7-1-2015

Defining and Assessing Problem Solving Across a Biochemistry Curriculum

Cheryl A. Sensibaugh

Cheryl A. Sensibaugh
Candidate

Biomedical Sciences
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Marcy Osgood, PhD, Chairperson

Steven Mitchell, MD

Erika Offerdahl, PhD

Jay Parkes, PhD

DEFINING AND ASSESSING PROBLEM SOLVING ACROSS A BIOCHEMISTRY CURRICULUM

by

CHERYL A. SENSIBAUGH

BS, Biochemistry, University of New Mexico, 2007

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

Biomedical Sciences

The University of New Mexico
Albuquerque, New Mexico

July, 2015

ACKNOWLEDGMENTS

First and foremost, I thank my advisor, committee members, and other scholarly educators too numerous to name for their unending support in pioneering doctoral training in discipline-based education research under the umbrella of the Biomedical Sciences Graduate Program at the University of New Mexico. Each of you played a role in building new paths in our community. I applaud your extensions above and beyond our work together, which truly would not have been possible without your advocacy.

Marcy Osgood and Bill Anderson, your influence as professional role models and researchers extends back to my undergraduate years. First, you introduced me to the wonders of biochemistry, and then to all of the potential and promise of biochemistry education. Your support and guidance has meant so much to my success. Thank you for helping me learn how I can best contribute to society, and for helping me determine my own most fulfilling career.

Erika Offerdahl, you were instrumental in my research process about pursuing doctoral research in biochemistry education. The experience you gained building your own new paths has served me well. I thank you for being one who paved the way before me. Steve Mitchell, you taught me so much about helping students learn, in ways that can only happen in a classroom. You also broadened my horizons into medical education, which grounded me in my own discipline by providing the perspective to compare and contrast the two settings. Jay Parkes, your expertise in educational research was invaluable. You served as a critical link in helping me form the connections necessary for biochemistry education research. I have been fortunate, indeed, for mentorship from all of you.

My thanks also goes to many others. At Tennessee Technological University, Barry Stein, Kevin Harris, and colleagues provided untold hours of scoring and compiling data for analysis. Barry, special thanks for your insights on my findings and your perspective on interesting new questions to ask. Jenny Knight at the University of Colorado at Boulder, I appreciate your discussion of the CLASS-Bio and support of its use. Without the work of others who develop instruments necessary for measuring the outcomes of teaching and learning, our understanding of the broader educational picture would quickly become limited.

My never ending gratitude goes to those who supported me in ways both large and small. Martina, you made teaching fun, and even funny sometimes. Thanks for your comic relief to make the tough times more bearable. Ellis, here's to your introductions to all the new people and

experiences, which have contributed greatly to my success. Douglas, Janice, and all women with wisdom, you have been my champions, and I can hear your cheers. Jess, you've always been there, through thick and thin, with your heart as big as the Sandia Mountains. Stacy, Tommy, Donny, Lori, Kim, and Dawn, *now* you can call me doctor. Our friendships have stood the test of time, and I treasure them always.

DEFINING AND ASSESSING PROBLEM SOLVING ACROSS A BIOCHEMISTRY CURRICULUM

by

Cheryl A. Sensibaugh

BS, BIOCHEMISTRY

PhD, BIOMEDICAL SCIENCES

ABSTRACT

Undergraduate discipline-based education research has shown that scientific problem solving involves five domains, spanning the steps of the scientific method as well as metacognition: Hypothesize, Investigate, Evaluate, Integrate, and Reflect. Student performance in each domain is measured with the Individual Problem Solving Assessment (IPSA). Others developed the Critical thinking Assessment Test (CAT) to measure critical thinking, with the view that problem solving is a component of critical thinking. A third group took the perspective that student attitudes about learning science will influence performance, and developed the Colorado Learning Attitudes about Science Survey for Biology (CLASS-Bio) to detect student attitudes.

This study employed a framework of constructivism, cognitive dissonance, and scientific teaching to address the educational problem of facilitating process-oriented skills within an upper-level biochemistry curriculum. During IPSA development, research goals centered on establishing instrument validity and reliability, as well as describing typical ranges of individual student performance in each domain of problem solving across the junior year of our biochemistry curriculum. The evidence indicated that students could struggle in any IPSA domain, even after two semesters of deliberate practice of problem solving.

The next goal was to describe average performance across the junior and senior years of a biochemistry curriculum, and explain score variability using hierarchical linear regression to account for contributions from time, academic factors, and demographic factors. The average student required two semesters to achieve the objectives for three domains of problem solving,

two years for Evaluate, but did not achieve the Investigate objective. Regression equations explained that time, critical thinking, and learning attitudes promoted performance, yet in different ways across domains. Based on these results, our main pedagogical recommendation is to model and scaffold the problem solving process.

Finally, we initiated a nomological network, or representation of relatedness among problem solving (IPSA), critical thinking (CAT), learning attitudes (CLASS-Bio), and biochemistry content knowledge (course exams), to visualize relationships among alternative perspectives of defining and assessing problem solving. Score correlations determined that the three process-oriented assessments converged when asking students to form a conclusion, weakly converged with content knowledge, and diverged from content when measuring metacognition and critical thinking.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1
Introduction

1.1 Study significance

The field of discipline-based education research (DBER) is grounded in a wide range of disciplines, including STEM (science, technology, engineering, and mathematics), educational psychology, and cognitive science. According to a consensus report commissioned by the National Research Council (2012), DBER addresses research questions about learning and teaching within those disciplines. Furthermore, high-quality DBER requires expertise in three areas: the core discipline, learning and teaching in the discipline, and the science of learning and teaching. All these areas are complex; therefore, collaborations among researchers with specific areas of expertise are recommended. This work takes place within the core discipline of biochemistry, and is termed biochemistry education research.

Given the importance of training biochemists in the processes of scientific problem solving and discovery, as set forth by multiple broad communities of life science researchers and educators (American Association for the Advancement of Science, 2011; American Society for Biochemistry and Molecular Biology, 2012; Association of American Medical Colleges and AAMC-HHMI Committee, 2009), this study is significant in three main ways. First, the study develops and validates a novel approach to assessing student performance in problem solving. Secondly, this work describes discipline-specific performance across undergraduate junior and senior years, and investigates putative contributors to performance, in an attempt to explain the observed performance. Finally, the project generates new understandings of the relationships between biochemistry content knowledge, problem solving, critical thinking, and learning attitudes about science.

1.2 Rationale and Experimental Design for Research Goal I

Goal I: To develop and validate an assessment tool that measures student performance in scientific problem solving.

Research question I.A: How closely related are the scores in each domain of problem solving to each other and to scores on content knowledge?

Specific aim I.A.1: To determine the degree of correlation between IPSA domain scores and content exam scores.

Evidence: Pearson correlation coefficients (*r*)

Method: Score correlation analysis

Reasoning: Students who are novice problem solvers could be quite skillful in one domain, while not demonstrating proficiency in others. Additionally, content knowledge is integral to the tasks of stating results in the Evaluate domain, and of forming conclusions in the Integrate domain.

Hypothesis I.A: Scores in domains that test independent problem solving skills (*i.e.*, Hypothesize, Investigate, and Evaluate) will not correlate with each other or with scores of content knowledge. However, scores in domains that are dependent upon content knowledge (*i.e.*, Evaluate and Integrate) will correlate with scores of content knowledge and with each other.

Research question I.B: Do the IPSA scoring rubrics promote consistent scoring, among faculty raters as well as graduate students?

Specific aim I.B.1: To determine the degree of correlation between IPSA domain scores assigned by three different faculty raters and one graduate student.

Evidence: Pearson correlation coefficients (*r*)

Method: Score correlation analysis

Reasoning: Rubrics that are clear enough to be applied by diverse experts will yield consistent scores, regardless of rater identity.

Hypothesis I.B: Scores assigned by different raters will correlate strongly, with an *r* value greater than 0.5.

1.3 Rationale and Experimental Design for Research Goal II

Goal II: To describe individual student performance in problem solving, when no longer collaborating in the setting of a group activity.

Research question II.A: Which domains of problem solving are most challenging for students to complete individually?

Specific aim II.A.1: To identify the domains in which student performance is consistently less than acceptable.

Evidence: Domain scores of individual students

Method: Visual depictions of representative individual performance patterns

Reasoning: Report common patterns in terms of whether domain scores were satisfactory or not, to maintain student confidentiality.

Hypothesis II.A: Consistent with our findings during development of group activities to promote problem solving skills, the strengths and weaknesses of individual students will also be apparent in domain scores, but will vary according to student.

Research question II.B: Are intervention strategies necessary for all students, when a two-semester course sequence incorporates multiple opportunities over time for deliberate practice of problem solving?

Specific aim II.B.1: To determine whether students are able to implement successful problem solving strategies on their own, over time.

Evidence: Representative longitudinal performance patterns

Method: Summarize longitudinal performance of individuals

Reasoning: Some students will be able to reach acceptable performance.

Hypothesis II.B: Not all students will require intervention to reach acceptable performance over time.

1.4 Rationale and Experimental Design for Research Goal III

Goal III: To quantitatively describe and explain student performance in problem solving across a two-year biochemistry curriculum.

Research question III.A: What is the longitudinal performance pattern of an average student, and when do most students begin maintaining satisfactory performance in each domain?

Specific aim III.A.1: Describe average student performance over two years.

Specific aim III.A.2: Describe rates of satisfactory performance over two years.

Evidence: Longitudinal IPSA domain score means and achievement rates

Method: Descriptive statistics

Reasoning: Taken together, previous results on the ranges of individual performance along with results from a small pilot study reporting on means and achievement rates inform the hypothesis.

Hypothesis III.A: The average student struggles in all domains then improves to satisfactory performance over time, the duration of which varies by domain. Consistent satisfactory performance by most students (> 50%) begins in the second semester, but not in all domains.

Research question III.B: Which contributors – among time, academic background, and demographic background – most consistently explain the observed problem solving performance, and how much do the contributors impact performance?

Specific aim III.B.1: Explain the impact of various contributors to student performance

Evidence: Multivariate linear regression equations

Method: Multivariate Hierarchical Linear Regression

Reasoning:

Hypothesis III.B: The greatest contributor to IPSA domain performance is time, followed by academic backgrounds of students, then by demographic backgrounds. Even in sum, these contributors will explain less than half of the variability in scores across students.

1.5 Rationale and Experimental Design for Research Goal IV

Goal IV: To initiate an understanding of the nomological network of various problem solving skills and biochemistry content knowledge.

Research question IV.A: How closely related are the observable scores of problem solving, critical thinking, and learning attitudes about science, to each other as well as to scores of content knowledge?

Specific aim IV.A.1: Determine the degree of convergence and divergence among assessment scores.

Evidence: Pearson correlation coefficients (*r*)

Method: Score correlations

Reasoning: Scores from assessments that measure similar constructs would be expected to correlate at least moderately.

Hypothesis IV.A: IPSA Hypothesize, Investigate, Evaluate, and Integrate domain scores will correlate at least moderately with CAT scores, while IPSA Reflect domain scores will correlate at least moderately with CLASS-Bio scores. Previous findings also indicate that the IPSA Evaluate and Integrate scores will correlate with content exam scores.

Chapter 2

What really matters: Assessing individual problem-solving performance in the context of biological sciences

Steven M. Mitchell[1], William L. Anderson[2], Cheryl A. Sensibaugh[2], and Marcy Osgood[2]

[1] School of Medicine, University of New Mexico, Albuquerque, NM, USA

[2] Department of Biochemistry and Molecular Biology, University of New Mexico, Albuquerque, NM, USA

2.1 Abstract

The evaluation of higher-level cognitive skills can augment traditional discipline-based knowledge testing by providing timely assessment of individual student problem-solving abilities that are critical for success in any professional development program. However, the wide-spread acceptance and implementation of higher level cognitive skills analysis has been delayed by the lack of rapid, valid, and reliable quantified-scoring techniques. At the University of New Mexico School of Medicine, Department of Biochemistry & Molecular Biology, we have developed an examination format that can be routinely and sequentially implemented for both formative and summative assessments of individual students in large classes. Rather than providing results in terms of an individual student's knowledge base in a single academic discipline or group of disciplines, this type of examination provides information on performance in the application of specific problem-solving skills, which we term "domains," to a contextual clinical or scientific problem. These domains, derived from the scientific method, are tested across various academic disciplines, and are reported in terms of the following: Initial and sequential hypothesis generation, investigation of these hypotheses, evaluation of newly acquired data, integration of basic science mechanisms with new information to explain the basis of the problem, and reflection on one's own professional development in the context of the examination. The process for criterion- referenced quantified grading of the examination is outlined in this paper. This process involves relatively rapid scoring, and permits the timely use of the resulting information for individual student feedback as well as curricular improvement. Data regarding grading consistency and comparison with other measures of student performance is also presented in this paper. An analysis of the performance characteristics of this examination, which has been utilized for over 10 years in a variety of course settings, indicates that it is valid, reliable, and utilizable. As such, the methodology is now routinely used in several undergraduate and graduate level biochemistry classes to monitor the development of individual student problem-solving abilities.

Keywords: Problem-solving, critical-thinking, evaluation, assessment, performance.

2.2 Introduction

In 2003, the American Society of Biochemistry and Molecular Biology (ASBMB) published a recommended curriculum for undergraduate biochemistry and molecular biology students. A significant distinction of this curriculum was the inclusion of skills- or process-based learning objectives, in addition to the more traditional requirement for students to master a body of content knowledge. While content-oriented knowledge reflects the body of facts learned about a subject, process-oriented knowledge reflects the ability to apply content knowledge within a contextual situation (Mayer, 2002). The ASBMB's recommendation for an undergraduate biochemistry program (ASBMB, 2003) echoed the framework for reform of science education that was outlined in the Biology 2010 report (National Research Council, 2003). And more recently, the American Association of Medical Colleges (AAMC), in conjunction with the Howard Hughes Medical Institute (HHMI), proposed specific learning objectives for both medical and pre-medical students (AAMC, 2009), reiterating the importance of teaching and assessing problem-solving skills as one of several process-based learning objectives. The underlying message of all of these reports is that, while conceptual understanding, or discipline-specific content knowledge, is clearly one part of the development of a scientist, it needs to be paired with cognitive understanding, or knowledge about the (often) discipline-specific processes that govern appropriate and successful use of content (Mayer, 2002). Even more specifically, these reports all recommend that undergraduate students in the biomedical sciences be provided routine opportunities to develop and practice their scientific problem-solving strategies.

While the requirement for students to practice their problem-solving skills is a laudable goal, in the classroom this becomes a daunting task. Moreover, this endeavor requires that the faculty both detect defective problem-solving, and provide student-specific feedback about strategies for improvement. This is feasible when a faculty member works with a limited number of students, but when an instructor is charged with implementing such an analysis and intervention strategy in large lecture classes, the job of teaching and evaluating student problem-solving rapidly becomes overwhelming. Consequently, it is not uncommon for faculty to state that, "It can't be done," and they will not even attempt any quantitative assessment of problem-solving skills, sometimes saying "I will know it when I see it," as their qualitative evaluation.

For the past 10 years, our undergraduate biochemistry students at the University of New Mexico have been required to apply their biochemistry content knowledge and concurrently practice their problem-solving strategies through online small group discussions of scientific problems (Anderson et al., 2008; Osgood et al., 2008). In these discussions, group problem-solving is routinely evaluated and the contribution of individual students to the successful solution of a biochemical dilemma can be tracked. These exercises provide students with routine opportunities to practice their problem-solving strategies; however, feedback to individual students is limited. Moreover, we have routinely observed that some students, who had appeared successful in contributing to the group solution of a biochemistry puzzle, were not subsequently able to succeed as individual problem-solvers, even when presented with very similar conceptual challenges. When such a student's contributions to the online group discussions were re-evaluated, it became evident that the student was not contributing broadly to the group solution, but instead tended to retreat to his/her "comfort zone" without confronting all aspects of an investigational strategy. We judged that it was necessary to provide regular opportunities for our students to address both group and individual problem-solving challenges within their biochemistry courses, thus encouraging them to apply the skills learned within the online group discussions to the solution of similar problems, but on their own. In order for these assignments to be useful, the assessment of the individual's problem-solving skills should provide novel information to the student that he/she can then use to successfully modify his/her own investigational strategies. This article describes the multiple iterative cycles over the 10-year development of this Individual Problem-Solving Assessment (IPSA) tool, and includes data on validation of the current version.

The authors, STEM education specialists, have been working together in biomedical sciences education for 16 years. Currently, two authors are course directors (WLA and MPO) in upper-level biochemistry classes. One author is a graduate student (CAS) focusing research efforts in biochemistry education and is responsible for facilitating small group exercises. The fourth author (SMM) is a MD who works with medical students and is also involved in the development and implementation of critical thinking exercises in both medical school and biochemistry classes.

2.3 Methods

2.3.1 Structure of an Individual Problem-Solving Assessment

The goal in this endeavor was to develop an easily implemented, reproducible method for evaluating a student's ability to apply content knowledge to the solution of a problem; in other words, this tool had to function as a novel means of evaluating process. Students should have multiple opportunities to practice their skills, succeed or fail, and then receive appropriate faculty feedback on their efforts. This iterative practice and assessment approach needed to allow students to develop a reliable and effective problem-solving strategy. The authors felt that in any problem-solving type of test, students should first, be able to learn process skills from the exam, and second, clearly see their content knowledge applied to the solution of a real-life problem. Finally, the authors wanted to ensure that any individualized problem-solving test would complement and enhance the student's small group learning experience.

The tool that was developed in this capacity is the Individual Problem-Solving Assessment (IPSA). IPSAs are provided to students electronically as multi-part, progressive-reveal essay exams, which are based on scientific dilemmas that capture student interest based on the contextuality of the problem. These scenarios are not discussed in other parts of the current course but require students to extrapolate their knowledge from online discussions, individual research, lecture material, and other components of the curriculum. The IPSAs require students to use the same problem-solving domains that are used in the online small group discussions and that are also integrated into the curriculum (Anderson et al., 2008; Osgood et al., 2008). The learning system development tool we use to construct our tests is Macromedia's Authorware©. Multiple other software packages are also potentially appropriate. Figure 2.1 schematically illustrates the structure of the IPSA scenarios. A complete IPSA, grading rubrics, one student's responses, and a corresponding visual representation of that student's performance are provided in Appendix B.

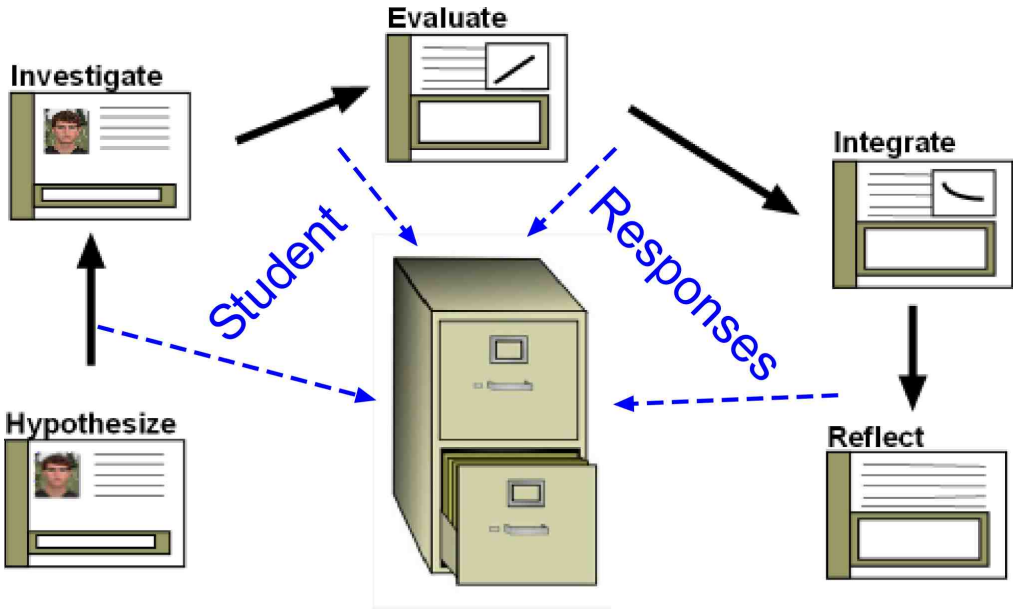


Figure 2.1: Individual Problem-Solving Assessment Structure. Each of the five domains of problem-solving are incorporated into the IPSA. To assess each domain on its own merit, student responses are collected in sequence and stored in a database. The software only allows forward progression through the assessment.

Each IPSA begins with a vague, two- to three-sentence presentation of the problem, shown on the first screen of the electronic presentation. The remainder of the exam is based on the problem-solving domains (Anderson et al., 2008) of **Hypothesize**, **Investigate**, **Evaluate**, **Integrate**, and **Reflect**. Students are directed to identify their initial **Hypotheses** as to the underlying cause of the problem, and submit that answer electronically. As the next screen comes up, students are then provided with a specific hypothesis to test, and asked to identify the data they feel would be most important to acquire in order to **Investigate** this hypothesis. After the students have submitted their answers to the **Investigate** question, they continue to be provided data in a progressive-reveal manner on successive screens and they must **Evaluate** the graphs, charts and other data in the context of the situation, while taking into account all previously acquired information about the case. Once students have attained enough information (through prompts in the exam), they are asked to **Integrate** their basic understanding of key concepts with the new knowledge presented in the IPSA scenario, and to provide a detailed description of the scientific mechanisms involved in the problem. Often, this **Integrate** challenge is presented to the students in the form of a controversy that they must resolve. Finally, students are asked to **Reflect** on their performance by generating a plan by which they can improve their own performance on later similar assessments and a strategy

for the resolution of the given problem. This is an attempt on our part to help the students develop a more metacognitive approach to their individual learning (Flavell, 1976).

The exam is structured as a progressive-reveal evaluation. Each new part of the exam is presented only after an answer to the previous question is submitted. Students are prevented from returning to a previous answer to alter it after they have accessed new information. Early on, we discovered that when students make a single mistake in answering the first or second question, it sends them in the wrong direction for the rest of the exam. Subsequent responses, although potentially correct based on the initial (wrong) answer, will earn inaccurate and low scores. To address this issue, as each new part of the exam is presented, we build in a teaching element to bring all students back on track as the case is progressively revealed.

In order to reassure ourselves that the IPSA results are truly providing novel information about student performance, we compared our problem-solving domains assessment to a classic evaluation of content knowledge. Two hundred forty first-year medical students were challenged with 6 different IPSA scenarios over a 3-year period with paper-and-pencil versions of the exams. Each of the IPSAs focused on different content. Concurrently with the IPSAs, these students were also challenged with the AAMC Shelf Boards, which are a well-established measure of content knowledge. All of the scores for each of the IPSA domains, as well as the content knowledge exam scores, were used to construct a correlation coefficient matrix.

All subsequent experiments used electronic versions of the exams.

