were found, the analysis that produced 7 groups was examined, and so on. One additional caveat used was to take notice of which group was deleted when the number of groups was reduced, to ensure a unique group was not being combined with other groups as the number of groups was reduced. This approach recommended 5 groups within the setting, which are described in Table 6.4.

Table 6.4 – Group Descriptions

	Surface	Deep	Achieving	Description
Group 1	6.0192	-3.1073	-2.9119	Surface
Group 2	-0.0817	3.2733	-3.1917	Deep
Group 3	9.0026	-10.6954	1.6928	Surface Achieving
Group 4	0.8961	-5.4632	4.5671	Achieving
Group 5	-4.3848	1.877	2.5078	Deep Achieving

Finding the number of students who employ each approach will detail the prevalence of each study approach within the sample, as presented in Table 6.4. To determine the extent the groups are distinguishable from each other, a predictive discriminant function analysis was run on the three study approach scores. Predictive discriminant function analysis creates a linear function that employs the three approach scores, and then uses this linear function to determine if the groups are distinguishable along this function. In doing so, a qualitative description of how distinct (mutually exclusive) the groups are. This approach found 89.1% agreement with the classification suggested with the cluster analysis method.

www.manaraa.com

Table 6.5 – Student Retention Based on Study Approach

Approach	Initial Sample N (%)	Finished Course N (%)	Percent of group finished course
Surface	174 (19%)	138 (18%)	79.3%
Deep	100 (11%)	73 (9.7%)	73.0%
Surface Achieving	255 (28%)	207 (27%)	81.2%
Achieving	231 (25%)	204 (27%)	88.3%
Deep Achieving	149 (16%)	131 (17%)	87.9%
Total	**909**	**753**	**82.8%**

The distribution of students among the groups seem to indicate that the surface achieving and achieving groups are the most populated, and either group with the deep construct is less populated. Among the sample who completed the survey, 27% belonged to one of the two deep approach groups, compared to 47% belonging to one of the two surface approach groups. The remaining 26% belongs to the achieving group, who do not have a strong surface or deep approach in this setting, but do work toward maximizing grades, representative of the interpretation from the correlations among the three constructs. Note that the low achieving group described by Biggs was not present in an appreciable amount in the sample to be recognized by the cluster analysis, though the deep group does coincide with a relatively low achieving approach score. This may explain the low 73.0% of deep students that finished the course as compared to the overall average of 82.8%. In a similar fashion, the surface approach scored achieving relatively low and also featured a somewhat reduced percent of finishing the course (79.3%) compared to the overall rate (82.8%). This seems to indicate that marking the achieving approach low can be related to leaving the course mid-semester. Among the remaining groups where achieving approach is marked higher, retention seems to be at the course average or higher.

www.manaraa.com

Table 6.6 – Student Performance Based on Study Approach

Approach	ACS Exam Percent	Conceptual Percent	Algorithmic Percent
Deep Achieving	58.1%	55.4%	61.0%
Deep	57.1%	54.3%	60.1%
Surface	51.3%	50.2%	52.6%
Achieving	50.5%	47.0%	54.4%
Surface Achieving	48.3%	46.1%	50.8%
Total	**52.0%**	**49.5%**	**54.8%**

From Table 6.6 it is shown that overall on the ACS Exam performance, both groups that feature a deep approach score markedly higher than the other groups. An ANOVA test indicates that the differences in performance are significant, $F = 9.141\ (748, 4)\ p < 0.05$, and a follow-up Tukey multiple comparison procedure, homogeneous variance assumed (variances were within 0.04 of each other), confirms that the significant difference is a result of the two deep approach groups scoring higher than the remaining groups. No other pair-wise differences were found among the groups.

To further investigate the reason for the differences among the study approaches, performance on the ACS exam was split into performance on just the conceptual questions and performance on the algorithmic questions. On the conceptual questions, ANOVA again indicates that the groups differ significantly, $F = 7.582\ (748, 4)\ p < 0.05$, and the follow-up procedure found that the two deep approaches were each scoring higher than the surface achieving and achieving groups. The algorithmic questions also featured significant group differences, $F = 8.383\ (748, 4)\ p < 0.05$ with the pair-wise test indicating both deep approaches scored higher than surface and surface achieving. The significant pair-wise differences are summarized in Table 6.7

www.manaraa.com

Table 6.7 – Significant Pair-wise Differences

Outcome Measure	ACS	Conceptual	Algorithmic
Higher Scoring Groups	Deep Deep Achieving	Deep Deep Achieving	Deep Deep Achieving
Lower Scoring Groups	Surface Surface Achieving Achieving	Achieving Surface Achieving	Surface Surface Achieving

These results suggest that the achieving approach represents a hindrance on conceptual questions except when it is combined with the deep approach, and the surface approach represents a hindrance on the algorithmic questions. For each of the ANOVAs, the effect size f is approximately 0.2, approaching a medium effect size. In other words, the size of the differences between groups is sufficient that it could be detected in the course of normal experience by reviewing assessment scores.[70]

The negative relation between the achieving approach and conceptual questions suggests that the perception among the sample is that conceptual knowledge is not an important piece in assessment. As these students would be principally concerned about their grades, they may not be focusing on conceptual knowledge, or may believe that focusing on algorithmic operations offers a greater pay-off in terms of assessment. This however only acknowledges the assessment portion of the classroom environment. As mentioned other factors may also contribute to this perception. For example, the teacher or teaching style may be diminishing the importance of conceptual knowledge to the students with the achieving approach. Or past student experience could play a role in this perception: possibly conceptual knowledge wasn't emphasized in past chemistry or science courses.

www.manaraa.com

The success of the deep approaches, compared to the remaining approaches, on both the conceptual and algorithmic questions is in contrast to existing research literature that finds multiple-choice question exams tending to reward the surface approach.[162] This may be a result of the nature of the multiple-choice questions, which may not be easily answered by memorization and recall of facts. Or it may be subject matter specific, with chemistry multiple-choice questions relying on a more integrated understanding that corresponds to the deep approach. Regardless, in this setting it does seem possible to reward a deep approach using multiple choice exams. The relative success of the deep approaches on the conceptual questions corresponds with the previous stated hypothesis that because of their effort to relate new information with previously learned knowledge these students would perform better with conceptual problems. Thus one may expect that steps which encourage the deep study approach in the classroom to improve conceptual understanding in chemistry. However, also notice that the algorithmic performance for the deep approaches were higher than the other approaches. Thus it may be that the students employing the deep approach are simply better students and their higher performance on the conceptual questions is just another reflection of their aptitude. There is most likely some truth to this possibility, it appears the deep approaches comprise a sub-set that is a sizable portion of the higher-performing students. But past research has also found that even good students tend to have a difficult time with conceptual questions, [135] which does not seem to be the case for those showing the deep approaches.

www.manaraa.com

Study Approaches in PLGI Reform

As previously mentioned, the study approaches students employ is thought to depend in part on the classroom setting. Previous studies have shown that assessment likely plays the largest role [163, 164] but teacher beliefs have been shown to have an impact on the study approaches students employ.[160] Because of this, it may be possible that the PLGI reform impacts the study approaches students employ. If this is true, the impact in study approaches may explain the improved chemistry understanding these students demonstrate compared to their lecture-only counter-parts. In particular, it may be that the use of cooperative learning causes negotiation of understanding among the students, and this negotiation may promote more of a desire to understand the underlying reasons. Because of this, the PLGI reform may lead to more students employing the deep approach in the classroom. However, the assessments used in the PLGI setting are un-altered from the lecture-only setting, with the exception of the use of weekly quizzes and homework in the PLGI reform class which played a small role (less than 10%) in the overall grade. Given that the assessments play a large role in study approaches, and that the assessments used in PLGI are similar to that used in the lecture setting, it is also likely that the use of PLGI does not affect the study approaches students employ. Adding to this, the study that found teaching beliefs impacted student study approach [160] did not investigate assessments used, though likely the assessments were impacted by teaching beliefs.

To investigate the potential impact of PLGI on study approaches students employ, the SPQ survey was administered in the one PLGI class during the fall semester of 2004. The procedures for administration were identical to that used in the lecture-only classes,

www.manaraa.com

however due to logistics, no make-up opportunity was offered for these students. The effect of this departure in procedure will be addressed later on in the discussion. Data analysis steps, including the transformation, and treatments of missing data are identical to the lecture-only students.