2.3.2 Implementation of the Exam

Typically four different IPSAs were presented to a class containing 80 to 100 students during one semester. Because of computer limitations we could only accommodate 30 students per testing session, requiring the IPSAs to be scheduled over a two-day period. It was important to emphasize that the same problem-solving domains that students were practicing in the online discussion component of the course were incorporated into each IPSA, which led to a more cohesive curriculum. Although we believe that simply taking the IPSAs was instructive for our students, and was an experience that students did not typically gain from a traditional lecture-based course, we also believe in the necessity of timely

feedback on individual performance. Accordingly, all students received scores for their performance on the domains within a week of taking the exam.

2.3.3 Grading the IPSA

We typically collect student responses for each part of an IPSA electronically, and transfer the responses into a database for grading ease as depicted in Figure 2.2. The two course instructors are responsible for grading the exams and providing feedback to students as necessary.

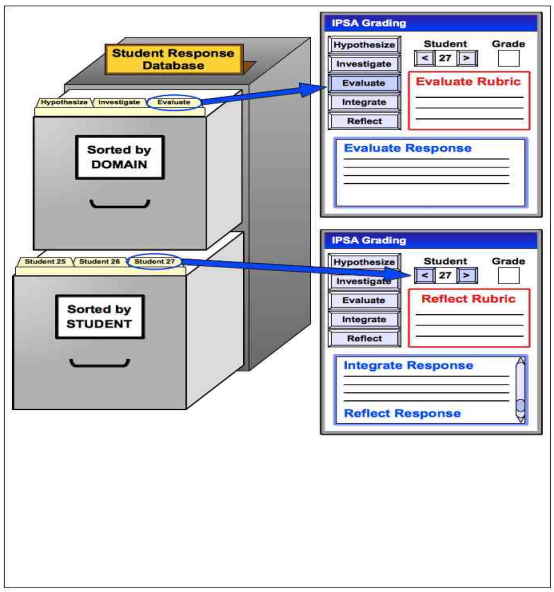


Figure 2.2: Database grading. Student responses to an IPSA may be retrieved from the database and sorted either by domain or by student. The grading rubric for each domain (red box) is also shown with the student records.

Using an electronic database to collect and grade student responses is preferable to grading hard copies because it increases speed, efficiency, and reproducibility in assigning grades. First and foremost, we can read the student responses without spending time deciphering cryptic handwriting. Moreover, we are able to limit student responses to a fixed number of characters which forces students to think first and then answer the specific question, rather than writing everything they know about the topic, hoping to produce an answer that will somehow include the correct response. Additionally, by taking advantage of

student name coding capabilities inherent to electronic databases, the element of bias is removed from scoring the essays.

Furthermore, using the database sorting capabilities, we can easily arrange responses either by domain or by students' complete responses to an IPSA as a whole (Fig. 2.2). For example, it is possible to grade a single domain for an entire class, which is typically how we grade the **Hypothesize**, **Investigate**, **Evaluate** and **Integrate** domains. In our experience this method decreases the time required for grading and improves the grading consistency. However, due to its dependence on metacognition, the **Reflect** domain must be graded in the context of all of one student's responses on that IPSA. Viewing the response in this way provides insight into the overall thinking of an individual student, which is particularly helpful when working with students who are having academic difficulty.

2.3.4 Development of Grading Rubrics

IPSAs are constructed around inherently difficult concepts and/or common misconceptions. These exams are not used for probing easily grasped items of content knowledge. The grading rubrics used to assess student performance on these complex exams thus require thoughtful development; as a result, this process is the most time-consuming and important step in the creation of an IPSA. Based on our own experience and on suggestions in the literature (Allen and Knight, 2009), we develop our grading rubrics in an iterative manner. The process involves multiple instructors, including some who are not involved in the initial construction of the IPSA scenario. In addition, upon the first use of a new IPSA, the students' domain responses to the new scenario are also used to re-evaluate both the clarity of questions and the applicability of the rubrics.

Specific rubrics are designed for each problem-solving domain. Establishment of clear benchmarks for each domain is essential for ease and accuracy in grading. We first design rubrics that delineate outstanding, acceptable, and failing performance criteria; and then assign numerical values to each of these benchmarks. As our experience with each IPSA grows, scores for performances that fall between the benchmarks are also assigned. For example, "outstanding" answers for the Hypothesize domain would include at least 3 logical, context-specific hypotheses, and be assigned a 10/10 value; an "acceptable" answer might include only two appropriate hypotheses, and be scored as a 7/10; and a "failing" answer

either misses something critical to the understanding of the concept, or includes irrelevant or factually incorrect ideas, and will earn less than a 7/10. When multiple instructors grade a student essay very differently, both the specific question and the grading rubrics are re-evaluated.

2.3.5 Evaluation of Rubrics

In order to evaluate the reproducibility and ease in applying the grading rubrics, a group of three faculty members independently graded all domain responses of 20 students in 8 IPSA scenarios over two semesters of an intensive biochemistry curriculum. All three instructors were intimately involved in the development of the questions and grading rubrics, and all had extensive prior experience in the implementation of IPSAs. The mean, standard deviation, and students t-test were used to compare the assigned grades.

In order to further probe the effectiveness of using the grading rubrics, and to determine if graduate students who are not involved in the construction of the IPSA can be reliable graders, a graduate student was provided the grading rubrics for a single IPSA and asked to grade all 5 domains for 10 different students. The graduate student was given 30 minutes training by a faculty member in the basic science of the case, and the grading rubrics were explained. Strict adherence to the rubrics was required. The student-grader was blinded to the instructor's responses and the two response sets were statistically compared as was done with the previously described faculty evaluations.

2.3.6 Reporting Data

Early in our evaluation of IPSA student data, we decided that we did not want to compress student responses on all domains into a single score. We view the **individual** use of each of the domains (**Hypothesize**, **Investigate**, **Evaluate**, **Integrate**, and **Reflect**) as integral to the overall process: Application of each of the domains must be mastered in order for a student to become a successful scientific problem-solver. Therefore, like we do in the online case discussion (Anderson et al., 2008), we score each domain separately, which creates a more complete picture of a student's problem-solving strategy. Reporting individual domain scores also provides the faculty with specific information that can be used to identify where students should focus in order to improve their skills. We present results from these

exams by using a radar plot in which each of the axes of the diagram represents the earned score within a single domain. This allows us, and our students, to see performance patterns on all five domains simultaneously. We find that students and instructors grasp a performance pattern more easily than a set of five different numerical scores. Figure 2.3 illustrates how student problem-solving domain patterns, or profiles, are depicted.

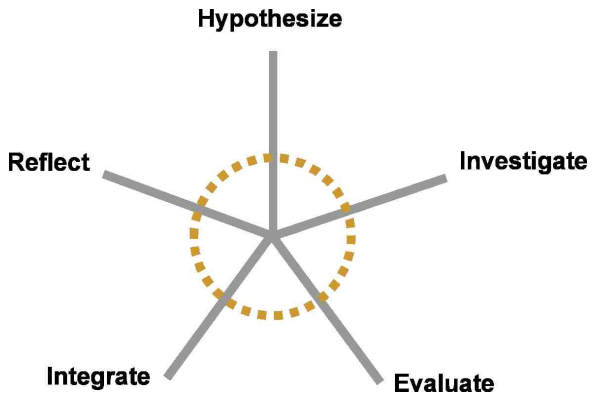


Figure 2.3: Radar Plot. A visual representation of the scoring ranges on an IPSA, with axes for each of the problem-solving domains (gray). Expected performance scores are indicated by the circular target (yellow dotted outline).

Low scores are at the periphery of the axes, and outstanding scores are in the center. Though this arrangement of scores may seem counterintuitive, we have found that students readily grasp the idea that they need to "try for the bull's-eye" in their domain scores. A circumscribing line is used to connect the domain scores between the axes to create a shape profile. The faculty expectation (the score for each domain that represents an "acceptable" grade) is indicated by the dotted circle toward the center of the diagram. Although there are other methods to report this type of data, we have found that the graphical representation shown in Fig. 2.3 is the clearest, and changes in student performance over time are readily seen as changes in the pattern, so that students and faculty alike can follow progress.

To evaluate a change in student performance over time on this type of exam, the same twenty students were evaluated with 8 different IPSA scenarios over the course of two semesters in the same undergraduate biochemistry courses that were analyzed in the evaluation of the rubrics. All student essays were independently graded by the same three instructors. In an effort to minimize the effect of content familiarity on a single question, a

rolling average of student domain scores on the most recent three exams was used for this analysis.

2.3.7 Student Populations

Two different student populations participated in these studies: 60 undergraduate biochemistry majors and 240 pre-clinical medical students. All students were experiencing a hybrid curriculum, which employed both small group cooperative-learning opportunities along with standard lecture presentations. Student populations were evenly split between male and female students and contained approximately 45% under-represented minority students. All students had successfully completed the prerequisite courses.

2.4 Results and Discussion

2.4.1 Exam Logistics

We have experimented with many different logistical ways of implementing the IPSAs that have ranged from paper and pencil execution to electronic assessment methods - either online or in a more secure computer center. All methods have worked, but we prefer the electronic format because it increases grading consistency and allows us to easily build in a teaching component into the exam.

Since IPSAs and their accompanying grading rubrics are difficult and time-consuming to construct, the exams are kept secure so that we are able to use the same IPSA for several years. However, this is a new type of exam for most of our students, and they lack experience in solving problems. Moreover, for reasons discussed previously, the online group discussions do not always allow for individual problem-solving practice. To address this issue, we typically present multiple different practice IPSAs to our students throughout their coursework, and some of these practice scenarios then serve as the conceptual basis of course lectures. We also role-model problem-solving strategies based on the practice exams in order to help the students become comfortable with the process. Even given all of this preparation for the first graded IPSA, these first exam results are usually not weighted heavily for the students' final grades as the approach to critical-thinking is often very novel to our students and may require multiple encounters in order to be conceptualized and utilized.

Depending on the pedagogical nature of the course, the number of IPSAs varies between 4 and 6 per semester. Students have one hour in a computer-testing center to complete each exam. Because students are taking other courses at the same time and have different schedules, the IPSAs are typically scheduled over a 2 to 3 day period. An alternate approach that we have tried is to let the students take the IPSA during one of the scheduled lecture periods. Although that approach works well, it requires that all students come with their own computers, which has obvious limitations.

2.4.2 Evaluating IPSA Structure

As stated previously, the objective of this endeavor was to create an assessment that probed a student's problem-solving strategies and did not simply provide the same kind of performance information that is available from tests of content knowledge. In addition, we continue to view each of the domains as independent skills, all of which are necessary for problem-solving. We hypothesized that students just beginning to practice problem-solving could be quite skillful in one domain, while not demonstrating proficiency in others. Consequently, we did not expect to find correlations between the student responses to the **Hypothesize**, **Investigate** and **Evaluate** domains, as we considered them to be independent skills. On the other hand, we found it difficult to imagine how a student could successfully **Integrate** their conclusions from an IPSA data set into their basic science understanding without first possessing an accurate comprehension of the relevant disciplinary content knowledge. This led us to predict a connection between the **Evaluate** and **Integrate** domains with each other, and with an independent measure of content knowledge. Table 2.1 presents a correlation matrix between student scores for the domains and scores from a content knowledge examination, the Comprehensive Basic Science Examination (CBSE), which was given to all of our pre-clinical medical students at this time.

Table 2.1: Correlation Coefficient Matrix Across Individual Problem-Solving Assessment Domain Scores and Content Knowledge Performance Scores

	Hypothesize	Investigate	Evaluate	Integrate	Content Knowledge (CBSE[t])
Hypothesize	1.00	0.21 ± 0.16	0.27 ± 0.07	0.24 ± 0.12	0.09 ± 0.03
Investigate		1.00	0.20 ± 0.12	0.12 ± 0.05	0.12 ± 0.18
Evaluate			1.00	**0.37* ± 0.01**	**0.53* ± 0.05**
Integrate				1.00	**0.44* ± 0.09**
Content Knowledge					1.00

N = 240 medical students; 18 IPSAs each, 3 CBSEs each, administered over 18 months.
* $p < 0.02$
[t] Comprehensive Basic Science Exam

The results demonstrate little correlation between the **Hypothesize**, **Investigate** and **Evaluate** domains. As expected, there was a modest but significant correlation between the **Evaluate** and **Integrate** domains. Student responses on both the **Evaluate** and **Integrate** domains exhibited a correlation with the results for the test of content knowledge.

Because of the unique and variant skills involved in the **Reflect** domain, and because its grading criteria were different from the other domains, student results for the **Reflect** domain were not included in this analysis.

2.4.3 Evaluation of the Rubrics - Development

As described earlier, the development of IPSA rubrics was an iterative and team-based process, which depended on the input from several disciplinary content experts. This was the most labor-intensive element of exam construction. This teamwork reinforced the cross-disciplinary nature of the IPSA scenarios, and improved the contextual relevance of the exams and helped students see the application of classroom training to their eventual careers.

We have found that the iterative process of developing rubrics tends to provide a method for identifying problems in the IPSAs. In the Biochemistry course for example, we have utilized the same 8 IPSAs for over 4 years. We evaluate the IPSAs after each iteration and make alterations based on student responses. This process has reinforced the importance of obtaining student input (through their early responses) that can improve IPSA quality and allow the same IPSA to become easier to implement after each iteration. Finally, the developmental process provides us with the confidence to provide students with timely feedback to help them modify their problem-solving strategies.

2.4.4 Evaluation of the Rubrics - Effectiveness/Validity

The standard deviation in assigned grades from three different graders on 8 IPSA scenarios given to 20 different biochemistry students during a two-semester biochemistry course varied by less than 10% with a correlation coefficient greater than 0.75. This suggests that strict adherence to the grading rubrics leads to acceptable grading consistency. Figure 2.4 depicts the IPSA rubric-based scores assigned to two representative students by these three graders, with 2.4A and 2.4B showing differing levels of grading consistency. The results are presented in the radar type format with the mean and standard deviation for the grading results indicated on the figure.

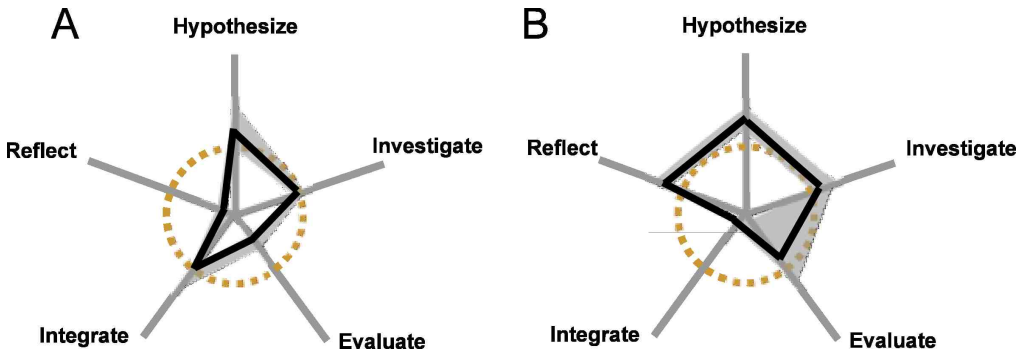


Figure 2.4: Inter-grader Reliability. Radar plots show mean scores assigned by three faculty members (black) and standard deviations (gray) for two representative students' IPSA results. The plot in **(A)** indicates standard deviations of less than 10%, while the plot in **(B)** indicates variability in grading the **Evaluate** domain.

Figure 2.4A illustrates an example of our typical grading consistency, with less than a 10% standard deviation between multiple graders. On the other hand, Figure 2.4B shows the

pattern of a student for whom the three graders disagreed on the **Evaluate** domain. In this case, the scores ranged from "acceptable" to "failure". When multiple student responses on this IPSA were evaluated using these rubrics, a similar lack of uniformity between instructor grades was persistently evident for the **Evaluate** domain. The rubrics were poorly defined in this case and the graders could not consistently apply the benchmarks. This led us to revisit our expectations, and also to use the student responses on the exam to help refine the grading rubrics.

We have identified three distinct reasons for a lack of grading consistency, and can now quickly recognize and rectify the problems. One reason, as illustrated in Fig. 2.4B, is that the rubrics are poorly defined. In such a case, the rubrics can be redefined and the question re-graded. A second reason for inconsistent grading is that the question itself is poorly worded, and is interpreted differently by students and graders. In this case, the question must be re-phrased for future use. The third source of grading inconsistency is an imprecise or ambiguous student response. In this case, the rubrics and question function acceptably for the majority of the class, but the graders have a difference in opinion on a single student's contribution because they are forced to "read between the lines" in order to assign any grade. This illustrates the real power of the iterative process for the development of grading rubrics.

An additional verification of validity of the grading rubrics was provided by the results of the comparison between the faculty graders and the graduate student grader, as illustrated in Figure 2.5. The domain scores given by the graduate student to ten student-generated performance patterns were within the experimental error set by the faculty. These data suggested to us that, once valid rubrics are established, graduate students or other instructors can assist in grading; and that it is not necessary to devote time of multiple faculty to grade student responses on the IPSAs. The authors acknowledge that the experiences and abilities of graduate students may vary considerably and that this experiment was only done once. However, coupled with our other experiences with multiple graders across various disciplines, this finding adds further evidence to the conviction that well-defined rubrics are the key to grading reliability, and that educators from different disciplines and varying levels of educational experiences can grade IPSAs accurately if sufficient time is spent developing the grading criteria.

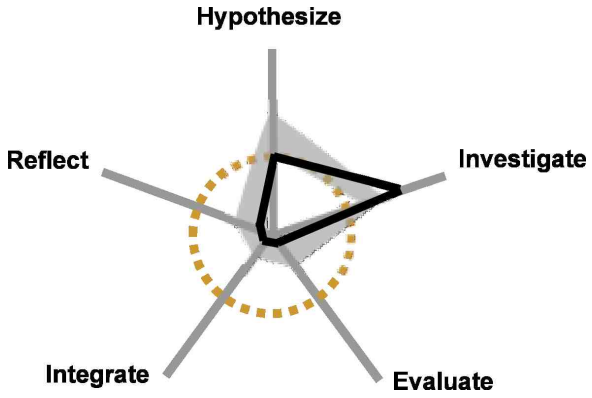


Figure 2.5: Graduate Student Grader Reliability. The radar plot of one student's IPSA grades, as assigned by a trained graduate student (black) and faculty (gray).

2.4.5 Reporting Grades

Because successful problem-solving requires mastery of all of the domains, we elected not to reduce all 5 domain scores into a single number as an indicator of performance. Instead, we reported student responses graphically as illustrated in Figures 2.4 and 2.5, which made clear student skills, or lack thereof, on individual domains. In order to provide the maximum reproducibility in pattern analysis from one IPSA to another, we standardized each domain axis independently, based on the rubrics, and defined minimal acceptable performance for each domain as "7", producing a symmetrical pattern when student performance is similar in all domains. Thus, performance patterns provide an easily understood visual tool that allows students to see their own progress relative to goals set by faculty.

2.4.6 Common Performance Patterns

We used this analysis to identify students with difficulty in problem solving and then to assist them in addressing their individual impediments. It was necessary to define the skills that an individual student possessed and those skills that the student was missing. Following this, appropriate intervention strategies were initiated. A first step in this long-term goal is the recognition of archetypal performance patterns. Four of the most common patterns that we have observed since the beginning of this endeavor are illustrated in Fig. 2.6. A full

library of archetypal performance patterns has not yet been defined, and is under investigation.

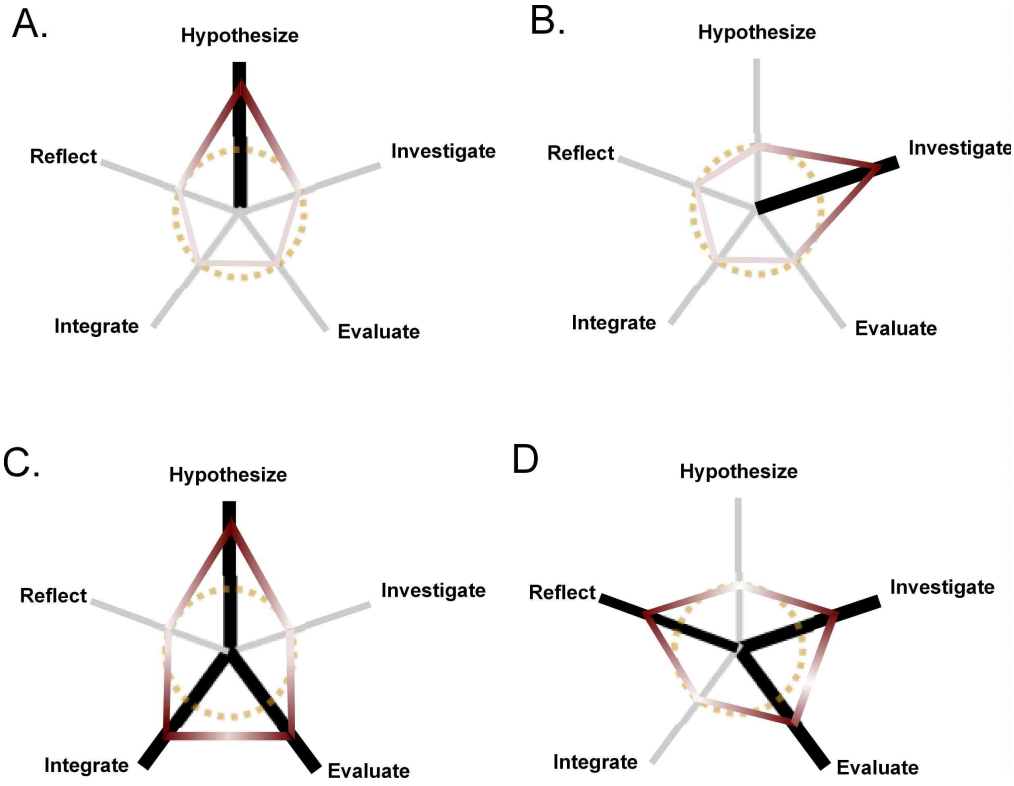


Figure 2.6: Four Common Student Performance Patterns on IPSAs. In **(A)**, students exhibit difficulty in the **Hypothesize** domain. In **(B)**, the low **Investigate** domain score indicates a challenge with contextualizing hypotheses within the scenario. In **(C)**, the low scores for both the **Evaluate** and **Integrate** domains correlate with a lack of content knowledge. In **(D)**, difficulty in the **Reflect** domain reflects poor metacognition.

Figure 2.6A depicts the most common patterns of student performance that we have seen over a 10 year period of implementing IPSAs. As shown by the low score on the **Hypothesize** axis, it is clear that one of the most difficult domains for our medical and biochemistry students to initially master is the generation of appropriate hypotheses. Fortunately, this appears to be an easily learnable skill. In faculty discussions with individual students regarding their difficulties in this area, many students reported that they had simply never been asked to do this before. Single Best Answer questions, which students have become accustomed to throughout their academic careers, present students with a concept and ask them to fill in the details. IPSAs inherently require a different approach, presenting

students with the details and asking them to develop conceptual hypotheses. Intervention strategies used to date indicate that modeling performance may provide a simple remedy to poor performance on this domain, but further research is required.