The survey administration in the PLGI class resulted in complete surveys for 96 students in the PLGI section, out of approximately 190 students who were initially enrolled. Table 6.8 compares SAT scores and ACS exam scores for those who completed the survey versus those who did not complete or did not take the survey, showing very little difference between these two groups on these measures.

Table 6.8 – Survey Completion Comparison

Measure	Survey	N	Mean	Std. Dev.
SAT Math	Not Complete	74	567	77.0
	Completed	82	565	82.5
SAT Verbal	Not Complete	74	546	79.1
	Complete	82	543	80.7
ACS Exam	Not Complete	55	59.7%	18.6%
	Complete	77	57.7%	17.2%

Outlier analysis indicates that 2 students had scores above 3 standard deviations and were removed from the analysis. The approach scores in the PLGI reform were compared to the approach scores for the classes in the same fall semester in Table 6.9.

Table 6.9 – Comparison of Average Study Approach Scores

Study Approach	PLGI n = 94 (std. dev)	Lecture-only n = 385 (std. dev)	d-value (significance)
Surface	2.816 (5.781)	2.299 (5.78149)	-0.090 (n.s.)
Deep	-4.450 (4.885)	-3.470 (5.63402)	0.186 (n.s.)
Achieve	1.635 (4.484)	1.171 (4.088)	-0.108 (n.s.)

No significant differences were found between the PLGI section and the lecture-only section on the three approach scores. The d-value provides a standardized measure of the

www.manaraa.com

difference, with 0.2 indicating a small difference, or a difference that is just distinguishable from the noise in the data.[70] The deep approach score had a d-value of 0.186 indicating that the difference was near the criteria for a small difference, but could still be attributed to chance. Also note the direction of the difference indicates that the deep approach was scored less in the PLGI section compared to the Lecture-only section, which is opposite what may have been expected.

Based on Table 6.9 though, it appears the students in the PLGI reform had a negligible difference to the students in the lecture-only course in the same semester. One possibility is the time the survey was administered, which was relatively early in the semester, after the first test in the course. It may be that the PLGI reform did not have sufficient time to impact study approaches students employ at this point. Administering the survey later in the semester to both cohorts may provide a more noticeable difference. Finally, to establish a stronger causal relationship between PLGI reform and student study approaches, a pre/post administration of the survey would likely provide a stronger indication. Additionally, as the study approaches are self-reported by the students in the sample, the incorporation of other study approach measures to triangulate the results would also add to the confidence in the results.

Finally it is possible that the change in methodology, where the make-up was offered in the lecture only class but not in the PLGI class, affected the results. The exact impact this discrepancy has cannot be known, however, as a proxy for the effect, a comparison of the study approach groups between those who took the make-up and those who took it in-class among the non-PLGI sections is performed. The results of this comparison are shown in Table 6.10.

www.manaraa.com

Table 6.10 – Study Approach Dependence on In-class Versus Make-up

Study Approach	In Class	Make-up
Surface	18.2%	18.7%
Deep	10.0%	8.9%
Surface Achieving	27.8%	26.6%
Achieving	26.9%	27.6%
Deep Achieving	17.1%	18.2%
Total	**550**	**203**

From Table 6.10 it appears there are only minor differences, as no group differs between in class and make-up by more than 2 percent. Chi square analysis on the effect of the make-up on the distribution showed no significant effect, $\chi^2 = 0.435$ (4) $p > 0.05$. Based on the results of the non-PLGI sections, there seems to be no effect in the group distribution for offering the make-up or not offering the make-up. Whether this is true for the PLGI section is not known, but there is no evidence to believe that the students who took the survey in the PLGI section are not representative of the responses in the PLGI section had the make-up been offered.

Implications

This study indicates that for the ACS Exam, the deep approaches are more successful and should be encouraged, but within the sample the deep approach represents only 27% of the students. Thus efforts to promote the deep approach within the classroom seems to be warranted. Past research has indicated that study approaches in the tertiary level are fairly stable, and that while they seem to be dependent on the discipline (psychology students show different scores from science students for example) within a discipline they tend not to change.[156, 165] However, these studies make this claim by comparing a first-year science classroom with a senior science classroom and

www.manaraa.com

finding them the same. This does not investigate whether study approaches could be altered by classroom interventions. Among the possibilities, the earlier mentioned work of Trigwell, *et. al*., suggests that the use of more student-oriented teaching may promote the deep approaches in this setting.[160] This describes a teaching intervention, but other classroom interventions are also plausible.

In this study, it was shown that the multiple-choice exam rewards those who employ a deep study approach, by way of their higher scores. It has been proposed that this relationship encourages the deep approach among students, [159] but other studies have found that even when multiple choice assessments measure more than factual recall, it still spurs the employment of surface approaches in students.[162, 166] This may be the case in the current setting, where the final exam measure seems to reward the deep approach, yet 47% of the sample employ one of the surface approaches and another 26% employ the achieving approach. Varying the assessment procedures may spur the deep approach, but with large class sizes often the range of assessments is limited. Other interventions may assist those employing the surface approach, such as emphasis on note-taking or time management skills.[167]

Finally, take note of the interesting relationship between the achieving construct, student retention and performance on the ACS Exam. As has been noted students with a low achieving score have a higher tendency to leave the course. However students with a high achieving score tend to perform worse on the conceptual portion of the assessment that was given, though reasons for this may just as well result from the classroom contexts as it does from student factors, as discussed. Suggestions for interventions to the achieving approach are understandably limited. The achieving approach was designed to

www.manaraa.com

be decontextualized, [162] minimally dependent on classroom environment, and to indicate the competitive nature of students.[156] Extremes on either end of the achieving construct may benefit from counseling, but the effect of classroom interventions on the achieving construct at this time are uncertain. One possibility, proposed by Trigwell and Sleet is that the students with high achieving scores tend to perform better on assessments they are familiar with.[159] Thus using a variety of assessment techniques in the classroom may promote the repertoire of assessments these students are familiar with, and subsequently improve their performance across multiple types of assessment of student understanding.

Conclusions

Student study approaches in chemistry can provide important information for both teachers and researchers on the resulting understanding students show in the course. This study has found a wide variety of study approaches that students employ within the setting, and has found which approaches are more successful as determined by a nationally available ACS exam. In particular, concern is given for the large percent of students employing the surface approach with unsuccessful results, and some suggested interventions to reduce this percent. The achieving approach has been linked to poor conceptual performance and may provide a rationale for the discrepancy in conceptual understanding found in other studies. Future work could focus on the effect of altering the classroom environment on the prevalence of study approaches students employ.

www.manaraa.com

VII. Conclusions and Future Directions

The principal focus of this work was to determine if the reform was successful in improving students understanding and in what ways was it successful. To address this, multiple angles of consideration were examined, with a particular emphasis on pre-existing achievement gaps. First, and foremost, the reform was found to be a more effective teaching technique than a traditional lecture-based approach. This finding was, on average, true for all students in the setting, across the three years the reform ran, even when controlling for student SAT sub-scores or student performance on a formal thought measure. Two significant indications, in agreement with each other, lead to this conclusion: first, students in the reform performed progressively better in the semester than their non-reform counterparts, and second, the students in the reform outperformed their counterparts on an external-to-the-institution exam which the students and instructors likely had no prior experience with. The congruence of these two findings leads to the conclusion that the reform is a more effective teaching pedagogy than the traditional lecture.

This evaluation also considered a unique perspective by examining the pre-existing achievement gaps presented by students' SAT sub-scores and performance on a formal thought measure. First, it was demonstrated that formal thought and SAT sub-scores represent independent factors in determining chemistry success, so that the achievement gaps present on the different constructs are distinct from each other.

www.manaraa.com

Second, the reform was evaluated for its effect on the achievement gaps present, taking advantage of the large sample size. The reform was found to have a limited impact on the SAT sub-score achievement gap that is the reform did not noticeably widen or shrink the pre-existing achievement gap. Regarding the formal thought achievement gap, there are some promising indications that the reform assists students with low formal thought, but this could not be satisfactorily concluded.

Finally this work investigated the study approaches students employed in the setting, by use of the Study Processes Questionnaire. This investigation was meant to provide a descriptive indication of the students' approaches to the course, the usefulness of each study approach toward students' performance on the ACS Exam, and serve as a baseline for evaluating any impact the reform may have on the approaches employed. Cluster analysis created meaningful groups in this context, indicating the questionnaire is appropriate for the course and viable for future studies. There were indications that the deep approach was generally the most successful approach employed, and is relatively rare in this setting. No indication was present that the reform spurred the deep approach; the impact of the reform on study approaches as a whole remains an area of open, and possibly fruitful, investigation.