Students who exhibit the pattern illustrated in Fig. 2.6B appeared to have a difficulty putting their hypotheses into the relevant context of the scenario, as shown by the low score on the **Investigate** axis. For example, a student exhibiting this pattern will, when presented with the sudden onset of an enzyme deficiency in an adult, develop a complicated investigational strategy to probe possible genetically inherited inborn errors in metabolism, completely ignoring the fact that the patient has reached adulthood without manifesting any common symptoms of that metabolic deficiency. Like the student with difficulty defining relevant hypotheses, the intervention strategy for the problem-solving pattern illustrated in Fig. 2.6B was to increase the student's sensitivity to the environment of the problem.

Students exhibiting the pattern illustrated in Fig. 2.6C, showing low scores on the **Evaluate** and **Integrate** axes, typically earned overall grades that placed them at the bottom of the class, and have had significant difficulty in improving their performance on IPSAs. As discussed previously (Table 2.1), performance on the **Evaluate** and **Integrate** domains generally correlated with students' fundamental understanding of basic science concepts. Deficiencies in these domains may therefore reflect either a problem with a grasp of the basic sciences behind the presented problem, or an inability to mechanistically relate these basic science concepts to the context of the problem. Remediation of the academic difficulties underlying this pattern is potentially more problematic than those illustrated by Figures 2.6A and 2.6B. The authors are continuing to identify strategies to address problems in this area, but feel that it is important to first work on the content knowledge issue.

In our experience, students who exhibit the pattern shown in Figure 2.6D, with a low score on the **Reflect** axis, tend to be the most difficult to remediate as this domain is heavily dependent on metacognition. However, other work has suggested that deficiencies in this area can be remediated. (Ref. Clayton) Reflection, by definition, requires students to examine their own performance and develop appropriate strategies for improvement. In discussions with the faculty about exam performance, students who exhibit difficulty in this area claim that the exam scenarios do not really represent real life and are "unfair" or

"unrealistic". We have identified these students at all academic levels, and are continuing to explore new intervention strategies.

2.4.7 Change in Performance Patterns Over Time

When we began using the first version of these exams in the late 1990s, specific feedback on problem-solving domains was not provided to individual students; instead, training on problem-solving skills was a component of multiple course lectures. Improving our ability to recognize and more finely resolve symptomatic profiles is an ongoing investigation. We are continually refining and assessing remediation strategies to promote improved student performance, and this endeavor is currently our salient research objective. At this point, the authors believe that simply presenting students with their own performance profiles, and thus providing students with feedback on their individual strengths and weaknesses, gives them an initial and fundamental start in addressing difficulties in becoming successful at scientific problem-solving.

Figure 2.7 illustrates IPSA performance patterns for two representative students over the course of 2 semesters from the set of 20 students previously described. Neither student received specific feedback during this time. With the exception of an improvement of the **Hypothesize** domain, the student represented by Fig. 2.7A failed to achieve significant improvement in problem-solving skills. We have regularly identified students who do not improve their skills and do not seek advice. On the other hand, the student represented by Fig. 2.7B, was able, without intervention, to develop an individual strategy and to optimize an approach to problem-solving. This type of analysis provides the basis for the evaluation of future intervention strategies.

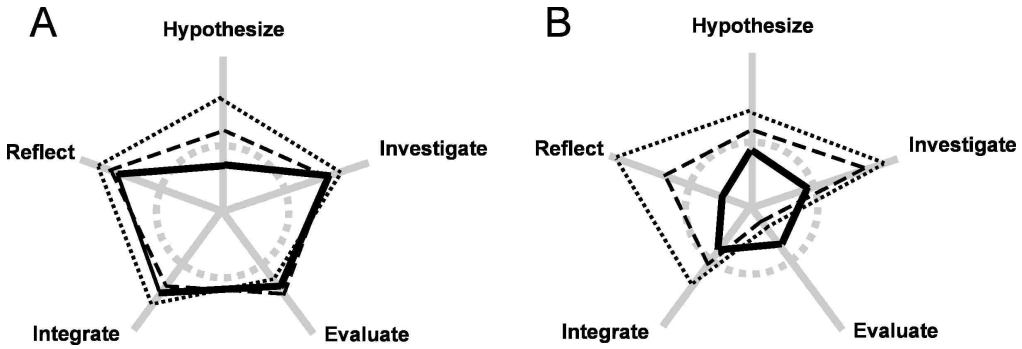


Figure 2.7: Longitudinal Performance Patterns. The change in two students' IPSA performance patterns over two semesters, at three points in time: initial (fine dashed line), midway (broad dashed line), and final (continuous line). Student **(A)** was only able to significantly improve in the **Hypothesize** domain, while student **(B)** made substantial strides and eventually exceeded expectations in all the domain scores.

2.5 Conclusion

At the University of New Mexico, our curricular approaches emphasize the integration of process and content, both at the undergraduate biochemistry level and in the School of Medicine. This paper describes a novel assessment tool, the IPSA, which provides practice to students in problem-solving, is relatively easy for faculty to administer and grade, and provides individualized assessment information to the student. The IPSAs, and the online group discussions of biomedical problems that are connected to them (Anderson *et al*., 2008; Osgood *et al*., 2008), have become integral to our efforts to "multicontextualize" biomedical education (Ibarra, 2001). These pedagogies support learners with a diversity of thinking and learning styles. They promote each learner's ability to recognize and develop their individual approach to problem-solving, in a context that honors the importance of content knowledge and its application to the career skills that will be needed by the student.

2.6 References

Allen, S. & Knight, J. (2009). A Method for Collaboratively Developing and Validating a Rubric, *Int. J. So. Teaching and Learning*. 3(2), 1-17.

American Society for Biochemistry and Molecular Biology. (2003). Recommended Curriculum for a Program in Biochemistry and Molecular Biology. *Biochemistry and Molecular Biology Education,* 31(3), 161-162. doi: 10.1002/bmb.2003 .494031030223

Anderson, W. L., Mitchell, S. M., & Osgood, M. P. (2008). Gauging the Gaps in Student Problem-Solving Skills: Assessment Of Individual And Group Use Of Problem-Solving Strategies Using On-Line Discussions. *Cell Biology Education—Life Sciences Education,* 7(2), 254-262. doi: 10.1187/cbe.07–06–0037

Ash, S. and Clayton, P. (2009). Generating, Deepening, and Documenting Learning: The Power of Critical Reflection in Applied Learning. Journal of Applied Learning in Higher Education, Vol. 1, 25-48.

Association of American Medical Colleges. (2009). *Scientific Foundations for Future Physicians.* Retrieved August 15, 2009, from AAMC Web site: http://www.aamc.org/scientificfoundations

Flavell, J. H. (1976) Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), The nature of intelligence (pp. 231-236). Hillsdale, NJ: Erlbaum

Ibarra, R. A. (2001) Beyond Affirmative Action: Reframing the Context of Higher Education, University of Wisconsin Press, Madison, WI, 43-78.

Mayer, R.E. (2002) Rote versus Meaningful Learning. *Theory & Practice* 41: 226-232.

National Research Council. (2003). *Bio2010: Transforming undergraduate education for future research biologists.* Washington, D.C.: National Academies Press.

Osgood, M. P., Mitchell, S. M., & Anderson, W. L. (October 13, 2008). Tracking student problem-solving strategies in online PBL case discussions: a method to target interventions to individuals and groups most in need of help. Commissioned Paper. National Academy of Sciences, Board on Science Education, Workshop on Linking Evidence and Promising Practices in STEM Undergraduate Education. http://www7.nationalacademies.org/bose/Osgood_Commissioned_Papers.html

Chapter 3

Scientific problem solving within an undergraduate biochemistry and molecular biology curriculum

Cheryl A. Sensibaugh, William L. Anderson[1], Marcy Osgood

Department of Biochemistry and Molecular Biology, University of New Mexico, Albuquerque, NM, USA

[1] Professor emeritus

3.1 Abstract

Discipline-based education research has produced varied perspectives on defining and assessing scientific problem solving at the undergraduate level. The goals of this study were to describe and explain longitudinal performance across two upper-level years of our biochemistry curriculum, and to initiate a nomological network of problem solving. Student performance was measured using the Individual Problem Solving Assessment (IPSA), generating scores in five domains: Hypothesize, Investigate, Evaluate, Integrate, and Reflect. Our average biochemistry student required two semesters to perform satisfactorily in three domains, two years for the Evaluate domain, and did not perform satisfactorily in the Investigate domain. Hierarchical linear regression explained performance by identifying significant contributors to performance as time, critical thinking skills (measured by the Critical thinking Assessment Test, CAT), and attitudes about learning science (measured by the Colorado Learning Attitudes about Science Survey for Biology, CLASS-Bio). All three contributors promoted problem solving, and accounted for up to eighteen percent of IPSA domain score variability. First efforts at building a nomological network revealed that aspects of problem solving converged when probing the ability to synthesize results into a conclusion, but diverged from content knowledge. Our primary pedagogical recommendations are to model successful problem solving and scaffold activities across time.

Keywords: Assessment, skills, problem solving, undergraduate, biochemistry

3.2 Introduction

As a broad learning goal for undergraduate life science majors, the ability to apply the process of science (*i.e.*, the scientific method) is an important competency for science students to develop, regardless of discipline (American Association for the Advancement of Science, 2011, p. 14). The process of science is one form of problem solving, what we term "scientific problem solving." While science is based upon problem solving, there are "disciplinary differences in what problem solving entails," (National Research Council, 2012, p. 5-15). For example, a biologist brings a different perspective, skills, and methods to

an observed problem than a chemist or physicist would bring. While the learning objectives for the overall process of scientific problem solving may be similar across disciplines, the specific criteria for meeting the objectives may be different in biology than in chemistry or physics. This work addresses problem solving as it occurs in the context of biochemistry and molecular biology (BMB).

This study is rooted in the evolving theory of constructivism, which asserts that learners construct their own learning from the building blocks of their own unique backgrounds (Bodner, 1986). Constructivism guides studies that investigate "conceptual change over time or the construction of knowledge," (Bodner and Orgill, 2007). The concept of cognitive dissonance, which refers to the conflict created when new information contradicts prior knowledge (Festinger, 1957), has also been incorporated into constructivist educational contexts. For example, a cross-disciplinary meta-analysis found that pedagogies that create cognitive dissonance between prior experience and new, contradictory information can effectively uncover alternative conceptions and thus stimulate learning in both reading and science (Guzzetti *et al.*, 1993). In order to resolve cognitive dissonance, it follows that both prior and new experiences must be reflected upon and explained. During the process of problem solving, which is foundational to the nature of science, metacognition or reflection upon one's learning is instrumental in developing a new approach to yield greater future success (Stroulia, 1994).

To round out the framework for this study, the practice of scientific teaching follows the principle that undergraduate education should be approached in the same manner as scientific research (Handelsman *et al.*, 2004; Handelsman *et al.*, 2006). Specifically, learning objectives (aims) are aligned with assessments that measure attainment of the objectives (evidence), as well as with strategically designed learning activities (methods) for achieving the objectives. Scientific teaching also advocates for applying the principle of backward design, which is to "start with the end in mind," much as a scientist does during experimental design. Educationally, this refers to defining specific learning objectives *prior* to developing assessments or activities (Wiggins and McTighe, 2005). Scientific teaching thus implements the paradigm of science in order to further our educational efforts.

Undergraduate education researchers have taken varying perspectives on defining problem solving. Stein and colleagues indirectly did so by stating that problem solving and

learning combine to form one dimension of the broader construct of critical thinking (2006). Overall, they assert that critical thinking is comprised of four dimensions, with the other three being evaluating information, creative thinking, and effective communication.

Researchers of student attitudes about learning science also employ an indirect method of defining problem solving. For instance, based on statistical factor analysis of students' self-reported attitudes about learning how to solve problems, Semsar and coworkers indirectly define problem solving as consisting of four dimensions: reasoning, synthesis and application, strategies, and effort (2011). Additionally, their analysis supports defining problem solving as one of four components to be considered with all learning attitudes about science. The remaining attitudes relate to real world connection, enjoyment, and conceptual connections.

In our previous work, we more directly defined problem solving as consisting of the scientific method with a metacognitive component, and developed specific learning objectives aligned with that definition (Anderson *et al.*, 2008). Each objective addresses one aspect, or domain, of scientific problem solving. Refined to be more explicit and updated according to the principles of scientific teaching, the objectives are as follows:

- Hypothesize Domain – Given a set of observations, students should be able to generate hypotheses about potential biochemical mechanisms underlying biological phenomena.
- Investigate Domain – Given a testable and falsifiable hypothesis regarding one distinct biochemical mechanism, students should be able to propose an experimental design to test that hypothesis.
- Evaluate Domain – Given an experimental design and data, students should be able to deduce the experimental results.
- Integrate Domain – Given an experimental result, students should be able to interpret the result within the context of the original observations, integrating pertinent evidence to form a conclusion.
- Reflect Domain – Given a conclusion, students should be able to critically evaluate their own performance.

Researchers have also approached the assessment of problem solving in a variety of ways, aligned with their definitions. The Critical thinking Assessment Test (CAT) generates

one score, which is a measure of critical thinking, on a scale of 0 – 38 points (Stein *et al.*, 2006). The test is comprised of fifteen questions in multiple choice, short answer, and essay formats. A major benefit of the CAT is that it focuses on critical thinking without embedding specific disciplinary content. Thus, the score is indicative of the ability to transfer critical thinking processes across disciplines. Further details about the CAT are available in supplementary information (Appendix C.I).

A second approach to assessing problem solving involves the affective domain of learning. Previous work in physics and chemistry demonstrated that the learning of science is influenced by students' attitudes about learning science (Adams *et al.*, 2004; Adams *et al.*, 2006; Barbera *et al.*, 2008; Perkins *et al.*, 2004). Semsar and colleagues modified surveys from those disciplines to develop the Colorado Learning Attitudes about Science Survey for Biology (CLASS-Bio; 2011). Students indicate their level of agreement on a Likert scale with 31 statements such as, "I enjoy figuring out answers to biology questions." The overall CLASS-Bio score is calculated as a percentage of statements that students answered in the same way as experts. Sub-scores can also be calculated for different categories of attitudes listed above, such as those related to learning problem solving. All CLASS-Bio statements are available by category in supplementary information (Appendix C.II).

A third approach to assessing problem solving is embodied in the Individual Problem Solving Assessment (IPSA), developed by our group (Mitchell *et al.*, 2011). The IPSA is a computer-based summative assessment that measures individual student performance in problem solving. Each IPSA follows one biochemistry problem explicitly through each of the five domains. The mechanics of an IPSA involve progressively revealing each domain to students, with each domain containing its own part of the problem (Fig. 3.1). Students enter an essay response to the prompt in each domain, and may review – but not go back and alter – completed domains at any time.

An IPSA opens with a scenario describing observations about a biochemical problem (Fig. 3.1A). Only the **Hypothesize** domain is accessible to students at this point. After providing minimal information to supplement the observations, the IPSA prompts students to generate multiple hypotheses that explain the observed phenomenon. Once students enter their hypotheses, the **Investigate** domain becomes accessible, while subsequent domains remain inaccessible to students (Fig. 3.1B). Here, students are prompted to design an

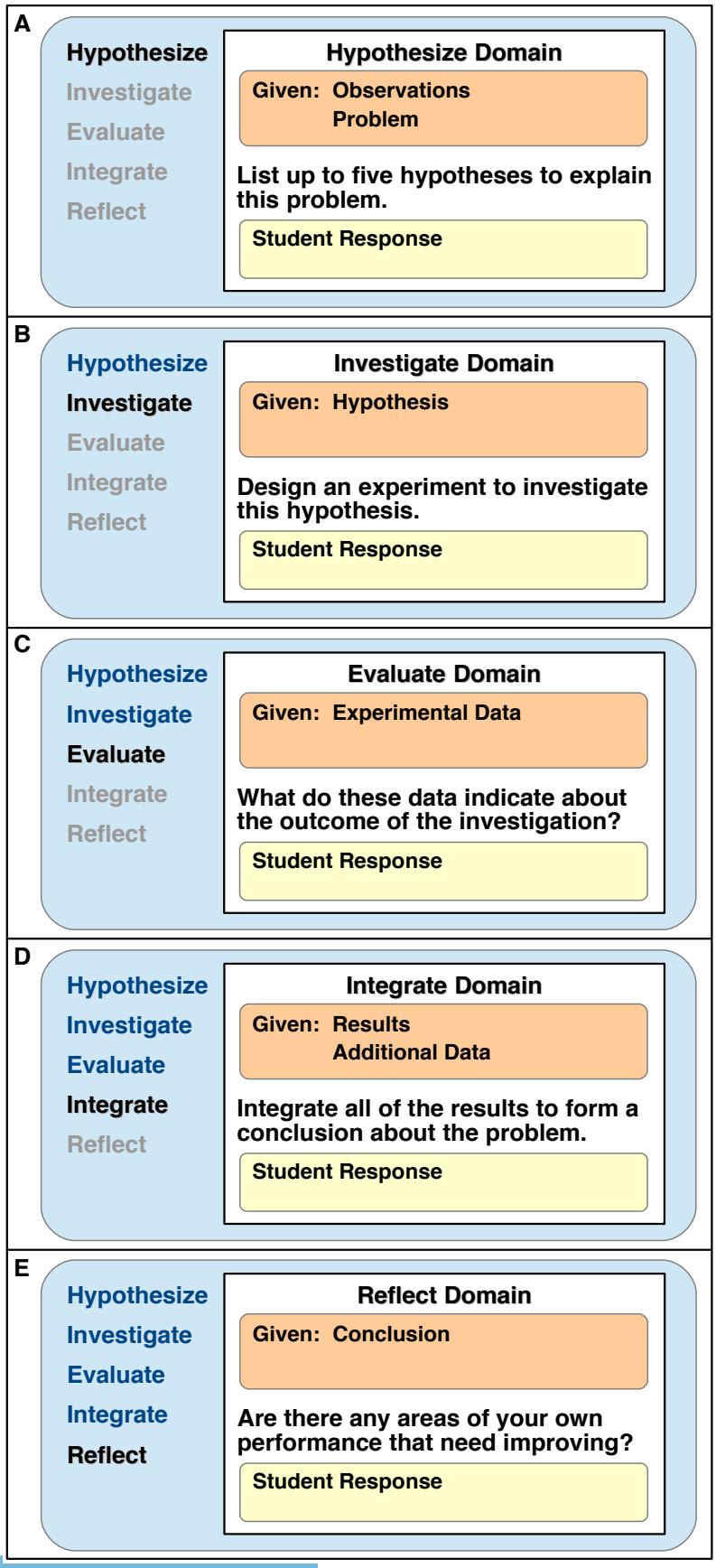


Figure 3.1: IPSA Mechanics. The progressive-reveal nature of an IPSA is captured in simplified versions of screen shots from each domain during computer administration. **(A)** Hypothesize, **(B)** Investigate, **(C)** Evaluate, **(D)** Integrate, and **(E)** Reflect. Black domain text on the left indicates the currently active domain, while gray text indicates inaccessible domains. Students may review the content and responses from previously completed domains (blue text), but cannot edit responses.

experiment that will test a single given hypothesis, which is written into the body of the assessment by the instructor. In the third section of an IPSA, the **Evaluate** domain, experimental results are provided in the form of figures, graphs, or tables, and students are prompted to evaluate the results (Fig. 3.1C). Then in the **Integrate** domain, an interpretation of the previous results are given, more data are provided, and students are prompted to integrate all available IPSA information, plus their content knowledge from their coursework, into the original context of the problem to come to a conclusion concerning the biochemical problem (Fig. 3.1D). Finally, when the **Reflect** domain is reached, the correct answer to the IPSA biochemical problem is provided, and students are asked to reflect on their responses (Fig. 3.1E).

Students typically complete an IPSA within 45-75 minutes. Rubrics for instructors to grade the student responses contain specific criteria for scoring each domain on a scale of 0 – 10, with a score of seven points defined as satisfactory performance (Appendix C.III). In this way, domain scores are generated for each student, on each IPSA.

To provide students with formative opportunities to practice solving problems, we also used online cases (OLCs) as activities that require students to go through the same set of problem solving steps as in the IPSAs, but OLCs were implemented in a group setting, via a web-based asynchronous discussion forum (Anderson *et al.*, 2008). The case discussion boards are available for one to two weeks, with group facilitators guiding students through the scientific ways of thinking about problem solving. OLC rubrics yield one overall case score for all members of a group, rather than domain scores for each student.

This study seeks to address some of the recommendations in the National Research Council's report on discipline-based education research (2012). The report states that the time is upon us to investigate more nuanced aspects of teaching and learning than the benefits of broadly-defined "active learning" over passive lecturing. Indeed, overwhelming evidence has established the benefits of active learning (Freeman *et al.*, 2014; Wieman, 2014). Specific areas of interest to the discipline-based education research community include generating evidence about learning that concerns: (1) upper-level science courses, rather than focusing primarily on introductory courses, (2) entire science curricula, beyond single courses, and (3) student adeptness not only with factual knowledge, but also with applying it to the processes of science.

Accordingly, the research goals of this study are two-fold: first, to quantitatively describe and explain student performance in scientific problem solving across a two-year biochemistry curriculum; and second, to initiate an understanding of the nomological network of various problem solving skills and biochemistry content knowledge. We also discuss recommendations for pedagogical practice, to maintain student-centered learning as a crucial underpinning for the research and to inform scholarly educators.

The first research question regarding performance is, what is the longitudinal performance pattern of an average student, and when do most students begin maintaining satisfactory performance, in each domain of problem solving? Longitudinal IPSA domain score means describe average performance over two years, while consistent satisfactory performance for most students (more than half) is better described by domain achievement rates across time. The achievement rate is the proportion of students who perform satisfactorily.

Previous work (Mitchell *et al.*, 2011) revealed two important findings about problem solving performance in our curriculum. First, students exhibited difficulties in every domain of problem solving. Second, some students made only limited improvements during their junior year (when they take multiple IPSAs), while others were able to reach satisfactory performance during that year without additional formal instructional strategies. A preliminary analysis of eleven biochemistry majors during the second semester of their junior year started to probe likely means and achievement rates, at the beginning and end of that semester (Fig. C.IV.1). Means were satisfactory by the end of the semester in the Integrate domain, and at both time points in the Reflect domain. Achievement rates were most similar to those for content exams in the Integrate and Reflect domains, but were markedly lower in other domains. Taken together, these findings inform our hypothesis that the average student struggles in all domains, and then improves to satisfactory performance over time, the duration of which varies by domain. Furthermore, we expect that consistent satisfactory performance by most students would begin in the second semester, but not in all domains.

The second research question is: Which contributors – among time, academic background, and demographic background – most consistently explain the observed problem solving performance, and how much do the contributors impact performance? A major concern in studies of complex human subjects in open systems is that of examining the

impact of one variable at a time while controlling other variables. Quantitatively, hierarchical linear regression provides a means to do so, by yielding regression equations that explain the variability in scores from student to student, along with determining how much of the score variability is explained by the equations. For a more complete and practical explanation of the applications of regression analyses in science education research, see Theobald and Freeman (2014).

Our hypothesis is that the greatest contributor to IPSA domain performance is time, followed by academic backgrounds of students, then by demographic backgrounds. We defined academic background as consisting of content knowledge (as measured by biochemistry course grades and content exam scores), critical thinking ability (as measured by the CAT), learning attitudes about science (as measured by the CLASS-Bio), and disciplinary major. Demographic backgrounds take into consideration age, gender, race, and ethnicity. In the absence of previous data on the interplay of these contributors with problem solving performance in the context of biochemistry, we can only estimate that the regression equations generated from the identified contributors will explain less than half of the variability in scores across students.