Relevance of the Work Presented

First, as part of a broad effort to improve science education, this evaluation serves as an important piece in understanding ways to promote better understanding among science students. While in general it has been shown that reform style teaching improves academic achievement over traditional lecture-based styles (see Johnson & Johnson [14]

www.manaraa.com

for example), this is among the first to demonstrate a reform that is specifically tailored to a large lecture class is also successful. As large lecture classes at universities are plentiful, this is an important step in bridging the gap between reform practice and large universities.[168]

Second, this evaluation is unique in its focus on equity or the pre-existing achievement gaps present within the setting. Comparison of average scores, or even average scores while controlling for one variable, may ignore the fact that one group is not being assisted, or worse disadvantaged, by the reform. By examining the impact of the reform on the already-present relationship between pre-existing measures and academic achievement helps to elucidate the effect of the reform on the achievement gaps present at the beginning of the course. Finding no impact on the equity in the classroom by the reform, indicates that students with low high school preparation or formal thought do benefit from the reform, but no more than their better prepared counterparts. This finding highlights the need, and challenge, to offer a science curriculum better tailored to students who enter the setting with less preparation. Furthermore it also offers other educational researchers a means and rationale for examining the impact other reforms may have on equity.

Finally, by demonstrating the effectiveness of the reform, this work has served as a basis for dissemination efforts with practitioners across the nation. The reform is unique in that it still offers the majority of time for conventional teaching approaches, and still covers all the material conventionally associated with the course. In showing that such a limited reform improves students' academic performance, this work may present a convincing argument to those practitioners interested in reform, but hesitant to

www.manaraa.com

abandon conventional teaching practices. By promoting dissemination, the potential for improving science understanding on a larger scale may be realized.

Future Projects Suggested by This Work

As a result of this work, a number of future projects are readily available. First, as a result of understanding that the reform is successful and for whom, this work naturally leads to why is it successful? What aspects of the reform are promoting this improved understanding? Because of the nature of the reform, several reasons have been postulated, including cooperative learning, the ability of peer leaders to naturally hit the students' zones of proximal development or the inquiry nature of the activities. Future studies can begin to understand the role each of these factors play in students' chemistry understanding. Such projects may involve comparing the actions that take place in the PLGI reform to a more conventional cooperative learning setting, or by comparing the reform to a class where inquiry materials are assigned to students to work on individually. Additionally, the role of the peer leader in the PLGI reform can be investigated from an ethnographic perspective to better understand the impact this position has on the classroom environment. It is worth noting that investigations into how the reform works would not be possible without this prior work establishing that the reform is an effective means for teaching.

This work has also highlighted the achievement gaps that are present in this setting and can be described by pre-semester measures. As mentioned, there is a need to further understand the underlying reasons behind these achievement gaps so that effective remediation or in-course remedies can be developed. Finding that this reform did assist

www.manaraa.com

all students in the course is a positive first step, but there was no indication that the reform began reducing these achievement gaps. Understanding the roles that students with poor preparation exhibit while in the reform setting may serve as a foundation for altering the reform to better assist these students. Doing so, as mentioned, has a strong potential for promoting diversity in the sciences and approaching the Science Education Standards' goal of "science for everyone." [7]

As mentioned before, the reform is limited in the extent that it alters a traditional classroom. Only one class per week out of three was changed with the reform, and this was thought to make dissemination more feasible, particularly to instructors that are well-versed in the traditional teaching style. But, as the reform is found to be effective in even this limited fashion, a reasonable direction may be to determine if this effect can be maximized by devoting more class time to the reform, and less to the traditional teaching style. This could even be extrapolated to the point where an entire class could be devoted to the reform style teaching. Such a change would require a rebalancing of the tasks between peer leaders and the course instructor, but it would still be feasible. In addition to effectiveness, this change may also promote classroom equity more as classroom modifications to assist students with lower preparation would have more opportunity to take hold.

Another area for future study would be on the transferability of this reform to other institutions. One limitation of the current study is that it takes place in only one institution, and implementation and sustainability at other institutions may face alternative barriers that hinder success. Since the reform is explicitly designed for large lecture classes which are common at larger universities, of particular interest would be in

www.manaraa.com

implementing the reform at different institution types, such as four-year universities or community colleges. As these types of institutions tend to be more open to reform teaching ideas, [168] the successful implementation of this reform at these levels would greatly aid any dissemination effort. One potential barrier to this transfer would be the role of peer leaders. In the current setting, peer leaders are required as a result of the large class sizes. When class size is decreased, the need for peer leaders changes, and their role in the setting may directly change as a result. This potential effect would be an interesting area of research and help uncover how much of the reform benefits can be attributed to the role the peer leaders play in the current setting.

While this work focused largely on student benefits, in particular with academic achievement, future work could also focus on benefits the peer leaders receive. Peer leaders may improve their understanding of chemistry topics as a result of this experience, for example. Another possibility is the development of peer leaders' pedagogical breadth and beliefs as a result of this experience. As most peer leaders have little teaching experience, this reform could provide an ideal ground for exploring the experiences and beliefs of beginning teachers. Such a study may help guide teacher training programs, as well as the peer leader training performed in the reform.

Also building on student benefits that have been demonstrated, would be any investigation that probes long-term benefits from the reform. Initial indications are that students who attended the reform did perform better in a follow-on course [63], but this study was limited to only one semester of the reform. A more comprehensive investigation of long-term student performance would indicate transferability of the concepts learned in the reform setting, which is an important hallmark of meaningful

www.manaraa.com

learning.[169] Additionally, the reform may have a possibility of creating a support group for students that reduces student drop-out later in the college curriculum, or the reform may promote interest in science, both of which would be interesting topics for future studies, and add to the positive facets of the reform demonstrated here.

www.manaraa.com

References

1. Karplus, R. and H.D. Thier, *A New Look at Elementary School Science*. New Trends in Curriculum and Instruction Series, ed. J.U. Michaelis. 1967, Chicago: Rand McNally & Company. 204.
2. Kratochvil, D.W. and J.J. Crawford, *Science Curriculum Improvement Study*. 1971, American Institutes for Research in the Behavioral Sciences: Palo Alto. p. 44.
3. Farrell, J.J., R.S. Moog, and J.N. Spencer, *A Guided Inquiry General Chemistry Course*. Journal of Chemical Education, 1999. **76**(4): p. 570-574.
4. Johnson, M.A. and A.E. Lawson, *What are the Relative Effects of Reasoning Ability and Prior Knowledge on Biology Achievement in Expository and Inquiry Classes?* Journal of Research in Science Teaching, 1998. **35**(1): p. 89-103.
5. Keselman, A., *Supporting Inquiry Learning by Promoting Normative Understanding of Multivariate Causality*. Journal of Research in Science Teaching, 2003. **40**(9): p. 898-921.
6. Zoller, U., *Scaling-Up of Higher-Order Cognitive Skills-Oriented College Chemistry Teaching: An Active-Oriented Research*. Journal of Research in Science Teaching, 1999. **36**(5): p. 583-596.
7. National Research Council, *National Science Education Standards*. 1996, National Academy Press: Washington, D.C.
8. The POGIL Project, *www.pogil.org*.
9. Cracolice, M.S., *How Students Learn: Knowledge Construction in College Chemistry Courses*, in *Chemists' Guide to Effective Teaching*, N.J. Pienta, M.M. Cooper, and T.J. Greenbowe, Editors. 2005, Prentice Hall: Upper Saddle River. p. 12-27.
10. Spencer, J.N., *New Directions in Teaching Chemistry: A Philosophical and Pedagogical Basis*. Journal of Chemical Education, 1999. **76**(4): p. 566-569.
11. Dreyfuss, A., *The PLTL Workshop Project Web Pages*. 2003: New York.
12. Tien, L.T., V. Roth, and J.A. Kampmeier, *Implementation of a Peer-Led Team Learning Instructional Approach in an Undergraduate Organic Chemistry Course*. Journal of Research in Science Teaching, 2002. **39**(7): p. 606-632.
13. Vygotsky, L.S., *Thought and Language*. Translation newly rev. and edited ed. 1986, Cambridge: MIT Press. 287.
14. Johnson, D.W. and R.T. Johnson, *Cooperation and Competition: Theory and Research*. 1989, Edina, Minnesota: Interaction Book Company. 257.
15. Springer, L., M.E. Stanne, and S.S. Donovan, *Effects of Small-Group Learning on Undergraduates in Science, Mathematics, Engineering, and Technology: A Meta-Analysis*. Review of Educational Research, 1999. **69**(1): p. 21-51.

www.manaraa.com

16. Bowen, C.W., *A Quantitative Literature Review of Cooperative Learning Effects on High School and College Chemistry Achievement*. Journal of Chemical Education, 2000. **77**(1): p. 116-119.
17. Slavin, R.E., *Research on Cooperative Learning and Achievement: What We Know, What We Need to Know*. Contemporary Educational Psychology, 1996. **21**: p. 46-69.
18. Slavin, R.E., *When Does Cooperative Learning Increase Student Achievement?* Psychological Bulletin, 1983. **94**(3): p. 429-445.
19. Karau, S.J. and K.D. Williams, *Social Loafing: A Meta-Analytic Review and Theoretical Integration*. Journal of Personality and Social Psychology, 1993. **65**(4): p. 681-706.
20. Ross, J.A. and C. Rolheiser, *Student Assessment Practices in Co-operative Learning*, in *Co-operative Learning: The social and intellectual outcomes of learning in groups*, R.M. Gillies and A.F. Ashman, Editors. 2003, RoutledgeFalmer: London. p. 119-135.
21. Wittrock, M.C., *A Generative Model of Mathematics Learning*. Journal for Research in Mathematics Education, 1974. **5**: p. 181-196.
22. Webb, N.M., *Student Interaction and Learning in Small Groups*, in *Learning to Cooperate, Cooperating to Learn*, R.E. Slavin, et al., Editors. 1985, Plenum Press: New York. p. 147-172.
23. Webb, N.M., *Task-Related Verbal Interaction and Mathematics Learning in Small Groups*. Journal for Research in Mathematics Education, 1991. **22**(5): p. 366-389.
24. Bodner, G., *Constructivism: A Theory of Knowledge*. Journal of Chemical Education, 1986. **63**(10): p. 873-878.
25. Moog, R.S. and J.J. Farrell, *Chemistry: A Guided Inquiry, 2nd Edition*. 2002, Hoboken, NJ: John Wiley & Sons. 376.
26. Cohen, E.G., R.A. Lotan, and N.C. Holthuis, *Organizing the Classroom for Learning*, in *Working For Equity in Heterogeneous Classrooms: Sociological Theory in Practice*, E.G. Cohen and R.A. Lotan, Editors. 1997, Teachers College Press: New York. p. 31-43.
27. Webb, N.M., et al., *Equity Issues in Collaborative Group Assessment: Group Composition and Performance*. American Educational Research Journal, 1998. **35**(4): p. 607-651.
28. Webb, N.M., K.M. Nemer, and S. Zuniga, *Short Circuits or Superconductors? Effects of Group Composition on High Achieving Students' Science Assessment Performance*. American Educational Research Journal, 2002. **39**(4): p. 943-989.
29. Bianchini, J.A., *From Here to Equity: The Influence of Status on Student Access to and Understanding of Science*. Science Education, 1999. **83**: p. 577-601.
30. Webb, N.M., *Sex Differences in interaction and achievement in cooperative small groups*. Journal of Educational Psychology, 1984. **76**: p. 33-44.
31. Keys, C.W. and L.A. Bryan, *Co-Constructing Inquiry-Based Science with Teachers: Essential Research for Lasting Reform*. Journal of Research in Science Teaching, 2001. **38**(6): p. 631-645.

www.manaraa.com

32. Roehrig, G.H. and J.A. Luft, *Inquiry Teaching in High School Chemistry Classrooms: The Role of Knowledge and Beliefs*. Journal of Chemical Education, 2004. **81**(10): p. 1510-1516.
33. Lewis, S.E. and J.E. Lewis, *Departing from Lectures: An Evaluation of a Peer-Led Guided Inquiry Alternative*. Journal of Chemical Education, 2005. **82**(1): p. 135-139.
34. Johnson, D.W. and R.T. Johnson, *Assessing Students in Groups: Promoting Group Responsibility and Individual Accountability*. Experts in Assessment, ed. T.R. Guskey and R.J. Marzano. 2004, Thousand Oaks: Corwin Press. 206.
35. Committee on Undergraduate Science Education, *Science Teaching Reconsidered; A Handbook*. 1997: National Academy Press. 88.
36. Advisory Committee to the Directorate for Education and Human Resources, *Shaping the Future Volume II: Perspectives on Undergraduate Education in Science, Mathematics, Engineering and Technology*. 1998, National Science Foundation: Arlington, VA. p. 399.
37. Von Secker, C.E. and R.W. Lissitz, *Estimating the Impact of Instructional Practices on Student Achievement in Science*. Journal of Research in Science Teaching, 1999. **36**(10): p. 1110-1126.
38. Johnson, D.W., R.T. Johnson, and K.A. Smith, *Cooperative Learning Returns to College: What Evidence is There That It Works?* Change, 1998. **30**: p. 27-35.
39. Okebukola, P.A., *The Relative Effectiveness of Cooperative and Competitive Interaction Techiniques in Strengthening Students' Performance in Science Classes*. Science Education, 1985. **69**(4): p. 501-509.
40. Burron, B., M.L. James, and A.L. Ambrosio, *The Effects of Cooperative Learing in a Physical Sceince Course for Elementary/Middle Level Preservice Teachers*. Journal of Research in Science Teaching, 1993. **30**(7): p. 697-707.
41. Balfakih, N.M.A., *The effectiveness of student team-achievement division (STAD) for teaching high school chemistry in the United Arab Emirates*. International Journal of Science Education, 2003. **25**(5): p. 605-624.
42. Schachar, H., *Who gains what from co-operative learning: an overview of eight studies*, in *Co-operative Learning: The social and intellectual outcomes of learning in groups*, R.M. Gillies and A.F. Ashman, Editors. 2003, RoutledgeFalmer: London.
43. Kromrey, J.D., *Detecting Unit of Analysis Problems in Nested Designs: Statistical Power and Type I Error Rates of the F Test for Groups-Within-Treatments Effects*. Educational and Psychological Measurement, 1996. **56**(2): p. 215-231.
44. Saxe, G.B., M. Gearhart, and M. Seltzer, *Relations between Classroom Practices and Student Learning in the Domain of Fractions*. Cognition and Instruction, 1999. **17**(1): p. 1-24.
45. Luke, D.A., *Multilevel Modeling*. Quantitative Applications in the Social Sciences, ed. M.S. Lewis-Beck. Vol. 143. 2004, London: Sage Publications. 78.

www.manaraa.com

46. Willett, J.B., J.D. Singer, and N.C. Martin, *The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations*. Development and Psychopathology, 1998. **10**: p. 395-426.
47. Tai, R.H. and P.M. Sadler, *Gender differences in introductory undergraduate physics performance: university physics versus college physics in the USA*. International Journal of Science Education, 2001. **23**(10): p. 1017-1037.
48. Supovitz, J.A. and H.M. Turner, *The Effects of Professional Development on Science Teaching Practices and Classroom Culture*. Journal of Research in Science Teaching, 2000. **37**(9): p. 963-980.
49. Nolen, S.B., *Learning Environment, Motivation, and Achievement in High School Science*. Journal of Research in Science Teaching, 2003. **40**(4): p. 347-368.
50. Kane, R., S. Sandretto, and C. Heath, *Telling Half the Story: A Critical Review of Research on the Teaching Beliefs and Practices of University Academics*. Review of Educational Research, 2002. **72**(2): p. 177-228.
51. Wittrock, M.C., *Students' Thought Processes*, in *Handbook of Research on Teaching*, M.C. Wittrock, Editor. 1986, Macmillian Publishing Company: New York. p. 297-314.
52. Cohen, E.G., *Understanding Status Problems: Sources and Consequences*, in *Working For Equity in Heterogeneous Classrooms: Sociological Theory in Practice*, E.G. Cohen and R.A. Lotan, Editors. 1997, Teachers College Press: New York. p. 61-76.
53. Cohen, E.G. and R.A. Lotan, *Producing Equal-Status Interaction in the Heterogenous Classroom*. American Educational Research Journal, 1995. **32**(1): p. 99-120.
54. Cohen, E.G., et al., *Complex Instruction: Higher-Order Thinking in Heterogenous Classrooms*, in *Handbook of Cooperative Learning Methods*, S. Sharan, Editor. 1994, Greenwood Press: Westport. p. 82-96.
55. Bianchini, J.A., *Where Knowledge Construction, Equity, and Context Intersect: Student Learning of Science in Small Groups*. Journal of Research in Science Teaching, 1997. **34**(10): p. 1039-1065.
56. Bryan, L.A. and M.M. Atwater, *Teacher Beliefs and Cultural Models: A Challenge for Science Teacher Preparation Programs*. Science Education, 2002. **86**: p. 821-839.
57. Oakes, J., *Multiplying Inequalities: The Effects of Race, Social Class, and Tracking on Opportunities to Learn Mathematics and Science*. 1990: The RAND Corporation.
58. Darling-Hammond, L., *Inequality and Access to Knowledge*, in *Handbook of Research on Multicultural Education*, J.A. Banks and C.A. McGee Banks, Editors. 1995, Macmillan Publishing USA: New York. p. 465-483.
59. Seymour, E., *Tracking the Processes of Change in US Undergraduate Education in Science, Mathematics, Engineering, and Technology*. Science Education, 2001. **86**: p. 79-105.