Our second main goal, initiating an understanding of the nomological network of scientific problem solving, has not previously been attempted using the assessments described here. Similar in appearance to a concept map, a nomological network represents relationships between constructs (the characteristics intended to be measured) and what can actually be measured (Cronbach and Meehl, 1955). In this study, the constructs are problem solving, critical thinking, learning attitudes about science, and content knowledge. Measurements are in terms of scores on the IPSA, CAT, CLASS-Bio, and content exams, respectively. The network shows connections between related elements, with no links between unrelated elements, thus depicting areas of convergence and divergence in the network.

The third research question we address in this study is therefore: How closely related are the observable scores of problem solving, critical thinking, and learning attitudes about science, to each other as well as to scores of content knowledge? Scores from assessments that measure similar constructs would be expected to correlate at least moderately. Based on a review of test content, we hypothesize that IPSA Hypothesize, Investigate, Evaluate, and

Integrate domain scores will correlate at least moderately with CAT scores, while IPSA Reflect domain scores will correlate at least moderately with CLASS-Bio scores. Based on our previous work (Mitchell *et al.*, 2011), we also expect that the IPSA Evaluate and Integrate scores will correlate with content exam scores.

3.3 Methods

Data collection – The study was conducted at the University of New Mexico (UNM), pursuant to research protocol 12-634, approved by the Human Research Review Committee at the UNM Health Sciences Center. As shown in the study design (Fig. 3.2), three cohorts of students were included in the study. Each cohort entered the biochemistry curriculum in sequential academic years (*i.e.*, Cohort A in the first year, B in the second, and C in the third). Longitudinal data were collected across two years for Cohorts A and B (Fig. 3.2A), and across one year for Cohort C (Fig. 3.2B). All students were pooled to maximize statistical power (Fig. 3.2C).

Students completed two core biochemistry courses: one on structure and function (BIOC I), and the other on metabolism (BIOC II), taken during the junior year (Fig. 3.2A-C, timeline). Within each course, four content exams were administered (Fig. 3.2A-C, squares). Content exams primarily measured lower-order cognitive skills; *i.e.*, remembering and understanding, rather than higher-order cognitive skills, such as applying information. Four IPSAs (Fig. 3.2A-C, pentagons) and four OLCs were also administered in each course, with the combined points from these problem solving assessments and activities comprising no more than ten percent of overall course grades. At least ninety percent of course grades were determined by content exams, short quizzes, and content-oriented activities.

Students in Cohort C also took the CAT and CLASS-Bio at the beginning and end of their junior year (Fig. 3.2B-C, triangles and circles, respectively). The senior year included biochemistry elective courses, which did not incorporate OLCs and IPSAs. Then at program exit, Cohorts A and B completed the American Chemical Society's 2003 Biochemistry Exam© (a nationally standardized content exam) along with an exit IPSA. Assessment scores from the following time points were collected for Cohorts A and B: entry into BIOC I, after one semester, after one year, and after two years (Fig. 3.2A). For Cohort C, test-retest

reliability of the CAT limited data collection to the time points at entry and after one year (Fig. 3.2B).

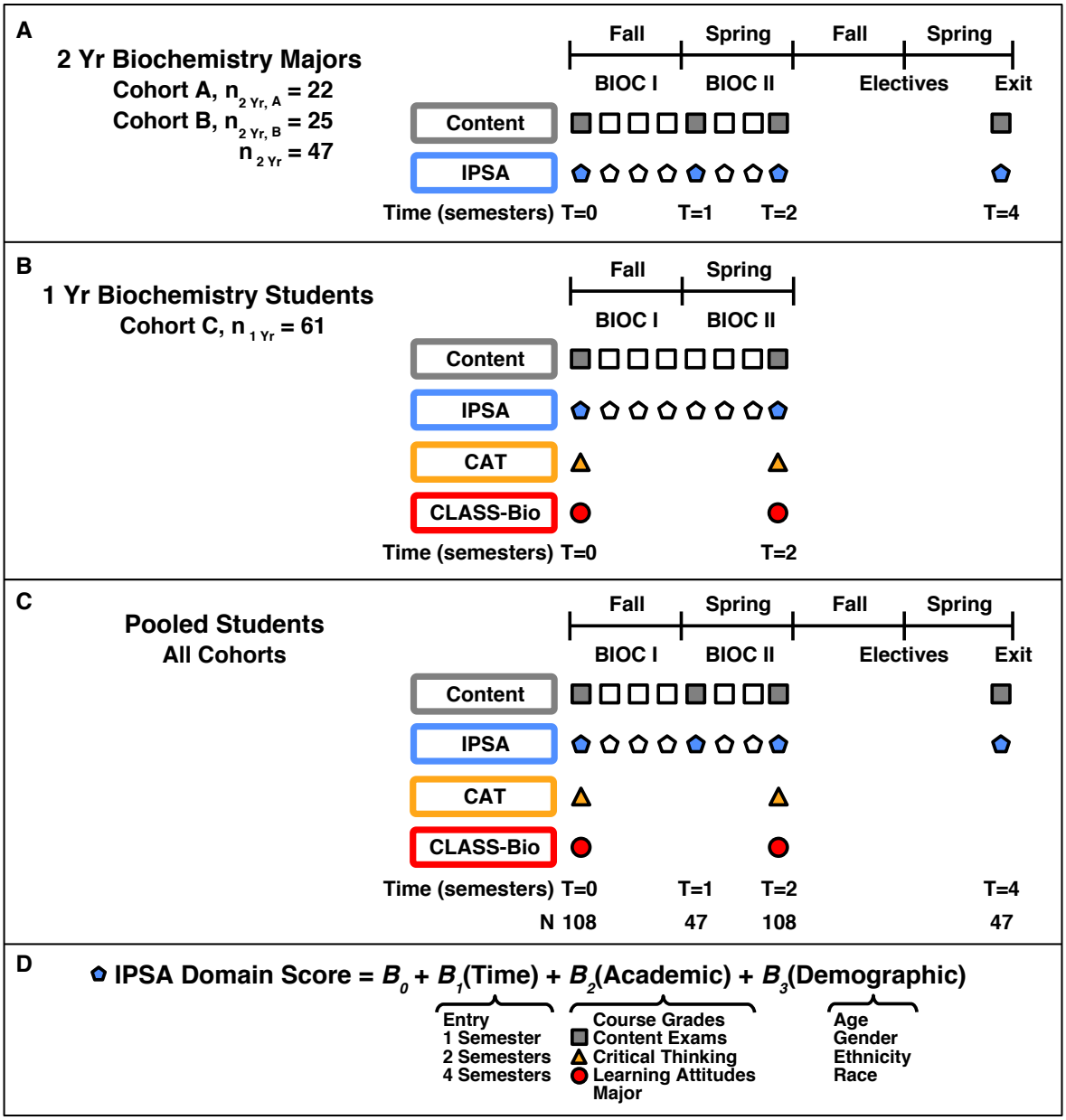


Figure 3.2: Study Design. Student cohorts and assessments are depicted for one- and two-year academic timelines. **(A)** Two cohorts of biochemistry majors took nine content exams (squares) and nine IPSAs (pentagons) during the two-year program. Scores were analyzed at four time points (filled polygons). **(B)** A third cohort of students completed the one-year sequence of biochemistry courses, as well as the CAT (triangles) and CLASS-Bio (circles). Due to limited test-retest reliability of the CAT, analyses were performed only at two time points (filled polygons). **(C)** All three cohorts were pooled for analyses, yielding the sample sizes shown for each time point. **(D)** The putative regression model being tested includes three main components of student experience that may impact IPSA domain scores.

Student backgrounds – Both academic (Fig. C.V.1) and demographic (Fig. C.V.2) aspects of student backgrounds were considered when evaluating IPSA performance. Approximately 86 percent of students in the study were biochemistry majors (Fig. C.V.1A). The mean biochemistry course grades were between 86 – 88 percent (Fig. V.1B), while mean scores on biochemistry content exams ranged more widely across time, between 57 – 82 percent (Fig. C.V.1C). Regarding demographics, most students were traditionally aged Caucasian males. However, 17 percent of the sample was comprised of returning students (Fig. C.V.2A), and 42 percent of all students were female (Fig. C.V.2B). The Hispanic or Latino/a population was represented by 31 percent of students, seven percent were Asian, two percent were African American, and two percent were American Indian (Fig. C.V.2C).

Statistical analyses – SPSS software (IBM Corp.) was used for all analyses, and the confidence level was set to 0.05 for tests of significance.

Instrument validity – One source of validity evidence is test content, which includes the scoring rubrics (American Educational Research Association *et al.*, 2014). Assessment scores are measurements, and the scales are an indicator of instrument sensitivity. The IPSA scoring rubrics detailed ten criteria for each domain response, with one point available per criterion, thus maximizing the sensitivity of the IPSA. The CAT is scaled similarly, with one point per criterion. The broad range of 31 items on the CLASS-Bio contributes to its sensitivity in detecting students' attitudes. Other validity measures for the CAT and CLASS-Bio were not determined in our study.

Longitudinal IPSA task variability – The entry IPSA was concerned with a protein purification protocol, unlike subsequent IPSAs that required experimental design in the Investigate domain. Consequently, the scoring criteria were fundamentally different for the Investigate domain of the entry IPSA, and the scales were not comparable across IPSAs. Therefore, Investigate scores at program entry were excluded from analysis.

Inter-rater reliability – In the absence of resources for administering and expecting students to complete multiple isomorphic IPSAs, inter-rater reliability was the primary measure considered, as estimated by the intraclass correlation coefficient (ICC). Pursuant to statistical power guidelines offered by Walter and colleagues (1998), two raters scored IPSA responses of 39 students at program entry, for calculating ICCs across IPSA domain scores. Inter-rater reliability was estimated with author MO as the true grader and author CS as the

additional grader. ICCs guided further scoring methods, to determine whether CS applied the domain scoring rubrics as consistently as MO, and would therefore reliably score all the IPSA responses for this study. To meet this expectation, a minimum ICC of 0.70 was established. The reasoning behind this minimum stemmed from two sources. Our previous work determined that scores assigned by three faculty raters (which included authors MO and WA) were strongly related, as evidenced by Pearson correlation coefficients greater than 0.75 with less than ten percent standard deviations (Mitchell *et al.*, 2011). The same work also showed that a graduate student rater (CS) scored within tolerance of the faculty raters. A second source of guidance was an interpretation of the range of ICC values (Cicchetti, 1994; Hallgren, 2012). An ICC of 0.7 is at the high end of the range for good agreement, while excellent agreement is reached at an ICC value of 0.75. Confidence intervals of 95% for the ICCs were also calculated.

The ICC was greater than 0.7 for all IPSA domains at entry (Fig. C.III.1). The 95% confidence interval of the ICC contained 0.8 for all domains. With excellent agreement between scores assigned by CS and MO, CS completed the scoring of all remaining IPSA responses, and scores assigned by CS were used in analyses.

Inter-rater reliabilities of the CAT and CLASS-Bio were not measured for this study. The CAT responses were scored by developers of that instrument; *i.e.*, true raters. For scoring of the CLASS-Bio, Likert scale responses are transformed to dichotomous scores to report whether or not students agreed with experts on each item.

Score means – Means were calculated with 95% confidence intervals for course grades, content exam scores, IPSA domain scores, CAT scores, and CLASS-Bio scores. As previously described, satisfactory performance in an IPSA domain was defined by a score of at least seven points (Mitchell *et al.*, 2011). To interpret mean IPSA domain scores, one-sample *t*-tests determined whether the scores were either at or below satisfactory levels at each time point. The null statistical hypothesis was that domain scores were at least seven points; the alternative was that means were lower than seven points. The assumption of normality was tested by visual inspection of distribution histograms.

Achievement rates – Achievement rates were calculated for IPSA domain scores, with 95% confidence intervals, to represent the percentage of students with satisfactory scores at each time point. To interpret achievement rates, binomial tests determined whether most

students performed satisfactorily. The null statistical hypothesis was that at least fifty percent of students earned satisfactory scores; the alternative was that achievement rates were lower than fifty percent. Normality was not assumed nor tested, in the context of this non-parametric test.

Regression models – To explain performance in problem solving, linear regression analyses generated equations that quantitatively model the contribution of various factors to variability in IPSA domain scores. A sequential hierarchical order of entry grouped the variables into three main qualities that may directly explain variability in IPSA scores: time, academic background, and demographic background (Fig. 3.2D). All three elements of the theoretical framework of this study contain elements of time, thus its contribution is accounted for first. Constructivism implies an element of time during the gathering of previous knowledge and experience. Resolving cognitive dissonance implies the need for time, especially with complex tasks such as problem solving. Scientific teaching includes time by emphasizing active practice of learning objectives and formative feedback. Student backgrounds, while also key, were intuitively secondary to time. In the context of scientific problem solving, we reasoned that academics (one aspect of student background) were more likely to explain performance than demographics.

Variables included within academic background were major, overall performance as measured by biochemistry course grades, content knowledge as measured by content exam scores, critical thinking as measured by CAT scores, and learning attitudes as measured by CLASS-Bio scores. Furthermore, the variable of research experience – as determined by enrollment in two semesters of honors research courses – was originally included as an academic factor for Cohorts A and B, yet it was not applicable for Cohort C as those students had not had a chance to complete the courses. Therefore, when the cohorts were pooled, research experience was no longer included in the regression analyses.

For demographic background, the variables were age group, gender, race, and ethnicity. Age groups were defined as either traditional (less than 26 years on September 1 of junior year) or returning (26 years or older), in an attempt to quantify any effect of life experience or maturity. For each domain model, R^2, adjusted R^2, and F values were reported, along with estimated regression coefficients (β), standard errors, and 95% confidence intervals of the coefficients.

Score correlations – To determine the degree of convergence and divergence among scores of problem solving (IPSA), critical thinking (CAT), attitudes about learning science (CLASS-Bio) and content exams, Pearson's correlation coefficients (*r*) were calculated and tested against the null hypothesis that the values equaled zero. For interpreting the size of *r* within the context of discipline-based education research, values of at least 0.1 indicate a weak association, 0.3 is moderate, 0.5 is strong, and 0.7 is very strong (Maher *et al.*, 2013).

3.4 Results and Discussion

3.4.1 Problem solving performance fluctuates across time and domains.

Average problem solving performance – Mean IPSA domain scores were highly variable overall, both increasing and decreasing through time (Fig. 3.3). To interpret IPSA scores, satisfactory performance in any domain is defined by a score of at least seven out of ten points (Mitchell *et al.*, 2011). In the Hypothesize domain, performance reached a satisfactory level for the average student only after one semester, then dropped. In the Investigate domain, average scores were well below satisfactory performance regardless of time. The Evaluate domain averages increased over time, reaching satisfactory levels only after two years in the program. Integrate domain scores exhibited no such trend, however, with scores declining over the first semester, but rebounding well into the satisfactory range

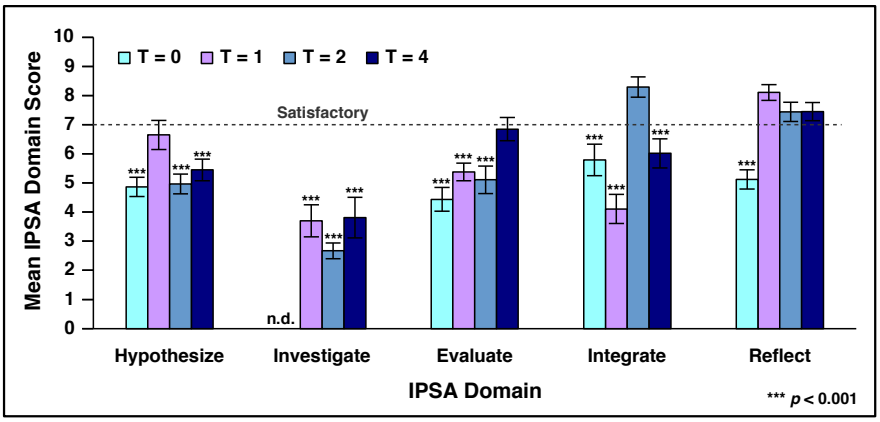


Figure 3.3: Longitudinal IPSA Domain Scores. Mean scores for each IPSA domain are shown across the two-year program. Error bars indicate 95% confidence intervals of the means. To target means that were below satisfactory performance (dotted line), one-sample t-tests determined which means were significantly lower than seven points. n.d., no data.

after two semesters, then declining again by program exit. Finally, mean scores in the Reflect domain reached satisfactory levels within one semester and were maintained. In summary, the average student in our biochemistry curriculum requires two semesters to achieve satisfactory performance in three domains of problem solving (Hypothesize, Integrate, and Reflect), two years for the Evaluate domain, but does not achieve satisfactory performance in the Investigate domain.

Problem solving achievement rates – Achievement rates for all domains were calculated to indicate the proportions of students who performed satisfactorily (Fig. 3.4). The trends mirror those seen in mean scores, with at least half the students in this study reaching satisfactory performance at least once in each domain, except the Investigate domain. However, for most time points in most domains, achievement rates were less than thirty percent. Most students maintained satisfactory performance after the first semester only in the Reflect domain, and after two semesters in the Integrate domain.

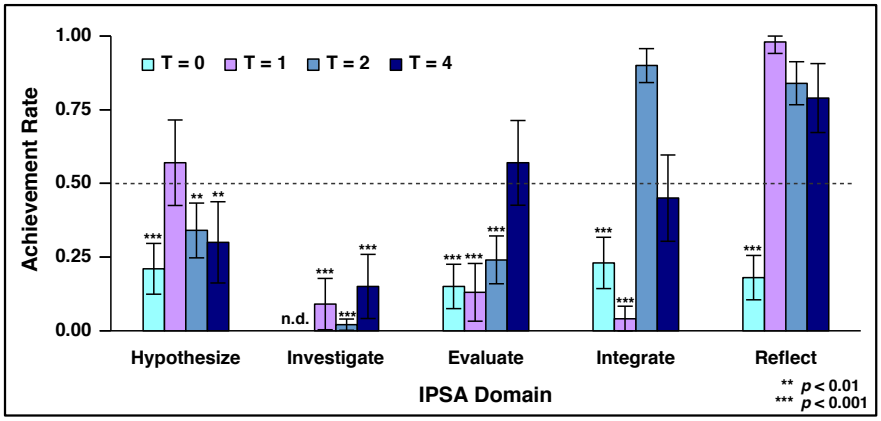


Figure 3.4: Longitudinal IPSA Domain Achievement Rates. The proportions of students who achieved a satisfactory score in each IPSA domain are shown across time. Error bars indicate 95% confidence intervals of the rates. To target rates that were below those for content exams (dotted line; Fig. IV.1B), binomial tests determined which rates were significantly lower than fifty percent. n.d., no data.

While IPSAs are designed to assess higher order cognitive skills, it is important to note that the cognitive level of an assessment (*e.g.*, as classified by Bloom's Taxonomy) is a different educational consideration from the difficulty of an assessment (Lemons and Lemons, 2013). Classical test theory (CTT) is the analytical framework used in this study,

and defines item difficulty as the proportion of students who respond with the correct answer. Given the structure of the IPSAs, item difficulty can be estimated as the proportion of students who respond satisfactorily to a domain prompt. Thus, the achievement rates reported here also serve as indices of the difficulty of each domain. From this standpoint, the fluctuations demonstrate irregular item difficulty from one IPSA to another, thus implying a need for standardization of the tasks to be performed in each domain.

While IPSAs are designed to assess higher order cognitive skills, it is important to note that the cognitive level of an assessment (*i.e.*, as classified by Bloom's Taxonomy) is a different educational consideration from the difficulty of an assessment (Lemons and Lemons, 2013). Classical test theory (CTT) is the analytical framework used in this study, and defines item difficulty as the proportion of students who respond with the correct answer. Given the structure of the IPSAs, item difficulty can be estimated as the proportion of students who respond satisfactorily to a domain prompt. Thus, the achievement rates reported here also serve as indices of the difficulty of each domain. From this standpoint, the fluctuations demonstrate irregular item difficulty from one IPSA to another, thus implying a need for standardization of the tasks to be performed in each domain.

3.4.2 Time, critical thinking, and learning attitudes promote performance in problem solving.

CAT Measurements of Critical Thinking – CAT scores were slightly above national averages at program entry and after two semesters (Fig. C.V.1D). Entering the junior year, students scored 18.7 points on average. At the end of that academic year, the mean increased to 20.9 points.

CLASS-Bio Measurements of Learning Attitudes – Biochemistry students scored similarly to other upper-level students on the CLASS-Bio, who were at the University of Colorado (CU) in the Departments of Integrative Physiology and Molecular, Cellular, and Developmental Biology (Fig. C.V.1E). The design of this study only allowed comparison with CU students at program entry, since the study of Semsar and colleagues (2011) was designed with an end point of one semester rather than two. The overall and subscale scores showed increasing trends across the junior year of our curriculum.

Table 3.1. Regression values for IPSA domain score equations

Domain	Model R^2	Adjusted Model R^2	Model F [a]	Coefficient	Estimate [a]	SE	95% Confidence Interval	
							Lower	Upper
Hypothesize	0.08	0.06	4.49 *	Intercept	2.6 **	0.92	0.8	4.4
				Time	0	0	0	0
				Academics				
				CAT	0.07 *	0.03	0.00	0.13
				CLASS-Bio	0.02	0.01	0.00	0.04
				Demographics	0	0	0	0
Investigate	0.01	- 0.01	0.45	Intercept	3.0 ***	0.39	2.2	3.8
				Time	0	0	0	0
				Academics	0	0	0	0
				Demographics	0	0	0	0
Evaluate	0.15	0.14	9.39 ***	Intercept	2.4 *	0.94	0.5	4.3
				Time	0.5 **	0.13	0.2	0.7
				Academics				
				CLASS-Bio	0.03 *	0.01	0.00	0.05
				Demographics	0	0	0	0
Integrate	0.19	0.17	12.01 ***	Intercept	0.5	1.27	- 2.0	3.0
				Time	0	0	0	0
				Academics				
				CAT	0.16 **	0.05	0.07	0.25
				CLASS-Bio	0.04 *	0.02	0.01	0.07
				Demographics	0	0	0	0
Reflect	0.20	0.18	12.68 ***	Intercept	4.6 ***	0.81	3.0	6.2
				Time	0.5 ***	0.11	0.3	0.7
				Academics				
				CLASS-Bio	0.02	0.01	0.00	0.04
				Demographics	0	0	0	0

[a] Significance levels for F test and two-sided t test: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Quantitative Models of Problem Solving Performance – Regression equations to explain performance in problem solving ranged widely in their fit with the data. Regression analyses yielded statistically significant models for all domains except Investigate. According to adjusted model R^2 values, the proportion of variability in scores explained by the regression models ranged from six to eighteen percent (Table 3.1). Time played a role in increasing Evaluate and Reflect domain scores. Academic factors were significant in explaining all domain scores except Investigate. Notably, demographic backgrounds did not

influence any IPSA domain score after accounting for time and academic factors. These results are summarized into a visual representation of scientific problem solving performance (Fig. 3.5A).

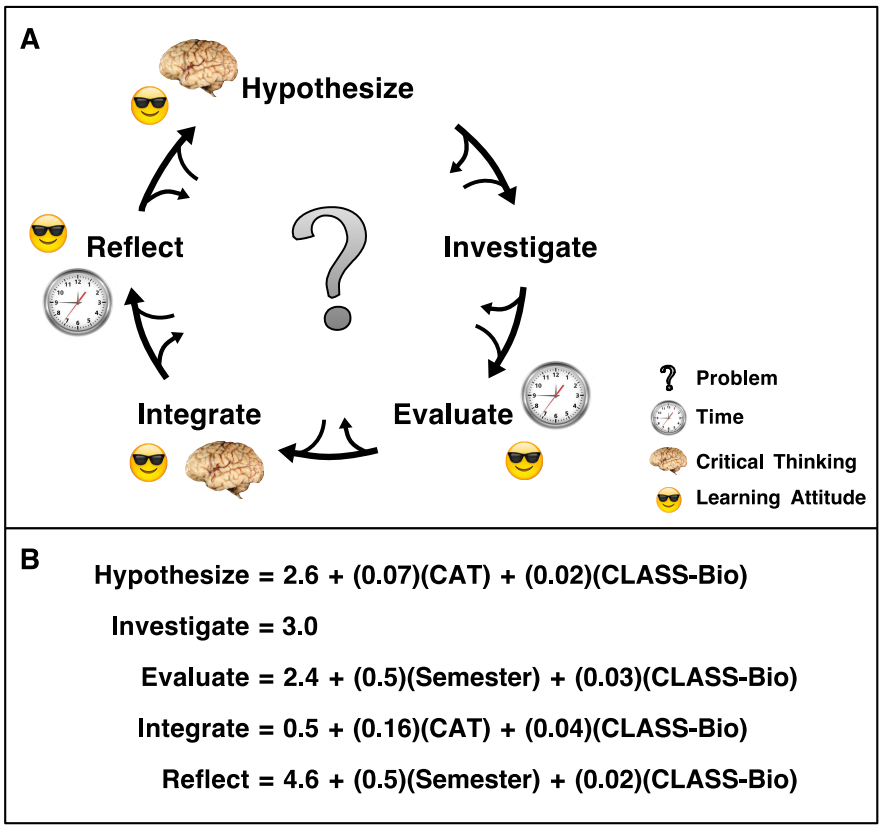


Figure 3.5: Influences Upon Scientific Problem Solving Performance. **(A)** The visual model summarizes the current quantitative findings across two curricular years. **(B)** Regression models are summarized for each IPSA domain score, when the model was statistically significant. Otherwise, only the intercept is included.

Regression equations provided quantitative models for performance (Fig. 3.5B). The model for the Hypothesize domain explained six percent of the variability in scores (F = 4.49, p = 0.013). When all factors were held constant, a Hypothesize domain score of 2.6 was predicted (p = 0.005). Yet critical thinking abilities and learning attitudes both impacted this domain score. For every point earned on the CAT (38 maximum), the Hypothesize score increased by 0.07 (p = 0.045). For example, a CAT score of 20 would raise a Hypothesize score by 1.4 points. Similarly, for each point on the CLASS-Bio (100 maximum), the Hypothesize score increased by 0.02 (p = 0.094); *e.g.*, a CLASS-Bio score of 60 would raise a Hypothesize score by 1.2 points. The model predicts a total domain score in this example

as 2.6 + 1.4 + 1.2 = 5.2 points (10 maximum).

For Investigate, none of the variables examined in this study were able to explain variability in these domain scores ($F = 0.45$, $p = 0.505$). Only the estimated y-intercept is statistically different from zero. The model predicts Investigate domain scores of 3.0 ($p < 0.001$), without contributions from time, academics, or demographics.

In Evaluate, a stronger model emerged, explaining fourteen percent of the score variability ($F = 9.39$, $p < 0.001$). When all else was equal, the model predicted Evaluate scores to be 2.4 ($p = 0.012$). For each semester increase in time when other factors were controlled, a half point gain was anticipated ($p = 0.001$). For each point on the CLASS-Bio, the Evaluate score was estimated to increase by 0.03 ($p = 0.031$); *e.g.*, a CLASS-Bio score of 60 would raise an Evaluate score by 1.8 points.

In Integrate, the model explained seventeen percent of the variability among scores ($F = 12.01$, $p < 0.001$). The model predicted domain scores to be nearly zero when all else was equal ($B = 0.5$, $p = 0.700$). Yet critical thinking made a large contribution to Integrate domain scores. Controlling for other factors, Integrate scores were expected to increase by 0.16 for every additional CAT point ($p = 0.001$). Thus, a CAT score of 20 would add 3.2 points to an IPSA Integrate domain score. The largest contribution to a domain by learning attitudes was also demonstrated in Integrate ($B = 0.04$, $p = 0.010$). Accordingly, a CLASS-Bio score of 60 was expected to raise an Integrate score by 2.4 points.

Finally, the Reflect domain score was influenced by both time and learning attitudes. The model explained eighteen percent of score variability in this domain ($F = 12.68$, $p < 0.001$). Controlling for all examined variables, the Reflect score was predicted to be 4.6 ($p < 0.001$). The time coefficient matched that in Evaluate, at a half point increase per semester ($p < 0.001$), when all else was equal. The model also predicted that for every CLASS-Bio point, the Reflect score would increase by 0.02 points ($p = 0.115$); *e.g.*, a CLASS-Bio score of 60 would raise the Reflect score by 1.2 points.

To summarize our quantitative explanation of problem solving as measured by the IPSA, time, critical thinking, and learning attitudes all promoted performance, yet in different ways across domains. Critical thinking ability (as measured by the CAT) impacted IPSA performance in only two domains, but with relatively large contributions. The affective domain of learning, as measured by the CLASS-Bio, played a pervasive role throughout

problem solving, via attitudes about learning science.

3.4.3 The nomological network of problem solving converges when forming conclusions.

To determine the extent of convergence and divergence among IPSA, CAT, CLASS-Bio, and content exam scores, score correlations are reported in Fig 3.6. Only some of the hypothesized correlations were demonstrated. First, the IPSA Evaluate score correlated moderately with a CLASS-Bio sub-score (Problem Solving, PS: Synthesis & Application). Likewise, IPSA Integrate scores correlated moderately with both the CAT and CLASS-Bio scores, as well as with the PS: Synthesis & Application sub-score of the CLASS-Bio. Additionally, CAT scores correlated moderately with one of the CLASS-Bio sub-scores, again that of PS: Synthesis & Application.

Our hypothesis that correlations would be present between the IPSA Evaluate and Integrate domains, and content exam scores, was not supported. No relationship was demonstrated for the Evaluate domain, and the Integrate domain only correlated weakly, which can be explained by the fact that biochemistry content is incorporated into the IPSAs. These results indicate that the IPSAs indeed assessed something different than the content measures assessed, which is consistent with our results during development (Mitchell *et al.*, 2011).

	IPSA Hypothesize	IPSA Investigate	IPSA Evaluate	IPSA Integrate	IPSA Reflect	Content Exam	CAT	CLASS-Bio Overall
Content Exam	0	0	0	0.18 **	- 0.16 **	1	0	0.23 *
CAT	0.23 *	0	0.23 *	*0.36 ******	0	0	1	0.23 *
CLASS-Bio Overall	0.21 *	0	0.25 *	*0.31 **	0.20 *	0.23 *	0.23 *	1
Real World Connection	0	0	0	0	0	0	0	**0.80 *** **
Personal Enjoyment	0	0	0	0.22 *	0	0	0	**0.69 *** **
Conceptual Connections	0	0	0	0.22 *	0.21 *	0.24 *	0	**0.81 *** **
PS: Reasoning	0	0	0	0	0	0	0	**0.63 *** **
PS: Synthesis & Application	0.21 *	0	*0.31 **	*0.36 ******	0	0.23 *	*0.30 **	**0.77 *** **
PS: Strategies	0	0	0	0.22 *	0	0	0	**0.56 *** **
PS: Effort	0	0	0	0	0	0	0	**0.76 *** **

* $p < 0.05$
** $p < 0.01$
*** $p < 0.001$

Figure 3.6: Score Correlations. Pearson's correlation coefficients are shown for correlations that were statistically different from zero. Italics indicate moderate correlations ($r > 0.3$) and bold indicates strong correlations ($r > 0.5$).

Taken together, our results indicate moderate relationships among problem solving (as measured by the IPSA), critical thinking (as measured by the CAT), and learning attitudes (as measured by the CLASS-Bio) when students are asked to integrate results to form a conclusion, thus demonstrating convergence across those skills. Yet the otherwise weak convergence in other domains of problem solving reveals that the three constructs, and their corresponding assessments, are distinct from one another, and merit use as defined.

Correlations with zero or negative values identified areas of divergence in the nomological network of problem solving. As predicted, CAT scores did not correlate with content scores, while CLASS-Bio scores were only weakly related to content scores, which may be due to the CLASS-Bio having been designed specifically for use in the discipline of biology. No correlations existed when relating content exam scores to the IPSA Hypothesize, Investigate, and Evaluate domain scores. Notably, the Reflect domain scores exhibited a weak, negative correlation with content exam scores. These findings indicate that content knowledge (as measured by content exams) diverges from process-oriented skills, when those skills involve the scientific method and metacognition (as measured by the IPSA), and critical thinking (as measured by the CAT).

The lack of correlations with the Investigate domain was unexpected. The CAT items minimally probed aspects of experimental design, yet enough to hypothesize that a correlation would exist, perhaps within the inherent skill of Evaluating Information. Yet unlike the IPSA, the CAT did not explicitly prompt students to design an experiment. A different approach would be to compare Investigate domain scores with those generated by the Experimental Design Ability Test (Sirum and Humburg, 2011), which is more closely aligned based on a review of test content. The Investigate domain was expected to correlate with CLASS-Bio scores based partly on our experience with the IPSA: Anecdotally, students expressed great, ongoing concern about their lack of laboratory experience, and what methods they should describe in an IPSA, and the level of detail to include in their descriptions. Given these apprehensions, attitude was expected to relate in some way to the Investigate domain, yet our evidence does not support such a conclusion.

The emergent nomological network represents the correlations visually (Fig. 3.7). Convergence is readily seen in the circular pattern of heavy arrows that connect all three process-oriented assessments. Divergence of those assessments from tests of content

knowledge is apparent in the sole weakly negative correlation, as well as in the reduced number of connections. These findings provide additional instrument validity evidence based on the relationships among scores, and inform how we, as discipline-based education researchers, can define and assess problem solving.

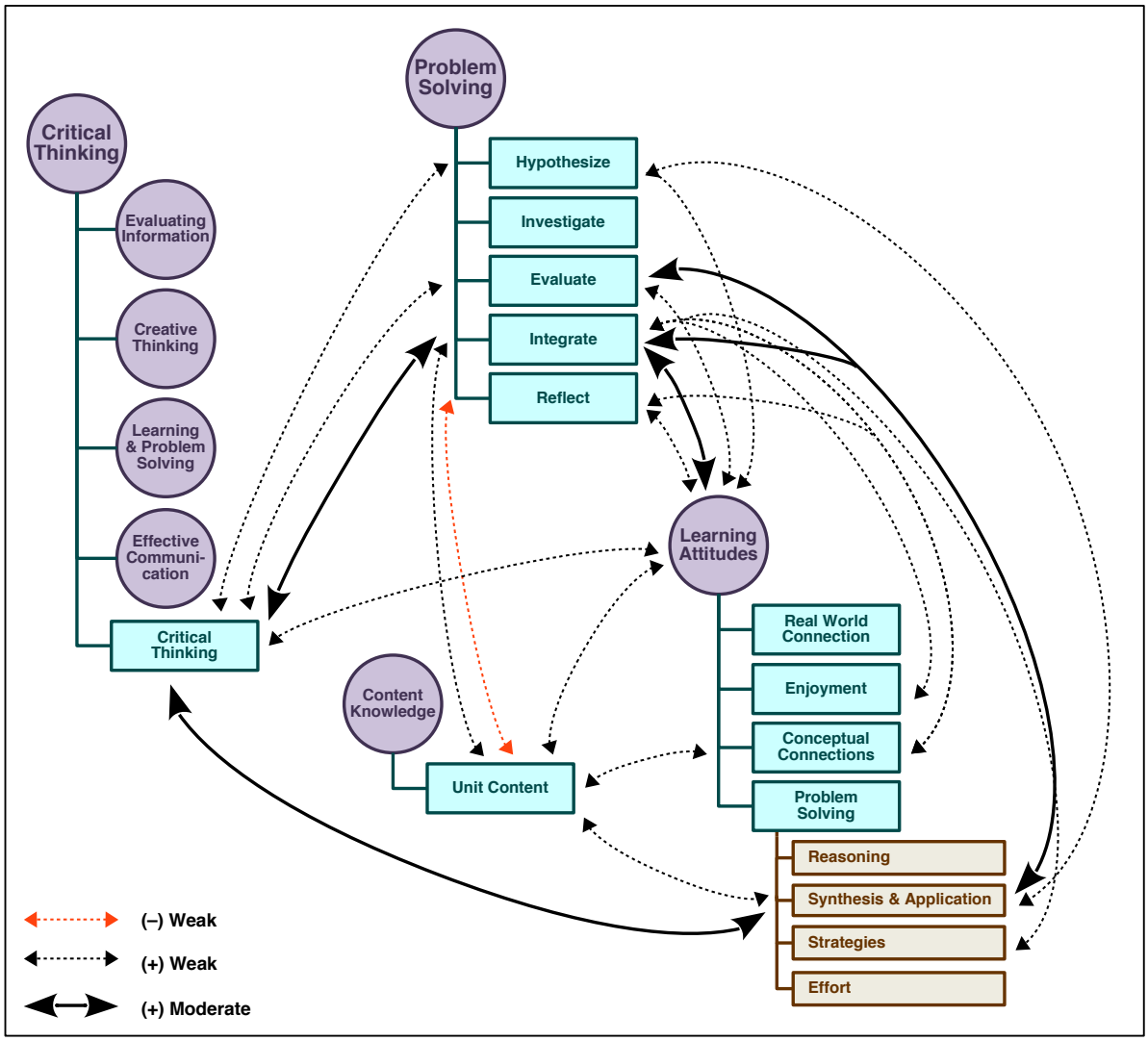


Figure 3.7: Nomological Network of Problem Solving. Relationships among problem solving, critical thinking, learning attitudes, and content knowledge (circles) are summarized based upon correlations among observable scores (boxes). All moderate correlations were positive (thick arrows). Most weak correlations were positive (black dotted arrows), with one weakly negative correlation between IPSA Reflect scores and content exam scores (orange dotted arrow).

3.5 Current Limitations and Future Research

Expanding IPSA implementation beyond its current environment promises multiple benefits. Research questions regarding transferability across disciplines, course levels, and institutions could be addressed. Analytically, item response theory (IRT) would be possible as a guiding framework only with larger samples, since it requires a minimum sample size of 200 to 500 students. Whereas CTT is primarily concerned with an overall assessment, the benefit of IRT is that it analyzes individual items within an assessment. Under IRT, the mathematical function that emerges from the analysis contains parameters for both item difficulty and item discrimination.

Item difficulty is an important measure, and would be informative during standardization efforts. Due to the faculty expectation of students accumulating scientific knowledge throughout a curriculum, it is likely that IPSA item difficulty increased over time. For example, the IPSA at program exit incorporates much more biochemistry than the early IPSAs. Analytically, item difficulty could be used either to normalize IPSA scores, or as an additional factor that influences performance.

3.6 Pedagogical Implications

A goal in using the IPSAs in this study was to describe and explain longitudinal student performance in problem solving. Here, we discuss potential applications of our findings, for any scholarly educator interested in promoting improvement in student problem solving performance. A key finding – demonstrated in three different ways – is that most students struggle with most domains of scientific problem solving across the entire upper-level curriculum, rather than improving to satisfactory levels over time. We suggest several explanations for this phenomenon.

A primary concern is that problem solving activities and assessments were not heavily weighted in course grades. The impetus for this was to provide structure for practice and feedback in problem solving, while reducing the stakes to allow for mistakes to be made. Furthermore, the IPSA and content exam at program exit were required to be completed prior to graduation, but the scores were not incorporated into any course grade. The importance of

this first concern is that sufficient incentives to motivate students' best efforts (*i.e.*, course points) were largely missing during this study.

Another explanation for these performance outcomes would be that our curriculum and course designs do not provide enough scaffolded practice attempts for full development of well-rounded expertise in all domains of problem solving. These results may represent a lag phase that we as educators can work to shorten by incorporating more opportunities for deliberate practice of problem solving within our curricula. Again within the framework of scientific teaching, it is essential to design learning activities that incorporate mechanisms for providing feedback to students. That feedback should include information about where and how to improve in order to meet the targeted learning objective. Furthermore, we understand that problem solving is a complex process; therefore, students will need additional opportunities for deliberate practice after receiving feedback, to determine whether they meet the objective yet in a low stakes setting.

This study provides a snapshot of performance over a two-year BMB curriculum, allowing scholarly educators to develop learning activities that target different domains at different times, to promote longitudinal improvements. We recommend an approach that first models, then scaffolds, exemplary problem solving for students. This aligns with the need to target the factor of time overall, and turn it to the advantage of students. Modeling the problem solving process would provide a full example for students to study and dissect. Scaffolding the process would incorporate more intermediate steps at the beginning of the curriculum, which would be removed over time as students become more practiced. Taken together, modeling and scaffolding would also be expected to reduce the item difficulty (as measured by achievement rates, discussed above), since those formative activities should better prepare students for successful completion of a summative IPSA.

Opportunities for students to make common mistakes are also recommended during the early phases of scaffolding. Building in such opportunities to fail would be expected to reduce the lag time mentioned above (Cannon and Edmondson, 2005; Coelho and McClure, 2005; Stroulia, 1994). Furthermore, this type of structure has the potential to inform our understanding of student misconceptions about solving scientific problems.

Our exclusion of data for the Investigate domain at program entry was a final testament to the emerging need to standardize IPSA prompts and rubrics, while allowing for

different biochemical content in each IPSA scenario. Accordingly, we have started creating a library of standardized IPSAs. We envision that regardless of content, a domain prompt will be the same across IPSAs, and that a domain rubric for scoring responses will be the same for all IPSAs.

To pedagogically address the influence of critical thinking skills upon problem solving performance, we recommend incorporating these aspects into formative learning activities across a BMB curriculum. One way to achieve this would be to develop CAT analogs (Stein and Haynes, 2011). For the development of expert-like learning attitudes, short activities would not be expected to have an appreciable impact (Hansen and Birol, 2014). The work of Hansen and Birol interestingly culminated in a recommendation to include "activities that emphasize the nature of scientific knowledge, the scientific method, and metacognition," in order to promote attitudes that are similar to those held by experts. This is consistent with the convergence demonstrated in our nomological network of scientific problem solving, and shows the utility of the network in pedagogical efforts.

In summary, this study provides some evidence necessary to promote improvement. At the course level, learning activities must go beyond practice to guided practice, with modeling and scaffolding of the problem solving process. Scaffolding should also include opportunities for students to fail safely, and to learn from common mistakes. Standardized IPSAs with rubrics based on problem solving criteria rather than on the content within each IPSA will provide more structure for students and make the requirements of each domain more transparent. At the curricular level as well as the course level, problem solving activities and assessments should be incentivized similar to those for content knowledge, and offered with increased frequency.

3.7 Conclusion

Our description of average student performance in problem solving, as measured by the IPSA, indicates that longitudinal trends in scores are inconsistent from domain to domain. Therefore, efforts to promote performance in this critical skill will need to be tailored to both domain and stage in the curriculum. At exit from the curriculum, without deliberate practice incorporated into the senior year, the average student continued to employ strategies that

yielded less than satisfactory performance in three domains. Based upon longitudinal achievement rates, we conclude that for most students, their own strategies were sufficient for achieving most objectives at least once during the two-year curriculum, but success was not widely maintained across domains.

While our regression equations modeling problem solving performance begin to explain the roles of time, academics, and demographics, much variability in domain scores remains to be explained. Additional longitudinal variables are necessary to explain performance in problem solving, as well as to explain the impact of time. Novel considerations are also necessary to identify variables that impact experimental design ability, as measured in the Investigate domain.

Finally, the nomological network of problem solving suggests that a definition of the construct extends beyond the scientific method and metacognition to include critical thinking and holding expert-like attitudes. Since the main area of convergence among the IPSA, CAT, and CLASS-Bio was in the IPSA Integrate domain, where students form conclusions, this suggests that much more is going on with students' reasoning at this stage of problem solving than is currently identified by the criteria in the IPSA scoring rubrics. Our hope is that this network will continue to be refined in future collaborations, to inform our understanding of defining and assessing scientific problem solving.

3.8 Additional materials

Supplementary information is available with this article as a separate file, which includes details about the CAT (Appendix C.I) and CLASS-Bio (Appendix C.II). The topics, prompts, and scoring rubrics for the IPSAs used in this study are part of Appendix C.III, along with results on IPSA inter-rater reliability. Additional sections are comprised of results of the preliminary study on IPSA performance (Appendix C.IV) and student backgrounds (Appendix C.V). Methods and results on score distributions (Appendix C.VI) and cohort differences (Appendix C.VII) are also part of the materials.

3.9 Acknowledgments

We thank CLRC for the use of previously developed IPSAs and administration software. Barry Stein, Ada Haynes, and Kevin Harris at Tennessee Technological University provided much support for use of the CATs. Our gratitude for manuscript review also goes to Steven Mitchell, Erika Offerdahl, and Jay Parkes.

3.10 References

Adams WK, Perkins KK, Dubson M, Finkelstein ND, Wieman C. (2004). *The design and validation of the Colorado Learning Attitudes about Science Survey.* Paper presented at the Physics Education Research Conference.

Adams WK, Perkins KK, Podolefsky NS, Dubson M, Finkelstein ND, Wieman CE (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Phys Rev ST Phys Educ Res 2*, 010101.

American Association for the Advancement of Science. (2011). Vision and Change in Undergraduate Biology Education: A Call to Action.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, D.C.

Anderson WL, Mitchell SM, Osgood MP (2008). Gauging the gaps in student problem-solving skills: assessment of individual and group use of problem-solving strategies using online discussions. *CBE Life Sci Educ 7*(2), 254-262.

Barbera J, Adams WK, Wieman CE, Perkins KK (2008). Modifying and validating the Colorado Learning Attitudes about Science Survey for use in chemistry. *J Chem Educ 85*(10), 1435-1439.

Bodner GM (1986). Constructivism: A Theory of Knowledge. *J Chem Educ 63*(10), 873-878.

Bodner GM, Orgill M (2007). *Theoretical frameworks for research in chemistry/science education*. Upper Saddle River, NJ: Pearson Education, Inc.

Cannon MD, Edmondson AC (2005). Failing to Learn and Learning to Fail (Intelligently). *Long Range Planning 38*(3), 299-319.

Cicchetti DV (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment 6*(4), 284-290.

Coelho PRP, McClure JE (2005). Learning from failure. *Mid-American Journal of Business 20*(1), 13-20.