www.manaraa.com

60. Tobias, S., *They're Not Dumb They're Different; Stalking the Second Tier*. Fifth ed. 1994, Tuscon: Research Corporation; A Foundation for the Advancement of Science. 94.
61. Tai, R.H., P.M. Sadler, and J.F. Loehr, *Factors Influencing Success in Introductory College Chemistry*. Journal of Research in Science Teaching, 2005. **42**(9): p. 987-1012.
62. National Science Teachers Association, *Standards for Science Teacher Prepartion*. 2003.
63. Lewis, S.E., T.M. Eckart, and J.E. Lewis, *Inquiry Teaching for Large Classrooms: Possibilities for Lasting Benefits and Best Practices for Teaching*. Journal of College Science Teaching, 2005. Submitted for Publication.
64. Raudenbush, S.W. and A.S. Bryk, *Hierarchical Linear Models: Applications and Data Analysis Methods*. Advanced Quantitative Techniques in the Social Sciences Seeries. 2002, Thousand Oaks, California: Sage. 485.
65. Stevens, J.P., *Intermediate Statistics: A Modern Approach*. Second ed. 1999, Mahwah, New Jersey: Lawrence Erlbaum Associates. 424.
66. Spencer, H.E., *Mathematical SAT Test Scores and College Chemistry Grades*. Journal of Chemical Education, 1996. **73**(12): p. 1150-1153.
67. Wagner, E.P., H. Sasser, and W.J. DiBiase, *Predicting Students at Risk in General Chemistry Using Pre-Semester Assessments and Demographic Information*. Journal of Chemical Education, 2002. **79**(6): p. 749-755.
68. Bunce, D.M. and K.D. Hutchinson, *The Use of the GALT (Group Assessment of Logical Thinking) as a Predictor of Academic Success in College Chemistry*. Journal of Chemical Education, 1993. **70**(3): p. 183-187.
69. Lewis, S.E. and J.E. Lewis. *Identifying At-Risk Students in General Chemistry: A Comparison of a Formal Thought Measure and a General Aptitude Measure*. in *18th Biennial Conference on Chemical Education*. 2004. Ames City.
70. Cohen, J., *Statistical Power Analysis for the Behavioural Sciences*. Second ed. 1988, Hillsdale: Lawrence Erlbaum Associates. 567.
71. Singer, J.D., *Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models*. Journal of Educational and Behavioral Statistics, 1998. **24**(4): p. 323-355.
72. Examinations Institute of the American Chemical Society Division of Education, *First Term General Chemistry (Special Examination)*. 1997, Clemson Univeristy: Clemson, SC.
73. Cohen, S.J. and L.J. Cronbach, *College Board Scholastic Aptitude Test and Test of Standard Written English*, in *The Mental Measruements Yearbook*, B.I.o.M. Measurement, Editor. 1985, Rutgers University Press: New Brunswick.
74. Lewis, S.E. and J.E. Lewis, *The Same or Not the Same: Equivalance as an Issue in Educational Research*. Journal of Chemical Education, 2005. **82**(9): p. 1408-1412.
75. Kreft, I.G.G., J.d. Leeuw, and L.S. Aiken, *The Effect of Different Forms of Centering in Hierarchical Linear Models*. Multivariate Behavioral Research, 1995. **30**(1): p. 1-21.

www.manaraa.com

76. Littell, R.C., et al., *SAS System for Mixed Models*. 1996, Cary, NC: SAS Institute Inc.
77. Snijders, T.A. and R.J. Bosker, *Modeled Variance in Two-Level Models*. Sociological Methods & Research, 1994. **22**(3): p. 342-363.
78. Gibson, N.M. and S. Olejnik, *Treatment of Missing Data at the Second Level of Hierarchical Linear Models*. Educational and Psychological Measurement, 2003. **63**(2): p. 204-238.
79. Toland, M.D. and R.J. De Ayala, *A Multilevel Factor Analysis of Students' Evaluations of Teaching*. Educational and Psychological Measurement, 2005. **65**(2): p. 272-296.
80. Institute of Education Sciences, *IES Biennal Report to Congress*. 2005, Institute of Education Sciences: Washington, DC.
81. McIntosh, W.G., *Teaching Standards*, in *College Pathways to the Science Education Standards*, E.D. Siebert and W.G. McIntosh, Editors. 2001, NSTA Press: Arlington. p. 1-24.
82. Madigan, T., *Science Proficiency and Course Taking in High School: The Relationship of Science Course-Taking Patterns to Increases in Science Proficiency between 8th and 12th Grades*, in *ERIC*. 1997, ERIC Document Number ED 407 279: Suitland, Maryland.
83. Seymour, E. and N.M. Hewitt, *Talking About Leaving; Why Undergraduates Leave the Sciences*. 1997, Boulder, Colorado: Westview Press. 429.
84. Raudenbush, S.W. and X. Liu, *Statistical Power and Optimal Design for Multisite Randomized Trials*. Psycholgocial Methods, 2000. **5**(2): p. 199-213.
85. Pienta, N.J., *A Placement Examination and Mathematics Tutorial for General Chemistry*. Journal of Chemical Education, 2003. **11**: p. 1244-1246.
86. Orgill, M. and G. Bodner, *The Role of Analogies in Chemistry Teaching*, in *Chemists' Guide to Effective Teaching*, N.J. Pienta, M.M. Cooper, and T.J. Greenbowe, Editors. 2005, Prentice Hall: Upper Saddle River, NJ. p. 90-105.
87. Inhelder, B. and J. Piaget, *The Growth of Logical Thinking from Childhood Adolescence*. 1958, New York: Basic Books Inc. 356.
88. Shayer, M. and P.S. Adey, *Towards a Science of Science Teaching*. 1981, London: Heinemann Educational Books. 159.
89. Adey, P.S. and M. Shayer, *Really Raising Standards: Cognitive Intervention and Academic Achievement*. First ed. 1994, New York: Routledge.
90. Lawson, A.E., R. Karplus, and H. Adi, *The Acquisition of Propositional Logic and Formal Operational Schemata During the Secondary School Years*. Journal of Research in Science Teaching, 1978. **15**(6): p. 465-478.
91. Lawson, A.E. and F.H. Nordland, *The Factor Structure of Some Piagetian Tasks*. Journal of Research in Science Teaching, 1976. **13**(5): p. 461-466.
92. Lawson, A.E. and J.W. Renner, *Relationships of Science Subject Matter and Developmental Levels of Learners*. Journal of Research in Science Teaching, 1975. **12**(4): p. 347-358.
93. Vosniadou, S. and W.F. Brewer, *Theories of Knowledge Restructuring in Development*. Review of Educational Research, 1987. **57**(1): p. 51-67.

www.manaraa.com

94. Novak, J.D., *Meaningful Learning: The Essential Factor for Conceptual Change in Limited or Inappropriate Propositional Hierarchies Leading to Empowerment of Learners*. Science Education, 2002. **86**: p. 548-571.
95. Posner, G.J., et al., *Accomodation of a Scientific Concept: Toward a Theory of Conceptual Change*. Science Education, 1982. **66**(2): p. 211-227.
96. Russell, A.A., *A Rationally Designed General Chemistry Diagnostic Test*. Journal of Chemical Education, 1994. **71**: p. 314-317.
97. Ausubel, D.P., E.V. Sullivan, and S.W. Ives, *Theory and Problems of Child Development*. Third ed. 1980, New York: Grune & Stratton, Inc. 652.
98. Lawson, A.E., *A Review of Research on Formal Reasoning and Science Teaching*. Journal of Research in Science Teaching, 1985. **22**(7): p. 569-617.
99. Lawson, A.E., *Formal Reasoning, Achievement, and Intelligence: An Issue of Importance*. Science Education, 1982. **66**(1): p. 77-83.
100. Lawson, A.E., *Predicting Science Achievement: The Role of Develomental Level, Disembedding Ability, Mental Capacity, Prior Knowledge, and Beliefs*. Journal of Research in Science Teaching, 1983. **20**(2): p. 117-129.
101. Lawson, A.E., *Combining Variables, Controlling Variables, and Proportions: Is There a Psychological Link?* Science Education, 1979. **63**(1): p. 67-72.
102. Chandran, S., D.F. Treagust, and K.G. Tobin, *The Role of Cognitive Factors in Chemistry Achievement*. Journal of Research in Science Teaching, 1987. **24**(2): p. 145-160.
103. Beck, H.P. and W.D. Davidson, *Establishing an Early Warning System: Predicting Low Grades in College Students from Survey of Academic Orientation*. Research in Higher Education, 2001. **42**(6): p. 709-723.
104. Beck, H.P., S. Rorrer-Woody, and L.G. Pierce, *The Relations of Learning and Grade Orientations to Academic Performance*. Teaching of Psychology, 1991. **18**(1): p. 35-37.
105. Hackett, G., et al., *Gender, Ethnicity, and Social Cognitive Factors Predicting the Academic Achievement of Students in Engineering*. Journal of Counseling Psychology, 1992. **39**(4): p. 527-538.
106. Brown, N.W., *Cognitive, Interest, and Personality Variables Predicting First-Semester GPA*. Psychological Reports, 1994. **74**: p. 605-606.
107. Wolfe, R.N. and S.D. Johnson, *Personality as a Predictor of College Performance*. Educational and Psychological Measurement, 1995. **55**(2): p. 177-185.
108. Larose, S., et al., *Nonintellectual Learning Factors as Determinants for Success in College*. Research in Higher Education, 1998. **39**(3): p. 275-297.
109. Pederson, L.G., *The Correlation of Partial and Total Scores of the Scholastic Aptitude Test of the College Entrance Examination Board with Grades in Freshman Chemistry*. Educational and Psychological Measurement, 1975. **35**: p. 509-511.
110. Pickering, M., *Helping the High Risk Freshman Chemist*. Journal of Chemical Education, 1975. **52**(8): p. 512-514.

www.manaraa.com

111. Bender, D.S. and L. Milakofsky, *College Chemistry and Piaget: The Relationship of Aptitude and Achievement Measures*. Journal of Research in Science Teaching, 1982. **19**(3): p. 205-216.
112. Craney, C.L. and R.W. Armstrong, *Predictors of Grades in General Chemsitry for Allied Health Students*. Journal of Chemical Education, 1985. **62**(2): p. 127-129.
113. Nordstrom, B.H. *Predicting Performance in Freshman Chemistry*. in *American Chemical Society*. 1990. Boston, Massachusettes: ERIC.
114. Carmichael, J.W.J., et al., *Predictors of First-Year Chemistry Grades for Black Americans*. Journal of Chemical Education, 1986. **63**(4): p. 333-336.
115. House, J.D., *Noncognitive Predictors of Achievement in Introductory College Chemistry*. Research in Higher Education, 1995. **36**(4): p. 473-490.
116. Ozsogomonyan, A. and D. Loftus, *Predictors of General Chemistry Grades*. Journal of Chemical Education, 1979. **56**(3): p. 173-175.
117. Yager, R.E., B. Snider, and J.S. Krajcik, *Relative Success in College Chemistry for Students who Experienced a High-School Course in Chemistry and Those Who Did Not*. Journal of Research in Science Teaching, 1988. **25**(5): p. 387-396.
118. Lawson, A.E. and W.T. Wollman, *Encouraging the Transition from Concrete to Formal Cognitive Functioning - An Experiment*. Journal of Research in Science Teaching, 1976. **13**(5): p. 413-430.
119. Adey, P.S. and M. Shayer, *Accelerating the Development of Formal Thinking in Middle and High School Students*. Journal of Research and Development in Education, 1990. **27**(3): p. 267-285.
120. Shayer, M. and P.S. Adey, *Accelerating the Development of Formal Thinking in Middle and High School Students II: Postproject Effects on Science Achievement*. Journal of Research in Science Teaching, 1992. **29**(1): p. 81-92.
121. Shayer, M. and P.S. Adey, *Accelerating the Development of Formal Thinking in Middle and High School Students III: Testing the Permanency of Effects*. Journal of Research in Science Teaching, 1992. **29**(10): p. 1101-1115.
122. Shayer, M. and P.S. Adey, *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, 1993. **30**(4): p. 351-366.
123. Roadrangka, V., R.H. Yeany, and M.J. Padilla. *The construction and validation of Group Assessment of Logical Thinking (GALT)*. in *Annual Meeting of the National Association of Research in Science Teaching*. 1983. Dallas.
124. Tobin, K.G. and W. Capie, *The Development and Validation of a Group Test of Logical Thinking*. Educational and Psychological Measurement, 1981. **41**: p. 413-423.
125. Staver, J.R. and D.L. Gabel, *The Development and Construct Validation of A Group Administered Test of Formal Thought*. Journal of Research in Science Teaching, 1979. **16**(6): p. 535-544.
126. Treagust, D.F., *Development and Use of Diagnostic-Tests to Evaluate Student Misconceptions in Science*. International Journal of Science Education, 1988. **10**(2): p. 159-169.

www.manaraa.com

127. Yarroch, W.L., *The Implication of Content Versus Item Validity on Science Tests*. Journal of Research in Science Teaching, 1991. **28**(7): p. 619-629.
128. Williamson, V.M. and M.R. Abraham, *The Effects of Computer Animation on the Particulate Mental Models of College Chemistry Students*. Journal of Research in Science Teaching, 1995. **32**(5): p. 521-534.
129. Haidar, A.H. and M.R. Abraham, *A Comparison of Applied and Theoretical Knowledge of Concepts Based on the Particulate Nature of Matter*. Journal of Research in Science Teaching, 1991. **28**(10): p. 919-938.
130. Lawson, A.E., *The Development and Validation of a Classroom Test of Formal Reasoning*. Journal of Research in Science Teaching, 1978. **15**(1): p. 11-24.
131. *About ACS*. 2004, American Chemical Society.
132. Pickering, M., *Further Studies on Concept Learning versus Problem Solving*. Journal of Chemical Education, 1990. **67**(3): p. 254-255.
133. Sawrey, B.A., *Concept Learning versus Problem Solving: Revisited*. Journal of Chemical Education, 1990. **67**(3): p. 253-254.
134. Nakhleh, M.B., *Are Our Students Conceptual Thinkers or Algorithimic Problem Solvers?* Journal of Chemical Education, 1993. **70**(1): p. 52-55.
135. Nakhleh, M., K.A. Lowrey, and R.C. Mitchell, *Narrowing the Gap between Concepts and Algorithms in Freshman Chemistry*. Journal of Chemical Education, 1996. **73**(8): p. 758-762.
136. Cohen, J., et al., *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences*. Third ed. 2003, Mahwah, NJ: Lawrence Erlbaum Associates, Inc. 703.
137. Carroll, J.B., *The Nature of the Data, or How to Choose a Correlation Coefficient*. Psychometrika, 1961. **26**(4): p. 347-372.
138. Herron, J.D., *Piaget for Chemists*. Journal of Chemical Education, 1975. **52**: p. 146-150.
139. Sanger, M.J. and T.J. Greenbowe, *Addressing student misconceptions concerning electron flow in aqueous solutions with instruction including computer animations and conceptual change strategies*. International Journal of Science Education, 2000. **22**(5): p. 521-537.
140. Wu, H.-K., J.S. Krajcik, and E. Soloway, *Promoting Understanding of Chemical Representations: Students' Use of a Visualization Tool in the Classroom*. Journal of Research in Science Teaching, 2001. **38**(7): p. 821-842.
141. Shayer, M. and P.S. Adey, *Cognitive Acceleration Comes of Age*, in *Learning Intelligence: Cognitive Acceleration Across the Curriculum from 5 to 15 Years*, M. Shayer and P.S. Adey, Editors. 2002, Open University Press: Buckingham. p. 1-17.
142. Tien, L.T., V. Roth, and J.A. Kampmeier, *A Course to Prepare Peer Leaders to Implement a Student-Assisted Learning Method*. Journal of Chemical Education, 2004. **81**(9): p. 1313-1321.
143. BouJaoude, S., S. Salloum, and F. Abd-El-Khalick, *Relationships between selective cognitive variables and students' ability to solve chemistry problems*. International Journal of Science Education, 2004. **26**(1): p. 63-84.