Cronbach LJ, Meehl P (1955). Construct validity in psychological tests. *Psychological Bulletin 52*(4), 281-302.

Festinger L (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP (2014). Active learning increases student performance in science, engineering, and mathematics. *Proc Natl Acad Sci USA 111*(23), 8410-8415.

Guzzetti BJ, Snyder TE, Glass GV, Gamas WS (1993). Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. *Reading Research Quarterly 28*(2), 116-159.

Hallgren KA (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol 8*(1), 23-34.

Handelsman J, Ebert-May D, Beichner R, Bruns P, Chang A, DeHaan R, Gentile J, Lauffer S, Stewart J, Tilghman SM, Wood WB (2004). Scientific Teaching. *Science 304*, 521-522.

Handelsman J, Miller S, Pfund C (2006). *Scientific Teaching*. New York, NY: W. H. Freeman and Company.

Hansen MJ, Birol G (2014). Longitudinal Study of Student Attitudes in a Biology Program. *CBE Life Sci Educ 13*(2), 331-337.

Lemons PP, Lemons JD (2013). Questions for assessing higher-order cognitive skills: it's not just Bloom's. *CBE Life Sci Educ 12*(1), 47-58.

Maher JM, Markey JC, Ebert-May D (2013). The other half of the story: effect size analysis in quantitative research. *CBE Life Sci Educ 12*(3), 345-351.

Mitchell SM, Anderson WL, Sensibaugh CA, Osgood MP (2011). What really matters: Assessing individual problem-solving performance in the context of biological sciences. *Int J So Teaching and Learning 5*(1).

National Research Council (2012). *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering* (S. R. Singer, N. R. Nielsen & H. A. Schweingruber Eds.). Washington, DC: The National Academies Press.

Perkins KK, Adams WK, Pollock SJ, Finkelstein ND, Wieman CE. (2004). *Correlating student beliefs with student learning using the Colorado Learning Attitudes about Science Survey.* Paper presented at the Physics Education Research Conference.

Semsar K, Knight JK, Birol G, Smith MK (2011). The Colorado Learning Attitudes about Science Survey (CLASS) for use in Biology. *CBE Life Sci Educ 10*(3), 268-278.

Sirum K, Humburg J (2011). The Experimental Design Ability Test (EDAT). *Bioscene 37*(1), 8-16.

Stein B, Haynes A (2011). Engaging Faculty in the Assessment and Improvement of Students' Critical Thinking Using the Critical Thinking Assessment Test. *Change: The Magazine of Higher Learning 43*(2), 44-49.

Stein B, Haynes A, Redding M. (2006). *Project CAT: Assessing Critical Thinking Skills.* Paper presented at the National STEM Assessment Conference.

Stroulia E. (1994). *Failure-driven learning as model-based self-redesign.* (Doctor of Philosophy), Georgia Institute of Technology.

Theobald R, Freeman S (2014). Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research. *CBE Life Sci Educ 13*(1), 41-48.

Walter SD, Eliasziw M, Donner A (1998). Sample Size and Optimal Designs for Reliability Studies. *Stat Med 17*, 101-110.

Wieman CE (2014). Large-scale comparison of science teaching methods sends clear message. *Proc Natl Acad Sci USA 111*(23), 8319-8320.

Wiggins G, McTighe J (2005). *Understanding by Design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Chapter 4
Conclusion

This research produced a novel assessment tool, the IPSA, validated to measure student attainment of learning objectives that relate to understanding the process of scientific problem solving (Chapter 2). Through a major effort spanning ten years, the instrument went through many iterative cycles of developing problems, gathering expert reviews of test content, collecting student responses, clarifying scoring rubrics, and monitoring inter-rater reliability. My roles in this endeavor included all of the above during the last two years of IPSA development and validation.

In my leadership role during the next phase of the study (Chapter 3), longitudinal performance in problem solving within an upper-level biochemistry curriculum was described in detail, with snapshots of average student performance and achievement rates across two years. These descriptions were necessary in order to (a) explain the observed performance in terms of elements that influence variability in student domain scores, and (b) to provide evidence that informs decisions faced by scholarly educators regarding the learning and teaching of scientific problem solving.

Finally, at a conceptual level, a nomological network of problem solving was assembled to understand relationships among the content and processes involved in learning biochemistry. Armed with empirical evidence about the interplay of various components of scientific knowledge, educators can apply that information to their facilitation of student-centered learning in biochemistry.

In the broadest context of DBER, this work combines findings about problem solving in biochemistry, critical thinking across disciplines, and learning attitudes about biology. The outcome is a view of defining and assessing scientific problem solving that can be adapted, transferred, and tested in additional disciplinary educational contexts.

Appendices

Appendix A

Statistical Procedures

These statistical procedures provide additional details, beyond those summarized in the methods of Chapter 3. Specific steps for performing the analyses within SPSS software are included, along with relevant SPSS output.

A.1 Inter-rater reliability

The population parameter of inter-rater reliability is estimated by various sample statistics, depending on the type of data analyzed (Table A.1). The Intraclass Correlation Coefficient (ICC) is calculated for continuous variables when the reliability of one typical rater is in question. The ICC compares the variability of different ratings of the same subject to the total variation across all ratings and all subjects.

$$ICC = \frac{MS_{Students} - MS_{Raters \times Students}}{MS_{Students} + \left(df_{Raters}\right)\left(MS_{Raters \times Students}\right) + \frac{n_{Raters}\left(MS_{Raters} - MS_{Raters \times Students}\right)}{n_{Students}}} = \frac{s^2_{between}}{s^2_{between} + s^2_{within}}$$

Table A.1. Statistics to Estimate Inter-Rater Reliability

Measure	Qualitative Data		Quantitative Data
	Categorical	Ordinal	Continuous
2 Raters	Cohen's kappa	Cohen's weighted kappa	
3+ Raters	Fleiss's kappa Conger's exact kappa		
Correlation; not precise agreement			Pearson correlation Kendall's tau
1 Rater			Intraclass Correlation Coefficient (ICC) aka in SPSS: Single measure intraclass correlation $ICC = \frac{s^2_{Between}}{s^2_{Between} + s^2_{Within}}$
All Raters			Inter-rater Reliability Coefficient aka Spearman-Brown Correction of the ICC aka in SPSS: Average measure intraclass correlation $\frac{\left(n_{raters}\right)\left(ICC\right)}{1 + \left(n_{raters} - 1\right)\left(ICC\right)}$

SPSS Procedure for Inter-Rater Reliability:

Use IPSA Scoring Comparison data file
Analyze → Scale → Reliability Analysis
    Move all raters into the Items box.
Statistics: descriptives, inter-item statistics, and summaries
ANOVA Table by F test
Intraclass correlation coefficient
Model = Two-Way Random Effects ANOVA
    One source of variability is due to differences in students
    Second source of variability is due to differences in raters
        Raters are considered a random sample from the population of raters
Type = Absolute Agreement
    (not just consistency)
    Systematic differences in ratings ARE relevant
Confidence Interval = 95%
Test value = 0
Continue, OK.

SPSS Output for Hypothesize:

ANOVA

		Sum of Squares	df	Mean Square	F	Sig
Between People		254.718	38	6.703		
Within People	Between Items	.628	1	.628	.773	.385
	Residual	30.872	38	.812		
	Total	31.500	39	.808		
Total		286.218	77	3.717		

Grand Mean = 4.37

Intraclass Correlation Coefficient - Hypothesize

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.785[a]	.627	.881	8.251	38	38	.000
Average Measures	.879	.771	.937	8.251	38	38	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

SPSS Output for Investigate:

ANOVA

		Sum of Squares	df	Mean Square	F	Sig
Between People		125.949	38	3.314		
Within People	Between Items	.051	1	.051	.122	.729
	Residual	15.949	38	.420		
	Total	16.000	39	.410		
Total		141.949	77	1.843		

Grand Mean = 6.64

Intraclass Correlation Coefficient - Investigate

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.779[a]	.617	.878	7.897	38	38	.000
Average Measures	.876	.763	.935	7.897	38	38	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

SPSS Output for Evaluate:

ANOVA

		Sum of Squares	df	Mean Square	F	Sig
Between People		432.718	38	11.387		
Within People	Between Items	13.962	1	13.962	9.909	.003
	Residual	53.538	38	1.409		
	Total	67.500	39	1.731		
Total		500.218	77	6.496		

Grand Mean = 4.29

Intraclass Correlation Coefficient - Evaluate

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.742[a]	.505	.866	8.082	38	38	.000
Average Measures	.852	.671	.928	8.082	38	38	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

SPSS Output for Integrate:

ANOVA

		Sum of Squares	df	Mean Square	F	Sig
Between People		612.949	38	16.130		
Within People	Between Items	.115	1	.115	.066	.799
	Residual	66.385	38	1.747		
	Total	66.500	39	1.705		
Total		679.449	77	8.824		

Grand Mean = 6.47

Intraclass Correlation Coefficient - Integrate

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.808[a]	.663	.895	9.233	38	38	.000
Average Measures	.894	.798	.944	9.233	38	38	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

SPSS Output for Reflect:

ANOVA

		Sum of Squares	df	Mean Square	F	Sig
Between People		285.385	38	7.510		
Within People	Between Items	12.321	1	12.321	18.594	.000
	Residual	25.179	38	.663		
	Total	37.500	39	.962		
Total		322.885	77	4.193		

Grand Mean = 5.04

Intraclass Correlation Coefficient - Reflect

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.781[a]	.454	.901	11.334	38	38	.000
Average Measures	.877	.624	.948	11.334	38	38	.000

Two-way random effects model where both people effects and measures effects are random.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

A.2 Hierarchical linear regression

SPSS Procedure for Regressions:

Use long form master data file
Analyze → Regression → Linear
DV: IPSA Domain Score
IV Hierarchical Block 1 (Time):
Time
IV Hierarchical Block 2 (Academics):
BIOC I Grade, BIOC II Grade, Content Exam Score, CAT Score, CLASS-Bio Score
IV Hierarchical Block 3 (Academics, disciplinary major variables):
Non-major, Graduate
IV Hierarchical Block 4 (Demographics):
Gender, Age Group
IV Hierarchical Block 5 (Demographics, race/ethnicity variables):
Hispanic, Asian, American Indian, African American
Method is "Enter" for all blocks
Statistics: Estimates, 95% CI, Model fit, R squared change, Descriptives
Plots: Y=ZRESID, X=ZPRED, Histogram of standardized residuals, Normal probability plot
Save: Predicted Unstandardized & Standardized, Residuals Unstandardized & Standardized
Options: Default entry .05 & removal .10, Include constant, Missing values exclude pairwise
Case Labels: Study ID

Two assumptions were not tested for any IPSA domain:

- The correct IVs have been specified in the model, by evaluating scatter plots of:
  - Standardized residuals as a function of a putative predictor
  - Standardized residuals as a function of standardized residuals from a putative model

No other data were available to consider, due to IRB restrictions.

- The IV scores are reliable, by evaluating:
  - Cronbach's *a*, as a measure of the internal consistency of the scales
  - Cohen's k, as a measure of the inter-rater agreement of observations
  - Pearson's *r*, as a measure of correlation between test-retest scores

The reliability of IV scores was not determined (Section 3.3.6).

SPSS Output for Hypothesize:

Model Summary[f]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.045[a]	.002	-.007	1.760	.002	.212	1	106	.646
2	.313[b]	.098	.044	1.714	.096	2.149	5	101	.066
3	.314[c]	.099	.026	1.730	.001	.045	2	99	.956
4	.315[d]	.100	.007	1.747	.001	.039	2	97	.961
5	.365[e]	.133	.003	1.751	.034	.911	4	93	.461

a. Predictors: (Constant), Time (semesters)

b. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

c. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

d. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

e. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender, Asian, American Indian, African American, Hispanic

f. Dependent Variable: IPSA Hypothesize Domain

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.657	1	.657	.212	.646[b]
	Residual	328.263	106	3.097		
	Total	328.920	107			
2	Regression	32.228	6	5.371	1.828	.101[c]
	Residual	296.692	101	2.938		
	Total	328.920	107			
3	Regression	32.496	8	4.062	1.357	.225[d]
	Residual	296.424	99	2.994		
	Total	328.920	107			
4	Regression	32.736	10	3.274	1.072	.391[e]
	Residual	296.184	97	3.053		
	Total	328.920	107			
5	Regression	43.902	14	3.136	1.023	.438[f]
	Residual	285.018	93	3.065		
	Total	328.920	107			

a. Dependent Variable: IPSA Hypothesize Domain

b. Predictors: (Constant), Time (semesters)

c. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

d. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

e. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

f. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender, Asian, American Indian, African American, Hispanic

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	5.179	.263		19.698	.000	4.658	5.700
	Time (semesters)	.053	.115	.045	.461	.646	-.175	.281
2	(Constant)	.836	2.171		.385	.701	-3.471	5.142
	Time (semesters)	-.117	.134	-.099	-.872	.385	-.383	.149
	BIOC I Grade	.021	.028	.087	.727	.469	-.035	.077
	BIOC II Grade	.014	.021	.088	.667	.506	-.027	.054
	Content Exam Score	-.018	.014	-.190	-1.339	.184	-.045	.009
	CAT Score	.072	.034	.214	2.127	.036	.005	.140
	CLASS-Bio Overall	.021	.011	.193	1.876	.064	-.001	.043
3	(Constant)	.846	2.231		.379	.705	-3.581	5.272
	Time (semesters)	-.114	.136	-.096	-.835	.405	-.384	.156
	BIOC I Grade	.020	.029	.086	.711	.479	-.036	.077
	BIOC II Grade	.014	.021	.092	.683	.496	-.027	.056
	Content Exam Score	-.018	.014	-.185	-1.285	.202	-.045	.010
	CAT Score	.072	.034	.214	2.106	.038	.004	.141
	CLASS-Bio Overall	.020	.012	.186	1.689	.094	-.004	.044
	Major - Non-Biochemistry	-.225	.770	-.029	-.292	.771	-1.753	1.303
	Major - Graduate	-.050	.630	-.008	-.079	.938	-1.300	1.201
4	(Constant)	.942	2.282		.413	.681	-3.586	5.471
	Time (semesters)	-.110	.138	-.093	-.797	.427	-.384	.164
	BIOC I Grade	.020	.029	.086	.698	.487	-.037	.078
	BIOC II Grade	.013	.021	.086	.626	.533	-.029	.056
	Content Exam Score	-.017	.014	-.178	-1.207	.230	-.045	.011
	CAT Score	.071	.035	.211	2.049	.043	.002	.141
	CLASS-Bio Overall	.020	.012	.186	1.659	.100	-.004	.045
	Major - Non-Biochemistry	-.200	.783	-.026	-.256	.799	-1.754	1.354
	Major - Graduate	-.069	.647	-.011	-.106	.915	-1.353	1.215
	Gender	-.096	.356	-.027	-.271	.787	-.803	.610
	Age Group	-.029	.469	-.006	-.062	.951	-.959	.902
5	(Constant)	1.061	2.361		.449	.654	-3.627	5.749
	Time (semesters)	-.112	.139	-.094	-.803	.424	-.387	.164
	BIOC I Grade	.016	.030	.068	.542	.589	-.043	.076
	BIOC II Grade	.015	.022	.099	.711	.479	-.028	.059
	Content Exam Score	-.017	.014	-.179	-1.194	.235	-.046	.011
	CAT Score	.073	.036	.217	2.037	.044	.002	.145
	CLASS-Bio Overall	.020	.012	.185	1.649	.102	-.004	.045
	Major - Non-Biochemistry	-.186	.790	-.024	-.235	.815	-1.755	1.384
	Major - Graduate	.006	.675	.001	.010	.992	-1.335	1.348
	Gender	-.109	.357	-.031	-.306	.760	-.819	.600
	Age Group	.099	.480	.021	.206	.837	-.854	1.052
	Hispanic	.047	.395	.012	.118	.906	-.738	.831
	Asian	.749	.680	.112	1.102	.273	-.601	2.099
	American Indian	-1.080	1.299	-.083	-.832	.408	-3.661	1.500
	African American	-1.556	1.329	-.120	-1.170	.245	-4.195	1.084

a. Dependent Variable: IPSA Hypothesize Domain

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	BIOC I Grade	.078[b]	.806	.422	.078	1.000
	BIOC II Grade	.071[b]	.727	.469	.071	1.000
	Content Exam Score	.020[b]	.192	.848	.019	.866
	CAT Score	.233[b]	2.396	.018	.228	.955
	CLASS-Bio Overall	.205[b]	2.121	.036	.203	.978
	Major - Non-Biochemistry	-.081[b]	-.829	.409	-.081	1.000
	Major - Graduate	.041[b]	.422	.674	.041	1.000
	Gender	-.076[b]	-.779	.437	-.076	1.000
	Age Group	.013[b]	.129	.897	.013	1.000
	Hispanic	.028[b]	.285	.776	.028	1.000
	Asian	.074[b]	.764	.447	.074	1.000
	American Indian	-.091[b]	-.932	.353	-.091	1.000
	African American	-.135[b]	-1.395	.166	-.135	1.000
2	Major - Non-Biochemistry	-.029[c]	-.290	.772	-.029	.899
	Major - Graduate	-.006[c]	-.065	.949	-.006	.922
	Gender	-.029[c]	-.296	.768	-.030	.955
	Age Group	-.009[c]	-.091	.927	-.009	.947
	Hispanic	.013[c]	.133	.895	.013	.958
	Asian	.122[c]	1.263	.209	.125	.958
	American Indian	-.086[c]	-.896	.372	-.089	.968
	African American	-.121[c]	-1.264	.209	-.125	.972
3	Gender	-.027[d]	-.275	.784	-.028	.925
	Age Group	-.007[d]	-.073	.942	-.007	.934
	Hispanic	.012[d]	.118	.906	.012	.943
	Asian	.121[d]	1.237	.219	.124	.952
	American Indian	-.088[d]	-.907	.367	-.091	.960
	African American	-.127[d]	-1.280	.203	-.128	.919
4	Hispanic	.010[e]	.099	.921	.010	.923
	Asian	.121[e]	1.227	.223	.124	.951
	American Indian	-.091[e]	-.911	.365	-.093	.940
	African American	-.127[e]	-1.263	.210	-.128	.911

a. Dependent Variable: IPSA Hypothesize Domain

b. Predictors in the Model: (Constant), Time (semesters)

c. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

d. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

e. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

Model fitting: CAT & CLASS-Bio only

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	CLASS-Bio Overall, CAT Score[b]	.	Enter

a. Dependent Variable: IPSA Hypothesize Domain

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.281[a]	.079	.061	1.699	.079	4.493	2	105	.013

a. Predictors: (Constant), CLASS-Bio Overall, CAT Score

b. Dependent Variable: IPSA Hypothesize Domain

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	25.930	2	12.965	4.493	.013[b]
	Residual	302.990	105	2.886		
	Total	328.920	107			

a. Dependent Variable: IPSA Hypothesize Domain

b. Predictors: (Constant), CLASS-Bio Overall, CAT Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2.627	.915		2.870	.005	.812	4.441
	CAT Score	.066	.033	.195	2.026	.045	.001	.130
	CLASS-Bio Overall	.018	.011	.163	1.693	.094	-.003	.039

a. Dependent Variable: IPSA Hypothesize Domain

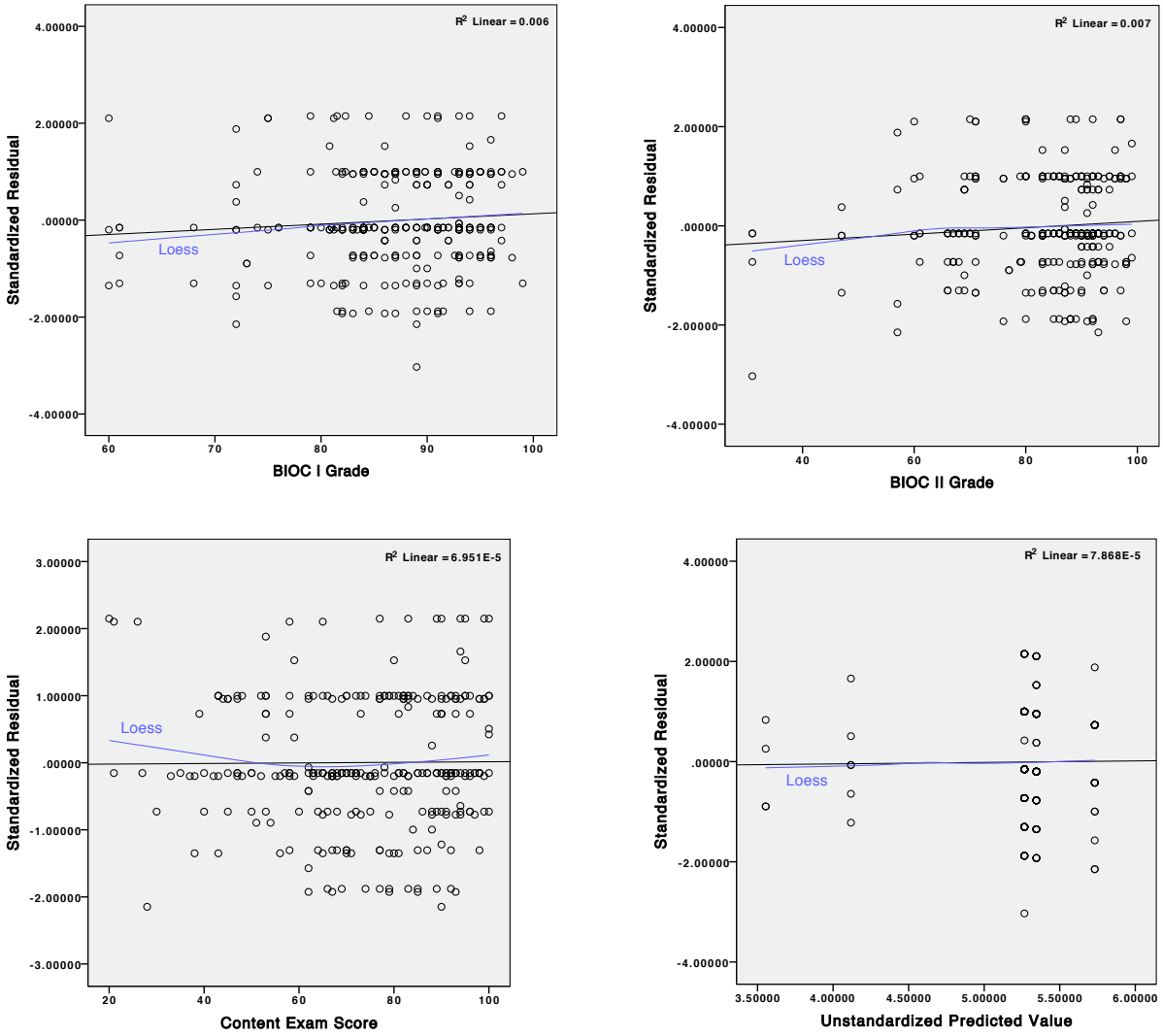
Hypothesize = 2.6 + (0.07)(CAT) + (0.02)(CLASS-Bio)

R2 = 0.08, Adj. R2 = 0.06, F = 4.493, p = 0.013
Intercept B = 2.6, p = 0.005, SE = 0.915, Lower = 0.8, Upper = 4.4
CAT B = 0.07, p = 0.045, SE = 0.033, Lower = 0.00, Upper = 0.13
CLASS-Bio B = 0.02, p = 0.094, SE = 0.011, Lower = 0, Upper = 0.04

Assumptions Testing for Hypothesize:

To test the assumption that the relationship between the IVs and DV has been correctly specified (*i.e.*, that a linear rather than non-linear model is appropriate), the following scatter plots may be evaluated:

- DV as a function of each IV
- Standardized residuals as a function of each IV
- Standardized residuals as a function of predicted Y values



If there is any relationship between predictor values and standardized residuals of the model, then it would indicate that the assumption has not been met. In this case, there is no evidence of a linear relationship between the standardized residuals and any of the three scaled predictors or predicted Y values (R^2 of the linear best fit lines < 0.01). The Loess best fit lines also show that there do not appear to be any non-linear relationships. Therefore,

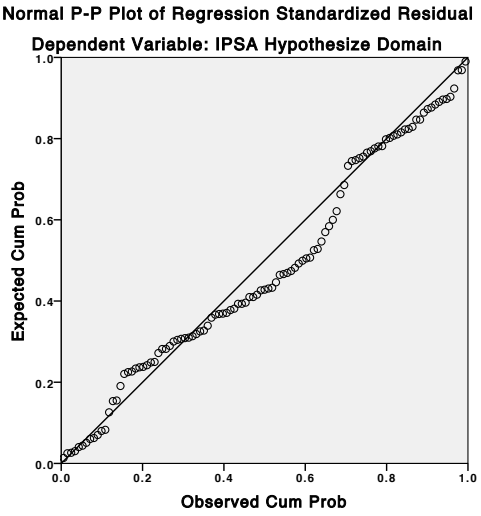
without any type of relationship between these variables, the assumption holds that a linear rather than non-linear model is appropriate.