www.manaraa.com

144. Novak, J.D., *Results and Implications of a 12-Year Longitudinal Study of Science Concept Learning*. Research in Science Education, 2005. **35**(1): p. 23-40.
145. Schoenfeld, A.H., *Making Sense of "Out Loud" Problem Solving Protocols*. The Journal of Mathematical Behavior, 1985. **4**: p. 171-191.
146. McMurry, J. and R.C. Fay, *Chemistry*. Fourth ed. 2004, Upper Saddle River: Prentice Hall. 1070.
147. Abraham, M.R. and V.M. Williamson, *A Cross-Age Study of the Understanding of Five Chemistry Concepts*. Journal of Research in Science Teaching, 1994. **31**(2): p. 147-165.
148. Lawson, A.E., et al., *Development of Scientific Reasoning in College Biology: Do Two Levels of General Hypothesis-Testing Skills Exist?* Journal of Research in Science Teaching, 2000. **37**(1): p. 81-101.
149. Shayer, M., *Not just Piaget; not just Vygotsky, and certainly not Vygotsky as alternative to Piaget*. Learning and Instruction, 2003. **13**: p. 465-485.
150. Zoller, U. and Y.J. Dori, *Algorithmic, LOCS and HOCS (chemistry) exam questions: performance and attitudes of college students*. International Journal of Science Education, 2002. **24**(2): p. 185-203.
151. Ebenezer, J.V. and G. Erickson, *Chemistry Students' Conceptions of Solubility: A Phenomenography*. Science Education, 1996. **80(2)**: p. 181-201.
152. National Science Board, *Science and Engineering Indicators 2006*. 2006, National Science Foundation, Division of Science Resources Statistics.
153. Torres, H.N. and D.L. Zeidler, *The Effects of English Language Proficiency and Scientific Reasoning Skills on the Acquisition of Science Content Knowledge by Hispanic English Language Learners and Native English Language Speaking Students*. Electronic Journal of Science Education, 2002. **6**(3): p. 1c-59c.
154. Biggs, J.B., *Study Process Questionnaire Manual*. 1987: Australian Council for Educational Research.
155. Trigwell, K. and M. Prosser, *Improving the quality of student learning: the influence of learning context and student approaches to learning on learning outcomes*. Higher Education, 1991. **22**: p. 251-266.
156. Biggs, J.B., *Individual Differences in Study Processes and the Quality of Learning Outcomes*. Higher Education, 1979. **8**: p. 381-394.
157. Biggs, J.B., *The revised two-factor Study Process Questionnaire: R-SPQ-2F*. British Journal of Educational Psychology, 2001. **71**: p. 133-149.
158. Zoller, U., et al., *Success on Algorithmic and LOCS vs. Conceptual Chemistry Exam Questions*. Journal of Chemical Education, 1995. **72**(11): p. 987-989.
159. Trigwell, K. and R. Sleet, *Improving the Relationship Between Assessment Results and Student Understanding*. Assessment and Evaluation in Higher Education, 1990. **15**(3): p. 190-197.
160. Trigwell, K., M. Prosser, and F. Waterhouse, *Relations between teachers' approaches to teaching and students' approaches to learning*. Higher Education, 1999. **37**: p. 57-70.
161. Adenderfer, M.S. and R.K. Blashfield, *Cluster Analysis*. Quantitative Analysis in the Social Sciences, ed. R.G. Niemi. Vol. 44. 1984, Beverly Hills: Sage Publications. 87.

www.manaraa.com

162. Scouller, K., *The influence of assessment method on students' learnign approaches: Mutliple choice question examination versus assignment essay*. Higher Education, 1998. **35**: p. 453-472.
163. Zoller, U. and B.-C. D., *Interaction Between Examination Type, Anxiety State, and Academic Achievement in College Science; An Action-Oriented Research*. Journal of Research in Science Teaching, 1988. **26**(1): p. 65-77.
164. Willis, D., *Learning and Assessment: exposing the inconsistencies of theory and practice*. Oxford Review of Education, 1993. **19**(3): p. 383-402.
165. Skogsberg, K. and M. Clump, *Do Psychology and Biology Majors Differ in their Study Processes and Learning Styles?* College Student Journal, 2003. **37**(1): p. 27-33.
166. Scouller, K. and M. Prosser, *Students' Experiences in Studying Multiple Choice Question Examinatinos*. Studies in Higher Education, 1994. **19**(3): p. 267-279.
167. Biggs, J.B., *Learning Strategies, Student Motivation Patterns, and Subjectively Perceived Success*, in *Cognitive strategies and educational performance*, J.R. Kirby, Editor. 1984, Academic Press, Inc.: Orlando, FL. p. 111-136.
168. Lewis, S.E. and J.E. Lewis, *Effectiveness of a Workshop To Encourage Action: Evaluation from a Post-Workshop Survey*. Journal of Chemical Education, 2006. **83**(2): p. 299-304.
169. Perkins, D.N. and G. Salomon, *Are Cognitive Skills Context-Bound?* Educational Researcher, 1989. **18**(1): p. 16-25.

www.manaraa.com

Appendices

www.manaraa.com

Appendix A: Commonly Used Acronyms

Table A.1 – Description of Commonly Used Acronyms

Acronym	Name
PLGI	Peer-Led Guided Inquiry
MSAT	Math portion of the SAT exam
VSAT	Verbal portion of the SAT exam
SATAVG	Class average score on the SAT measure
HLM	Hierarchical Linear Models
ACS	American Chemical Society
ACS Exam	American Chemical Society First Semester General Chemistry (Special) Examination
TOLT	Test of Logical Thinking
TOLT4AVG	Class average of students scoring over 4 on TOLT
SPQ	Study Processes Questionnaire

www.manaraa.com

Appendix B: Institutional Review Board Approval

EXEMPTION CERTIFICATION

MEMO: Scott Lewis, MA
Chemistry
SCA 400

FROM: Institutional Review Board, PGS/cas

SUBJECT: Exemption Certification for Protocol No. 101507

DATE: June 18, 2003

On June 17, 2003, it was determined that your project entitled, "Sustainable Reform for General Chemistry: Phase Implementation of Lecture-Based Reforms and Peer-Led Guided Inquiry", meets federal criteria to qualify as an exempt study.

Because the study has been certified as exempt, you will not be required to complete continuation or final review reports. However, it is your responsibility to notify the IRB prior to making any changes to the study. Please note that changes made to an exempt protocol may disqualify it from exempt status and may require an expedited or full review.

The Division of Research Compliance will hold your exemption application for five years. At least 90 days before the end of the fifth year, you will be notified that your file will be closed. If your project is still ongoing, you will need to contact the Division of Research Compliance upon receipt of that letter and follow the instructions for completing a new exemption application. It is, therefore, important that you keep your address current with the Division of Research Compliance.

If you have any questions, please contact the Division of Research Compliance at 813-974-5638.

cc: Dr. Lewis
FAO (NSF)

Office of Research, Division of Research Compliance
Institutional Review Boards, FWA No. 00001669
University of South Florida • 12901 Bruce B. Downs Blvd., MDC035 • Tampa, Florida 33612-4799
(813) 974-5638 • FAX (813) 974-5618

The University of South Florida is an Affirmative Action/Equal Access/Equal Opportunity Institution

www.manaraa.com

Appendix B: Continued

IRB Approval
FWA 00001669
IRB Number: 101507
From 04-05-2005
Thru 02-07-2006

Social Sciences/Behavioral Adult Informed Consent

University of South Florida

Information for People Who Take Part in Research Studies

The following information is being presented to help you decide whether or not you want to be a part of a minimal risk research study. Please read carefully. If you do not understand anything, ask the Person in Charge of the Study.

Title of Study:	Sustainable Reform for General Chemistry: Phased Implementation of Lecture-based Reforms and Peer-Led Guided Inquiry
Principal Investigator:	Scott Lewis, graduate student, with Dr. Jennifer Lewis
Study Location:	University of South Florida

You are being asked to participate because your experiences in chemistry courses will assist us in developing better methods for chemistry instruction both at your institution and across the nation.

General Information about the Research Study

The purpose of this research study is to gather information about the effectiveness of your General Chemistry I course. Your responses during the interview process will allow us to make recommendations for improvements in the course. Our goal is to help instructors make chemistry more understandable to students.