To test the assumption of constant variance of errors, or homoscedasticity, the same scatter plots may be evaluated:

- Standardized residuals as a function of each IV
- Standardized residuals as a function of predicted Y values

There are no relationships between standardized residuals and either the predictors or predicted values ($R^2 < 0.01$), which supports the assumption overall. The plots of residuals versus predictors show reasonable dispersions. The plot against predicted values shows a slight wedge pattern, which indicates that the assumption of homoscedasticity may not have been met (heteroscedasticity may exist). In this case, a weighted least squares (WLS) estimation could be applied, in order to reduce any bias in standard errors.

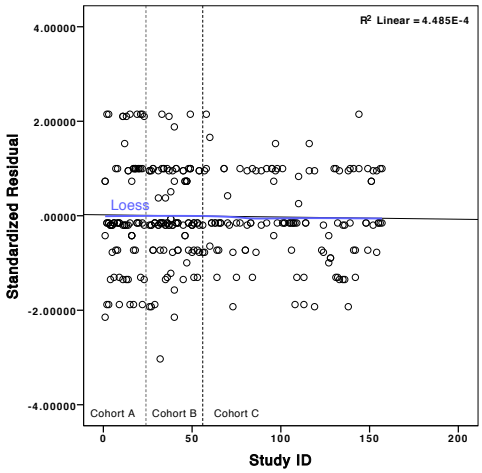
To test the assumption of normality of errors, a probability-probability (p-p) plot is evaluated.



Comparing the normal distribution (diagonal) with the residual distribution (circular markers), the p-p plot shows only slight departures from normality. Therefore, the assumption of normality of errors is met.

To test the assumption of independent errors, the following may be evaluated:

- Index plots of standardized residuals for each member, with the index ordered by any potential nesting factor
- Intraclass correlation coefficient (ICC), as a measure of the proportion of between-group variance to total variance



Considering that Cohorts A and B have four time points plotted per participant, while Cohort C participants only have two data points, the plot shows reasonable dispersion. The assumption of independent errors is met.

SPSS Output for Investigate:

Model Summary[f]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.092[a]	.008	-.010	1.862	.008	.451	1	53	.505
2	.200[b]	.040	-.080	1.925	.032	.317	5	48	.900
3	.265[c]	.070	-.092	1.935	.030	.742	2	46	.482
4	.277[d]	.077	-.133	1.972	.007	.160	2	44	.853
5	.333[e]	.111	-.200	2.029	.034	.386	4	40	.817

a. Predictors: (Constant), Time (semesters)

b. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

c. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

d. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

e. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender, Asian, American Indian, African American, Hispanic

f. Dependent Variable: IPSA Investigate Domain

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.562	1	1.562	.451	.505[b]
	Residual	183.725	53	3.467		
	Total	185.287	54			
2	Regression	7.434	6	1.239	.334	.915[c]
	Residual	177.854	48	3.705		
	Total	185.287	54			
3	Regression	12.991	8	1.624	.434	.895[d]
	Residual	172.296	46	3.746		
	Total	185.287	54			
4	Regression	14.234	10	1.423	.366	.955[e]
	Residual	171.053	44	3.888		
	Total	185.287	54			
5	Regression	20.588	14	1.471	.357	.980[f]
	Residual	164.699	40	4.117		
	Total	185.287	54			

a. Dependent Variable: IPSA Investigate Domain

b. Predictors: (Constant), Time (semesters)

c. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

d. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

e. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

f. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender, Asian, American Indian, African American, Hispanic

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2.968	.391		7.594	.000	2.184	3.752
	Time (semesters)	.115	.171	.092	.671	.505	-.228	.458
2	(Constant)	.092	3.432		.027	.979	-6.808	6.992
	Time (semesters)	-.008	.212	-.006	-.036	.972	-.434	.419
	BIOC I Grade	.030	.045	.121	.676	.502	-.060	.120
	BIOC II Grade	.018	.033	.113	.569	.572	-.047	.084
	Content Exam Score	-.023	.021	-.225	-1.060	.294	-.066	.020
	CAT Score	.023	.054	.065	.435	.666	-.085	.132
	CLASS-Bio Overall	.001	.018	.011	.073	.942	-.035	.037
3	(Constant)	-.388	3.512		-.110	.913	-7.457	6.682
	Time (semesters)	.002	.214	.002	.009	.993	-.429	.433
	BIOC I Grade	.031	.045	.124	.693	.492	-.059	.122
	BIOC II Grade	.024	.033	.144	.718	.477	-.043	.090
	Content Exam Score	-.021	.022	-.208	-.968	.338	-.065	.023
	CAT Score	.022	.054	.062	.409	.684	-.087	.131
	CLASS-Bio Overall	.001	.019	.007	.043	.966	-.037	.039
	Major - Non-Biochemistry	-.870	1.213	-.108	-.718	.477	-3.311	1.570
	Major - Graduate	-1.010	.992	-.151	-1.018	.314	-3.007	.987
4	(Constant)	-.188	3.624		-.052	.959	-7.491	7.116
	Time (semesters)	.011	.219	.009	.050	.960	-.431	.453
	BIOC I Grade	.033	.046	.133	.720	.475	-.060	.127
	BIOC II Grade	.021	.034	.128	.617	.540	-.047	.089
	Content Exam Score	-.020	.023	-.195	-.880	.384	-.065	.026
	CAT Score	.019	.055	.054	.348	.730	-.092	.131
	CLASS-Bio Overall	1.578E-5	.019	.000	.001	.999	-.039	.039
	Major - Non-Biochemistry	-.808	1.244	-.100	-.649	.519	-3.314	1.698
	Major - Graduate	-1.108	1.027	-.166	-1.079	.287	-3.179	.962
	Gender	-.301	.566	-.080	-.533	.597	-1.441	.839
	Age Group	.156	.745	.032	.210	.835	-1.344	1.657
5	(Constant)	-.441	3.852		-.114	.910	-8.226	7.345
	Time (semesters)	.004	.227	.003	.019	.985	-.454	.462
	BIOC I Grade	.032	.049	.127	.651	.519	-.067	.130
	BIOC II Grade	.019	.035	.119	.550	.586	-.052	.091
	Content Exam Score	-.019	.023	-.189	-.818	.418	-.066	.028
	CAT Score	.033	.059	.092	.559	.580	-.086	.151
	CLASS-Bio Overall	.000	.020	-.001	-.008	.994	-.041	.040
	Major - Non-Biochemistry	-.653	1.290	-.081	-.506	.616	-3.259	1.954
	Major - Graduate	-1.151	1.102	-.172	-1.045	.303	-3.379	1.076
	Gender	-.290	.583	-.077	-.498	.621	-1.469	.888
	Age Group	.229	.783	.046	.293	.771	-1.353	1.812
	Hispanic	.358	.644	.089	.555	.582	-.944	1.660
	Asian	1.263	1.109	.179	1.139	.262	-.978	3.504
	American Indian	.567	2.120	.041	.268	.790	-3.718	4.852
	African American	-.442	2.169	-.032	-.204	.840	-4.825	3.942

a. Dependent Variable: IPSA Investigate Domain

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	BIOC I Grade	.079[b]	.574	.569	.079	1.000
	BIOC II Grade	.058[b]	.421	.676	.058	1.000
	Content Exam Score	-.067[b]	-.449	.655	-.062	.866
	CAT Score	.044[b]	.313	.755	.043	.955
	CLASS-Bio Overall	-.009[b]	-.063	.950	-.009	.978
	Major - Non-Biochemistry	-.095[b]	-.692	.492	-.096	1.000
	Major - Graduate	-.124[b]	-.905	.370	-.125	1.000
	Gender	-.092[b]	-.667	.508	-.092	1.000
	Age Group	-.016[b]	-.114	.910	-.016	1.000
	Hispanic	.063[b]	.460	.648	.064	1.000
	Asian	.141[b]	1.032	.307	.142	1.000
	American Indian	.045[b]	.325	.747	.045	1.000
	African American	-.091[b]	-.659	.513	-.091	1.000
2	Major - Non-Biochemistry	-.100[c]	-.669	.507	-.097	.899
	Major - Graduate	-.146[c]	-.989	.328	-.143	.922
	Gender	-.068[c]	-.464	.645	-.067	.955
	Age Group	.008[c]	.054	.957	.008	.947
	Hispanic	.044[c]	.302	.764	.044	.958
	Asian	.158[c]	1.099	.277	.158	.958
	American Indian	.047[c]	.324	.747	.047	.968
	African American	-.079[c]	-.545	.588	-.079	.972
3	Gender	-.079[d]	-.531	.598	-.079	.925
	Age Group	.028[d]	.190	.850	.028	.934
	Hispanic	.052[d]	.352	.726	.052	.943
	Asian	.158[d]	1.089	.282	.160	.952
	American Indian	.032[d]	.218	.828	.033	.960
	African American	-.051[d]	-.344	.733	-.051	.919
4	Hispanic	.055[e]	.359	.722	.055	.923
	Asian	.160[e]	1.082	.285	.163	.951
	American Indian	.025[e]	.165	.869	.025	.940
	African American	-.054[e]	-.349	.729	-.053	.911

a. Dependent Variable: IPSA Investigate Domain

b. Predictors in the Model: (Constant), Time (semesters)

c. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

d. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

e. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

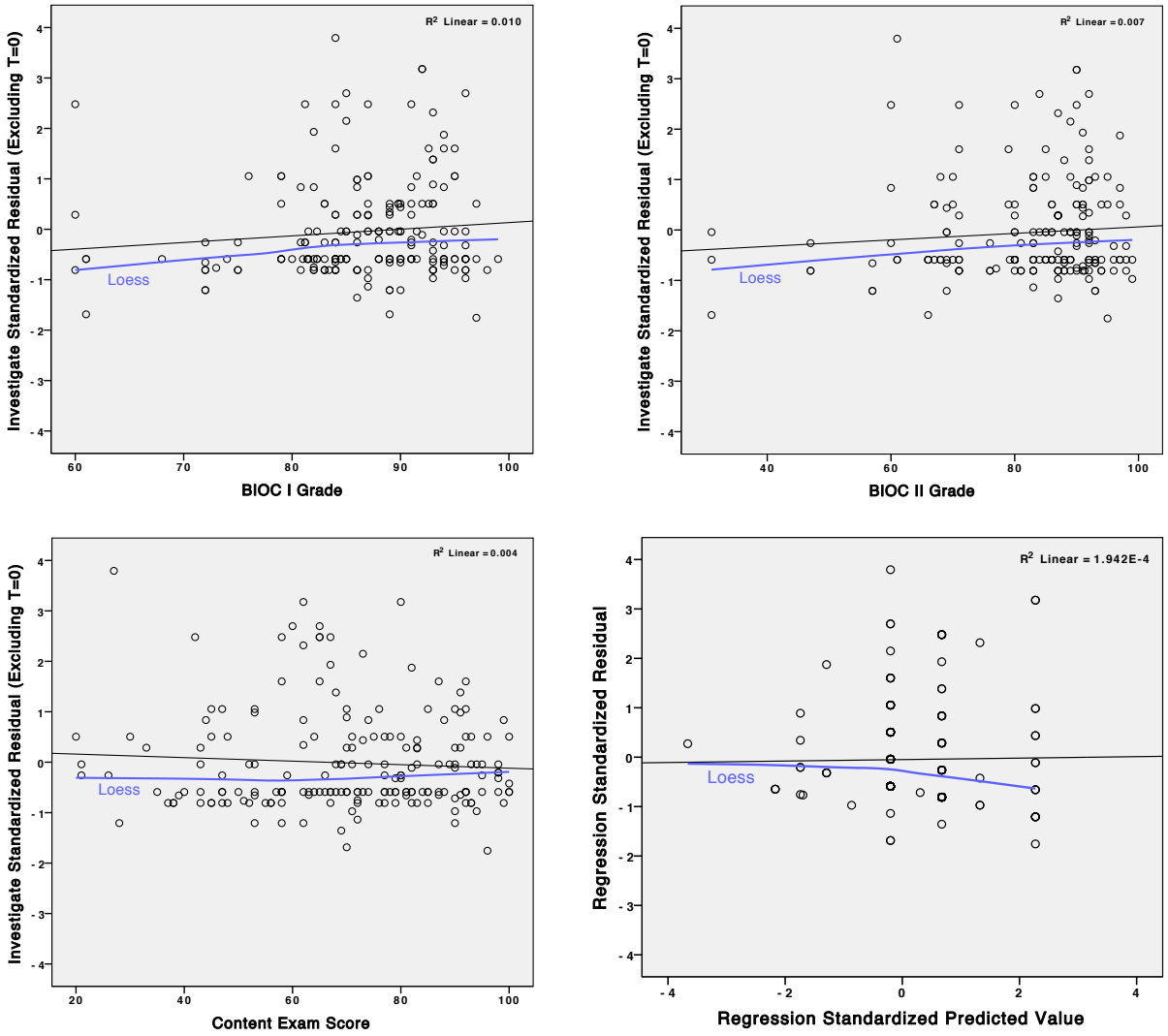
Use Model 1, but exclude time since it is not statistically significant:
Investigate = 3.0

R2 = 0.01, Adj. R2 = -0.01, F = 0.451, p = 0.505
Intercept B = 3.0, $p < 0.001$, SE = 0.39, Lower = 2.2, Upper = 3.8

Assumptions Testing for Investigate:

To test the assumption that the relationship between the IVs and DV has been correctly specified (*i.e.*, that a linear rather than non-linear model is appropriate), the following scatter plots may be evaluated:

- DV as a function of each IV
- Standardized residuals as a function of each IV
- Standardized residuals as a function of predicted Y values



If there is any relationship between predictor values and standardized residuals of the model, then it would indicate that the assumption has not been met. In this case, there is no evidence of a linear relationship between the standardized residuals and any of the three scaled predictors or predicted Y values (R^2 of the linear best fit lines ≤ 0.01). The Loess best fit lines also show that there do not appear to be any non-linear relationships. Therefore,

without any type of relationship between these variables, the assumption holds that a linear rather than non-linear model is appropriate.

To test the assumption of constant variance of errors, or homoscedasticity, the same scatter plots may be evaluated:

- Standardized residuals as a function of each IV
- Standardized residuals as a function of predicted Y values

There are no relationships between standardized residuals and either the predictors or predicted values ($R^2 \leq 0.01$), which supports the assumption overall. The plots of residuals versus predictors show reasonable dispersions. The plot against predicted values shows a slight wedge pattern, which indicates that the assumption of homoscedasticity may not have been met (heteroscedasticity may exist).

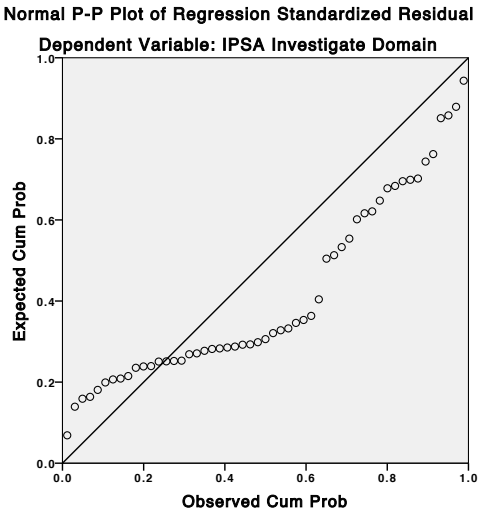
To evaluate the magnitude of nonconstant variance, residuals were divided at the median into two levels: low and high. Computing the ratio of the variances of the two groups determines the magnitude; more than a ten-fold difference calls for an alternate approach.

Descriptives

	Investigate (C) Standardized Residual (Excluding T=0) (Binned)			Statistic	Std. Error
Investigate (C) Standardized Residual (Excluding T=0)	<= -.27204	Mean		-.6497031	.04899599
		95% Confidence Interval for Mean	Lower Bound	-.7506122	
			Upper Bound	-.5487939	
		5% Trimmed Mean		-.6289689	
		Median		-.6124444	
		Variance		.062	
		Std. Deviation		.24983151	
		Minimum		-1.44748	
		Maximum		-.30958	
		Range		1.13791	
		Interquartile Range		.29269	
		Skewness		-1.315	.456
		Kurtosis		3.015	.887
	-.27203+	Mean		.6945170	.15946413
		95% Confidence Interval for Mean	Lower Bound	.3667338	
			Upper Bound	1.0223002	
		5% Trimmed Mean		.6425733	
		Median		.6259393	
		Variance		.687	
		Std. Deviation		.82859993	
		Minimum		-.27204	
		Maximum		2.81556	
		Range		3.08760	
		Interquartile Range		1.42542	
		Skewness		.686	.448
		Kurtosis		-.209	.872

With a ratio of 11.1 (0.687/0.062), a weighted least squares (WLS) regression could be performed, in order to reduce any bias in standard errors of the regression coefficients. However, the ordinary least squares (OLS) regression presented here is preferable to WLS with small sample sizes such as in this study. Moreover, the meaning of the model R^2 generated by WLS is inconsistent with the meaning under OLS. Therefore, WLS regression was not performed. The coefficient estimates are not biased, but their significance tests and confidence intervals may be biased.

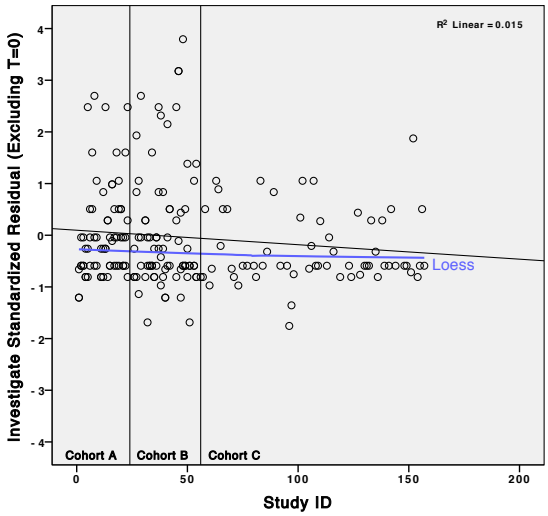
To test the assumption of normality of errors, a probability-probability (p-p) plot is evaluated.



Comparing the normal distribution (diagonal) with the residual distribution (circular markers), the p-p plot shows only slight departures from normality. Therefore, the assumption of normality of errors is met.

To test the assumption of independent errors, the following may be evaluated:

- Index plots of standardized residuals for each member, with the index ordered by any potential nesting factor
- Intraclass correlation coefficient (ICC), as a measure of the proportion of between-group variance to total variance



Considering that Cohorts A and B have four time points plotted per participant, while Cohort C participants only have two data points, the plot shows reasonable dispersion. The assumption of independent errors is met.