Plan of Study

- Nothing is required of you beyond discussing selected topics during a scheduled interview period, not to last longer than 30 minutes. The topics discussed during the interview will be related to your General Chemistry I experience. With your permission, this interview will be audio taped.
- **Payment for Participation**
  You will be paid $10 for participation in this study.

Benefits of Being a Part of this Research Study

- Although you will not directly benefit, by taking part in this research, you will contribute to improvements in chemistry instruction in the course you have taken.

Risks of Being a Part of this Research Study

- The only risk we have been able to identify is possible discomfort at being part of a group that is being studied. We assure you that you have been chosen to be a part of this study on the basis of your knowledge of the course in question and your willingness to participate.

Confidentiality of Your Records

- Your privacy and research records will be kept confidential to the extent of the law. Authorized research personnel, employees of the Department of Health and Human Services and the USF Institutional Review Board and its staff, and other individuals acting on behalf of USF may inspect the records from this research project.

The results of this study may be published. However, the data obtained from you will be combined with data from other people in the publication. The published results will not include your name or any other information that would in any way personally identify you. The information you provide will be seen only by authorized research personnel such as Dr. Jennifer Lewis and the graduate student in charge of the project and will be stored in a locked research laboratory.

IRB#_____

Rev 9/99

www.manaraa.com

Appendix B: Continued

Volunteering to Be Part of this Research Study

- Your decision to participate in this research study is completely voluntary. You are free to participate in this research study or to withdraw at any time. If you choose not to participate, or if you withdraw, there will be no penalty or loss of benefits that you are entitled to receive.

Questions and Contacts

- If you have any questions about this research study, contact Dr. Jennifer Lewis or Scott Lewis at 813-974-1286. You may also e-mail Dr. Lewis at jlewis@chuma1.cas.usf.edu or Scott Lewis at selewis@eng.usf.edu

- If you have questions about your rights as a person who is taking part in a research study, you may contact a member of the Division of Research Compliance of the University of South Florida at 813-974-5638.

Your Consent—By signing this form I agree that:

- I have fully read or have had read and explained to me this informed consent form describing a research project.
- I have had the opportunity to question one of the persons in charge of this research and have received satisfactory answers.
- I understand that I am being asked to participate in research. I understand the risks and benefits, and I freely give my consent to participate in the research project outlined in this form, under the conditions indicated in it.
- I have been given a signed copy of this informed consent form, which is mine to keep.

Signature of Participant	Printed Name of Participant	Date

Investigator Statement

I have carefully explained to the subject the nature of the above protocol. I hereby certify that to the best of my knowledge the subject signing this consent form understands the nature, demands, risks and benefits involved in participating in this study.

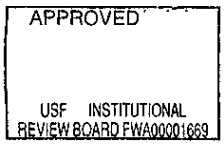
Signature of person obtaining consent Or Authorized research investigators designated by the Principal Investigator	Printed Name of person obtaining consent	Date

Institutional Approval of Study and Informed Consent

This research project/study and informed consent form were reviewed and approved by the University of South Florida Institutional Review Board for the protection of human subjects. This approval is valid until the date provided below. The board may be contacted at (813) 974-5638.

Approval Consent Form Expiration Date:

Revision Date:_____



IRB#_____

Rev 9/99

www.manaraa.com

Appendix C: First Day Survey

Please fill out your name, social security number and your answers to the following questions on a scan-tron bubble sheet. The scan-tron 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

www.manaraa.com

Appendix C: Continued

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

www.manaraa.com

Appendix D: Standardized Growth Model

Since the tests from one semester to another were different, the combination of test scores across semester may be problematic. For example, if the first test in each semester varied in difficulty, a 60% correct score on one test could mean a relatively high score in one semester or a relatively low score a different semester. One way to account for this possibility is to standardize the test for each semester. This way, each test in each semester has an average score of zero and a standard deviation of one. And correspondingly, a value of 0.2 on a test would mean 0.2 standard deviations above the class average of those who took that same test.

With this transformation made, an HLM was run using identical decisions as before with the unstandardized data. The full equation again suggested removing the classroom effects on individual SAT scores, suggesting little impact on equity as a result of the reform. With these coefficients removed, the model was run again, resulting in the coefficients in Tables A.2 through A.3.

www.manaraa.com

Appendix D: Continued

Table A.2 - Estimating the Intercept Coefficient (π_{0jk})

Symbol	Description	Estimate	Std. Error	Sig.
γ_{000}	Intercept	-0.006588	0.01874	n.s.
γ_{001}	Class SAT	0.003461	0.000911	<0.001
γ_{002}	PLGI	-0.02964	0.04561	n.s.
β_{01k}	Student Math SAT	0.005129	0.000254	<0.001
β_{02k}	Student Verbal SAT	0.001191	0.000255	<0.001

n.s. = non significant ($p > 0.050$)

Table A.3 - Estimating the Slope Coefficient (π_{1jk})

Symbol	Description	Estimate	Std. Error	Sig.
γ_{100}	Intercept	-0.05076	0.00795	<0.001
γ_{101}	Class SAT	-0.0003662	0.000397	n.s.
γ_{102}	PLGI	0.07316	0.01963	<0.001
β_{11k}	Student Math SAT	-0.00053	0.000110	<0.001
β_{12k}	Student Verbal SAT	0.000047	0.000109	n.s.

n.s. = non significant ($p > 0.050$)

The discussion will begin with Table A.2, which indicates the effect of these factors on Test 1 performance. First note the intercept is approximately zero, which is a direct result of the standardization procedure. However, apart from this change, the significance test of each parameter is similar to Table 2.8. The PLGI coefficient is negative, indicating that students in the PLGI section scored lower than their non-PLGI counterparts on the first test; though this difference is non-significant and can be attributed to chance.

Table A.3 indicates how the standardized scores change over time. Similar to Table 2.9, the intercept in this table is significant and negative, and the PLGI variable is significant and positive. In Table 2.9 the PLGI coefficient of 1.55 is just over half of the

www.manaraa.com

intercept value –3.08. However in the standardized values of Table A.3 the PLGI coefficient is actually larger than the intercept coefficient. This is a direct result of the standardizing procedure, which keeps the average for each test at zero. Such that when the PLGI reform produces a positive effect, resulting in PLGI students scoring in general above average, the non-PLGI students (reflected by the intercept) have to demonstrate a negative value in order to keep the average at zero. The relative weights of the coefficients are a result of the sample size of PLGI versus non-PLGI rather than an indication of the difference. The coefficients in Table 2.9 are more readily applicable in making this type of comparison.

However, the standardized values in Table A.3 can provide some descriptive value which isn't present in Table 2.9. The coefficients indicate the number of standard deviations above or below average for each group. Finding the score expected for a student in PLGI with SAT scores equal to the class average on test 4 (time = 3), the score would be 0.19 standard deviations above average, where a similar student without PLGI would be 0.16 standard deviations below average.

In the unstandardized model it was found that student SAT scores had a strong impact on the first test, but in terms of the slope coefficient that represents time student SAT scores did not impact the scores. This model features a similar result, with the exception of a significant negative student Math SAT score relating to the slope coefficient. This may be an indication of student drop-out in the semester disproportionately occurring for students with low SAT scores. As a result, students with

www.manaraa.com

Appendix D: Continued

higher SAT scores that remain would be approaching the class average. Nonetheless, the impact is also minimal, reducing the effect of Math SAT from an original value of 0.005129 to a Test 4 value of 0.003554, such that Math SAT still plays a role in course performance throughout the semester. Also note this additional significant coefficient does not impact the interpretation of the effect of the reform, and like before the reform was found to demonstrate effectiveness in the overall class performance, but no noticeable effect on the equity in the classroom that results from student SAT scores.

www.manaraa.com

About the Author

Scott Edwin Lewis was a member of the charter class of the International Baccalaureate program at King High School, Tampa, Florida, graduating in 1997. At the University of South Florida in 2001 he earned a Bachelor's of Science degree in Chemical Engineering, summa cum laude. During this year he was named co-student of the year in Inorganic Chemistry. Also at the University of South Florida in 2003 he earned a Master of Arts in Chemistry, with an emphasis in Chemical Education in detailing the effectiveness of several pedagogical workshops. This work has appeared in the *Journal of Chemical Education* along with other articles he has contributed on evaluating pedagogical reform. He has a faculty position at Kennesaw State University.

www.manaraa.com