SPSS Output for Evaluate:

Model Summary[f]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.336[a]	.113	.105	2.049	.113	13.508	1	106	.000
2	.421[b]	.177	.129	2.022	.064	1.582	5	101	.172
3	.423[c]	.179	.112	2.040	.001	.080	2	99	.923
4	.424[d]	.180	.095	2.060	.001	.075	2	97	.928
5	.448[e]	.201	.081	2.076	.021	.613	4	93	.654

a. Predictors: (Constant), Time (semesters)

b. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

c. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

d. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

e. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender, Asian, American Indian, African American, Hispanic

f. Dependent Variable: IPSA Evaluate Domain

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	56.728	1	56.728	13.508	.000[b]
	Residual	445.140	106	4.199		
	Total	501.868	107			
2	Regression	89.050	6	14.842	3.631	.003[c]
	Residual	412.818	101	4.087		
	Total	501.868	107			
3	Regression	89.716	8	11.215	2.694	.010[d]
	Residual	412.152	99	4.163		
	Total	501.868	107			
4	Regression	90.350	10	9.035	2.130	.029[e]
	Residual	411.518	97	4.242		
	Total	501.868	107			
5	Regression	100.927	14	7.209	1.672	.075[f]
	Residual	400.941	93	4.311		
	Total	501.868	107			

a. Dependent Variable: IPSA Evaluate Domain

b. Predictors: (Constant), Time (semesters)

c. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

d. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

e. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

f. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender, Asian, American Indian, African American, Hispanic

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.350	.306		14.207	.000	3.743	4.957
	Time (semesters)	.492	.134	.336	3.675	.000	.227	.757
2	(Constant)	-.473	2.561		-.185	.854	-5.552	4.607
	Time (semesters)	.435	.158	.298	2.749	.007	.121	.750
	BIOC I Grade	.020	.033	.069	.607	.545	-.046	.086
	BIOC II Grade	.003	.024	.016	.123	.902	-.045	.051
	Content Exam Score	.004	.016	.031	.232	.817	-.028	.036
	CAT Score	.055	.040	.131	1.366	.175	-.025	.134
	CLASS-Bio Overall	.020	.013	.150	1.529	.129	-.006	.047
3	(Constant)	-.459	2.631		-.175	.862	-5.679	4.760
	Time (semesters)	.441	.161	.301	2.745	.007	.122	.759
	BIOC I Grade	.020	.034	.068	.590	.557	-.047	.087
	BIOC II Grade	.004	.025	.020	.158	.875	-.045	.053
	Content Exam Score	.004	.016	.037	.272	.786	-.028	.037
	CAT Score	.055	.041	.131	1.354	.179	-.026	.135
	CLASS-Bio Overall	.019	.014	.141	1.343	.182	-.009	.047
	Major - Non-Biochemistry	-.354	.908	-.037	-.390	.698	-2.156	1.448
	Major - Graduate	-.081	.743	-.010	-.109	.913	-1.556	1.393
4	(Constant)	-.553	2.689		-.206	.838	-5.891	4.785
	Time (semesters)	.436	.163	.298	2.680	.009	.113	.759
	BIOC I Grade	.019	.034	.064	.544	.588	-.049	.087
	BIOC II Grade	.005	.025	.027	.207	.836	-.045	.055
	Content Exam Score	.004	.017	.032	.230	.818	-.029	.037
	CAT Score	.056	.041	.135	1.369	.174	-.025	.138
	CLASS-Bio Overall	.019	.014	.144	1.349	.181	-.009	.048
	Major - Non-Biochemistry	-.384	.923	-.041	-.416	.678	-2.216	1.447
	Major - Graduate	-.030	.762	-.004	-.040	.968	-1.544	1.483
	Gender	.150	.420	.034	.356	.722	-.684	.983
	Age Group	-.091	.553	-.016	-.164	.870	-1.188	1.006
5	(Constant)	-.432	2.800		-.154	.878	-5.992	5.129
	Time (semesters)	.432	.165	.295	2.621	.010	.105	.759
	BIOC I Grade	.017	.035	.057	.469	.640	-.054	.087
	BIOC II Grade	.002	.026	.012	.088	.930	-.049	.053
	Content Exam Score	.004	.017	.036	.248	.805	-.030	.038
	CAT Score	.066	.043	.158	1.547	.125	-.019	.151
	CLASS-Bio Overall	.019	.015	.139	1.288	.201	-.010	.048
	Major - Non-Biochemistry	-.281	.937	-.030	-.300	.765	-2.143	1.581
	Major - Graduate	.008	.801	.001	.010	.992	-1.583	1.599
	Gender	.186	.424	.042	.438	.662	-.656	1.027
	Age Group	-.171	.569	-.030	-.301	.764	-1.302	.959
	Hispanic	.241	.468	.051	.515	.608	-.689	1.172
	Asian	.188	.806	.023	.234	.816	-1.412	1.789
	American Indian	2.327	1.541	.145	1.510	.134	-.733	5.388
	African American	.277	1.577	.017	.175	.861	-2.854	3.407

a. Dependent Variable: IPSA Evaluate Domain

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	BIOC I Grade	.118[b]	1.293	.199	.125	1.000
	BIOC II Grade	.103[b]	1.131	.261	.110	1.000
	Content Exam Score	.153[b]	1.568	.120	.151	.866
	CAT Score	.169[b]	1.828	.070	.176	.955
	CLASS-Bio Overall	.199[b]	2.189	.031	.209	.978
	Major - Non-Biochemistry	-.067[b]	-.733	.465	-.071	1.000
	Major - Graduate	.047[b]	.511	.611	.050	1.000
	Gender	.007[b]	.078	.938	.008	1.000
	Age Group	.017[b]	.189	.850	.018	1.000
	Hispanic	.025[b]	.269	.788	.026	1.000
	Asian	-.019[b]	-.208	.836	-.020	1.000
	American Indian	.140[b]	1.546	.125	.149	1.000
	African American	-.011[b]	-.123	.902	-.012	1.000
2	Major - Non-Biochemistry	-.037[c]	-.387	.700	-.039	.899
	Major - Graduate	-.009[c]	-.091	.928	-.009	.922
	Gender	.030[c]	.320	.750	.032	.955
	Age Group	-.016[c]	-.176	.861	-.018	.947
	Hispanic	.037[c]	.402	.688	.040	.958
	Asian	.006[c]	.062	.950	.006	.958
	American Indian	.135[c]	1.481	.142	.147	.968
	African American	.006[c]	.071	.944	.007	.972
3	Gender	.033[d]	.352	.726	.035	.925
	Age Group	-.014[d]	-.151	.881	-.015	.934
	Hispanic	.036[d]	.383	.702	.039	.943
	Asian	.004[d]	.039	.969	.004	.952
	American Indian	.134[d]	1.451	.150	.145	.960
	African American	.008[d]	.083	.934	.008	.919
4	Hispanic	.036[e]	.374	.709	.038	.923
	Asian	.003[e]	.030	.976	.003	.951
	American Indian	.141[e]	1.492	.139	.151	.940
	African American	.009[e]	.091	.927	.009	.911

a. Dependent Variable: IPSA Evaluate Domain

b. Predictors in the Model: (Constant), Time (semesters)

c. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

d. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

e. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

Model fitting: Time, CAT, CLASS-Bio, American Indian

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Time (semesters)[b]	.	Enter
2	CLASS-Bio Overall, CAT Score[b]	.	Enter
3	American Indian[b]	.	Enter
4	.[b]	CAT Score[c]	Remove
5	.[b]	American Indian[c]	Remove

a. Dependent Variable: IPSA Evaluate Domain

b. All requested variables entered.

c. All requested variables removed.

Model Summary[f]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.336[a]	.113	.105	2.049	.113	13.508	1	106	.000
2	.410[b]	.168	.144	2.003	.055	3.449	2	104	.035
3	.435[c]	.189	.157	1.988	.021	2.633	1	103	.108
4	.410[d]	.168	.144	2.003	-.021	2.623	1	103	.108
5	.390[e]	.152	.136	2.014	-.017	2.067	1	104	.153

a. Predictors: (Constant), Time (semesters)

b. Predictors: (Constant), Time (semesters), CLASS-Bio Overall, CAT Score

c. Predictors: (Constant), Time (semesters), CLASS-Bio Overall, CAT Score, American Indian

d. Predictors: (Constant), Time (semesters), CLASS-Bio Overall, American Indian

e. Predictors: (Constant), Time (semesters), CLASS-Bio Overall

f. Dependent Variable: IPSA Evaluate Domain

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	56.728	1	56.728	13.508	.000[b]
	Residual	445.140	106	4.199		
	Total	501.868	107			
2	Regression	84.418	3	28.139	7.010	.000[c]
	Residual	417.450	104	4.014		
	Total	501.868	107			
3	Regression	94.823	4	23.706	5.999	.000[d]
	Residual	407.045	103	3.952		
	Total	501.868	107			
4	Regression	84.455	3	28.152	7.014	.000[e]
	Residual	417.412	104	4.014		
	Total	501.868	107			
5	Regression	76.158	2	38.079	9.392	.000[f]
	Residual	425.710	105	4.054		
	Total	501.868	107			

a. Dependent Variable: IPSA Evaluate Domain

b. Predictors: (Constant), Time (semesters)

c. Predictors: (Constant), Time (semesters), CLASS-Bio Overall, CAT Score

d. Predictors: (Constant), Time (semesters), CLASS-Bio Overall, CAT Score, American Indian

e. Predictors: (Constant), Time (semesters), CLASS-Bio Overall, American Indian

f. Predictors: (Constant), Time (semesters), CLASS-Bio Overall

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.350	.306		14.207	.000	3.743	4.957
	Time (semesters)	.492	.134	.336	3.675	.000	.227	.757
2	(Constant)	1.627	1.079		1.507	.135	-.513	3.768
	Time (semesters)	.413	.135	.282	3.067	.003	.146	.680
	CAT Score	.056	.039	.134	1.435	.154	-.021	.133
	CLASS-Bio Overall	.023	.012	.172	1.865	.065	-.001	.048
3	(Constant)	1.574	1.072		1.469	.145	-.551	3.699
	Time (semesters)	.410	.134	.280	3.071	.003	.145	.675
	CAT Score	.063	.039	.151	1.620	.108	-.014	.141
	CLASS-Bio Overall	.022	.012	.160	1.738	.085	-.003	.046
	American Indian	2.330	1.436	.145	1.623	.108	-.518	5.177
4	(Constant)	2.441	.936		2.609	.010	.586	4.296
	Time (semesters)	.450	.132	.308	3.404	.001	.188	.713
	CLASS-Bio Overall	.026	.012	.191	2.107	.037	.002	.050
	American Indian	2.067	1.438	.129	1.438	.153	-.784	4.918
5	(Constant)	2.401	.940		2.554	.012	.537	4.264
	Time (semesters)	.449	.133	.307	3.374	.001	.185	.712
	CLASS-Bio Overall	.027	.012	.199	2.189	.031	.003	.051

a. Dependent Variable: IPSA Evaluate Domain

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	CAT Score	.169[b]	1.828	.070	.176	.955
	CLASS-Bio Overall	.199[b]	2.189	.031	.209	.978
	American Indian	.140[b]	1.546	.125	.149	1.000
2	American Indian	.145[c]	1.623	.108	.158	.983
4	CAT Score	.151[d]	1.620	.108	.158	.905
5	CAT Score	.134[e]	1.435	.154	.139	.916
	American Indian	.129[e]	1.438	.153	.140	.996

a. Dependent Variable: IPSA Evaluate Domain

b. Predictors in the Model: (Constant), Time (semesters)

c. Predictors in the Model: (Constant), Time (semesters), CLASS-Bio Overall, CAT Score

d. Predictors in the Model: (Constant), Time (semesters), CLASS-Bio Overall, American Indian

e. Predictors in the Model: (Constant), Time (semesters), CLASS-Bio Overall

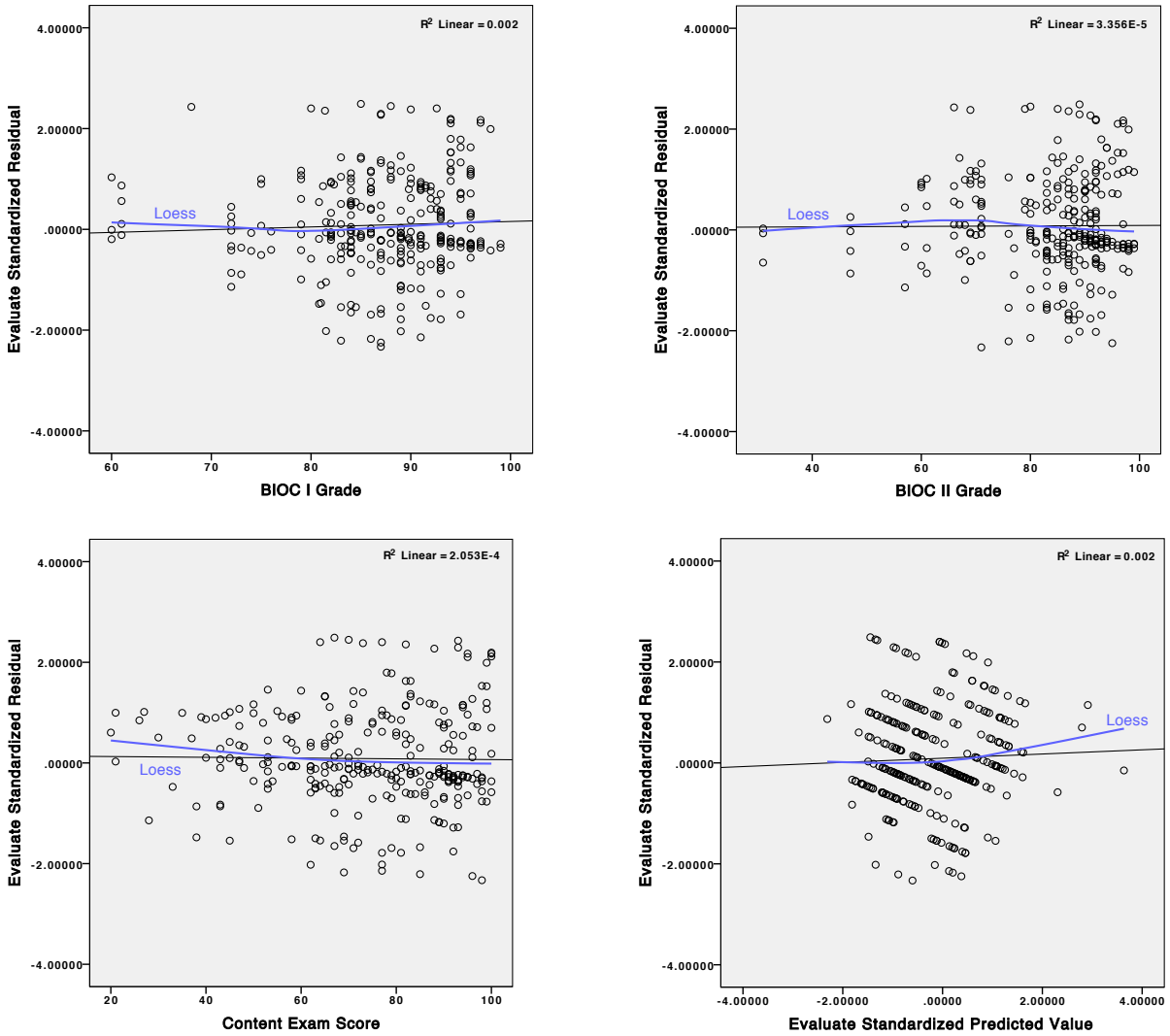
Evaluate = 2.4 + (0.4)(Semester) + (0.03)(CLASS-Bio)

$R2 = 0.15$, Adj. $R2 = 0.14$, $F = 9.392$, $p < 0.001$
Intercept B = 2.4, $p = 0.012$, SE = 0.940, Lower = 0.5, Upper = 4.3
Time B = 0.5, $p = 0.001$, SE = 0.133, Lower = 0.2, Upper = 0.7
CLASS-Bio B = 0.03, $p = 0.031$, SE = 0.012, Lower = 0.00, Upper = 0.05

Assumptions Testing for Evaluate:

To test the assumption that the relationship between the IVs and DV has been correctly specified (*i.e.*, that a linear rather than non-linear model is appropriate), the following scatter plots may be evaluated:

- DV as a function of each IV
- Standardized residuals as a function of each IV
- Standardized residuals as a function of predicted Y values



If there is any relationship between predictor values and standardized residuals of the model, then it would indicate that the assumption has not been met. In this case, there is no evidence of a linear relationship between the standardized residuals and any of the three scaled predictors or predicted Y values (R^2 of the linear best fit lines < 0.01). The Loess best fit lines also show that there do not appear to be any non-linear relationships. Therefore,

without any type of relationship between these variables, the assumption holds that a linear rather than non-linear model is appropriate.

To test the assumption of constant variance of errors, or homoscedasticity, the same scatter plots may be evaluated:

- Standardized residuals as a function of each IV
- Standardized residuals as a function of predicted Y values

There are no relationships between standardized residuals and either the predictors or predicted values ($R^2 < 0.01$), which supports the assumption overall. However, all plots show slight wedge patterns, which indicates that the assumption of homoscedasticity may not have been met (heteroscedasticity may exist).

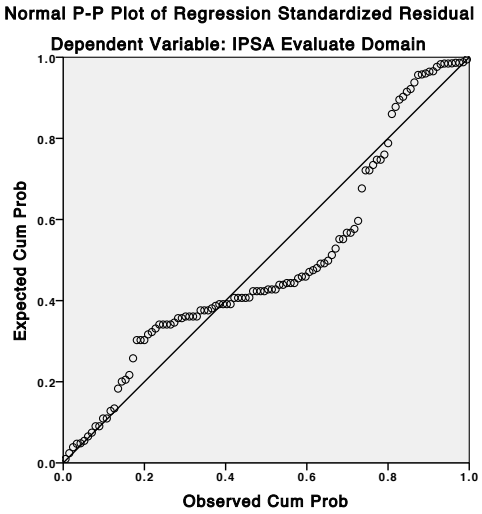
To evaluate the magnitude of nonconstant variance, residuals were divided at the median into two levels: low and high. Computing the ratio of the variances of the two groups determines the magnitude; more than a ten-fold difference calls for an alternate approach.

Descriptives

	Evaluate (C) Standardized Residual (Binned)			Statistic	Std. Error
Evaluate (C) Standardized Residual	<= -.16988	Mean		-.7135899	.06523142
		95% Confidence Interval for Mean	Lower Bound	-.8444863	
			Upper Bound	-.5826935	
		5% Trimmed Mean		-.6848809	
		Median		-.5348656	
		Variance		.226	
		Std. Deviation		.47489194	
		Minimum		-1.86435	
		Maximum		-.17923	
		Range		1.68511	
		Interquartile Range		.63189	
		Skewness		-.959	.327
		Kurtosis		-.296	.644
	-.16987+	Mean		.6958028	.10301088
		95% Confidence Interval for Mean	Lower Bound	.4890964	
			Upper Bound	.9025093	
		5% Trimmed Mean		.6502868	
		Median		.5242205	
		Variance		.562	
		Std. Deviation		.74993053	
		Minimum		-.16052	
		Maximum		2.46376	
		Range		2.62429	
		Interquartile Range		1.25261	
		Skewness		.803	.327
		Kurtosis		-.434	.644

With a ratio of 2.5 (0.562/0.226), a weighted least squares (WLS) regression is not necessary.

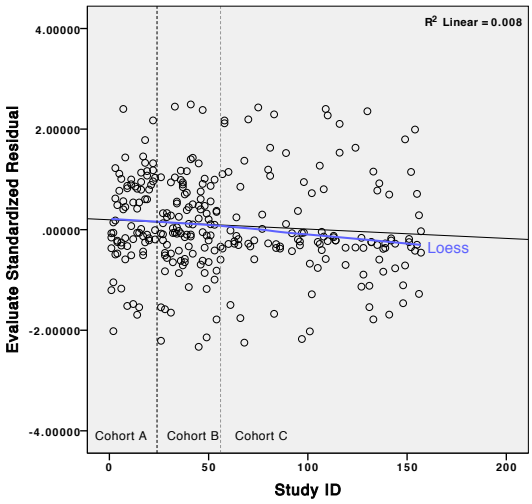
To test the assumption of normality of errors, a probability-probability (p-p) plot is evaluated.



Comparing the normal distribution (diagonal) with the residual distribution (circular markers), the p-p plot shows only slight departures from normality. Therefore, the assumption of normality of errors is met.

To test the assumption of independent errors, the following may be evaluated:

- Index plots of standardized residuals for each member, with the index ordered by any potential nesting factor
- Intraclass correlation coefficient (ICC), as a measure of the proportion of between-group variance to total variance



Considering that Cohorts A and B have four time points plotted per participant, while Cohort C participants only have two data points, the plot shows reasonable dispersion. The assumption of independent errors is met.

SPSS Output for Integrate:

Model Summary[f]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.190[a]	.036	.027	2.561	.036	3.950	1	106	.049
2	.475[b]	.226	.180	2.351	.190	4.962	5	101	.000
3	.482[c]	.232	.170	2.365	.006	.390	2	99	.678
4	.487[d]	.237	.159	2.381	.005	.324	2	97	.724
5	.493[e]	.243	.129	2.422	.006	.187	4	93	.945

a. Predictors: (Constant), Time (semesters)

b. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

c. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

d. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

e. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender, Asian, American Indian, African American, Hispanic

f. Dependent Variable: IPSA Integrate Domain

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	25.907	1	25.907	3.950	.049[b]
	Residual	695.224	106	6.559		
	Total	721.130	107			
2	Regression	162.997	6	27.166	4.916	.000[c]
	Residual	558.133	101	5.526		
	Total	721.130	107			
3	Regression	167.361	8	20.920	3.740	.001[d]
	Residual	553.769	99	5.594		
	Total	721.130	107			
4	Regression	171.033	10	17.103	3.016	.002[e]
	Residual	550.098	97	5.671		
	Total	721.130	107			
5	Regression	175.418	14	12.530	2.135	.016[f]
	Residual	545.712	93	5.868		
	Total	721.130	107			

a. Dependent Variable: IPSA Integrate Domain

b. Predictors: (Constant), Time (semesters)

c. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

d. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

e. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

f. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender, Asian, American Indian, African American, Hispanic

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	5.883	.383		15.375	.000	5.125	6.642
	Time (semesters)	.332	.167	.190	1.987	.049	.001	.664
2	(Constant)	-2.990	2.977		-1.004	.318	-8.896	2.916
	Time (semesters)	.276	.184	.157	1.497	.138	-.090	.641
	BIOC I Grade	.039	.039	.112	1.016	.312	-.037	.116
	BIOC II Grade	-.011	.028	-.049	-.399	.691	-.067	.045
	Content Exam Score	.021	.019	.147	1.119	.266	-.016	.058
	CAT Score	.143	.047	.286	3.072	.003	.051	.236
	CLASS-Bio Overall	.028	.015	.172	1.802	.075	-.003	.058
3	(Constant)	-2.622	3.049		-.860	.392	-8.672	3.429
	Time (semesters)	.272	.186	.155	1.463	.147	-.097	.641
	BIOC I Grade	.038	.039	.109	.983	.328	-.039	.116
	BIOC II Grade	-.014	.029	-.063	-.506	.614	-.071	.042
	Content Exam Score	.020	.019	.141	1.062	.291	-.017	.057
	CAT Score	.144	.047	.288	3.072	.003	.051	.238
	CLASS-Bio Overall	.027	.016	.169	1.662	.100	-.005	.060
	Major - Non-Biochemistry	.406	1.053	.036	.386	.700	-1.683	2.495
	Major - Graduate	.700	.861	.075	.812	.419	-1.009	2.409
4	(Constant)	-2.288	3.110		-.736	.464	-8.460	3.884
	Time (semesters)	.282	.188	.161	1.498	.137	-.091	.655
	BIOC I Grade	.035	.040	.100	.884	.379	-.044	.114
	BIOC II Grade	-.016	.029	-.071	-.557	.579	-.074	.042
	Content Exam Score	.022	.019	.158	1.163	.248	-.016	.061
	CAT Score	.143	.048	.285	3.003	.003	.048	.237
	CLASS-Bio Overall	.029	.017	.177	1.720	.089	-.004	.062
	Major - Non-Biochemistry	.475	1.067	.042	.445	.658	-1.643	2.592
	Major - Graduate	.730	.882	.078	.828	.410	-1.020	2.480
	Gender	-.164	.485	-.031	-.337	.737	-1.127	.800
	Age Group	-.458	.639	-.066	-.716	.476	-1.726	.810
5	(Constant)	-2.593	3.267		-.794	.429	-9.080	3.894
	Time (semesters)	.272	.192	.155	1.415	.160	-.110	.654
	BIOC I Grade	.038	.041	.107	.907	.367	-.045	.120
	BIOC II Grade	-.016	.030	-.071	-.544	.588	-.076	.043
	Content Exam Score	.022	.020	.152	1.087	.280	-.018	.061
	CAT Score	.150	.050	.299	3.011	.003	.051	.249
	CLASS-Bio Overall	.029	.017	.179	1.709	.091	-.005	.063
	Major - Non-Biochemistry	.530	1.094	.047	.484	.629	-1.642	2.702
	Major - Graduate	.623	.935	.066	.667	.507	-1.233	2.479
	Gender	-.164	.495	-.031	-.333	.740	-1.147	.818
	Age Group	-.534	.664	-.077	-.805	.423	-1.853	.784
	Hispanic	-.078	.546	-.014	-.142	.887	-1.163	1.007
	Asian	.342	.940	.035	.363	.717	-1.526	2.209
	American Indian	.596	1.798	.031	.332	.741	-2.974	4.167
	African American	1.240	1.839	.064	.674	.502	-2.413	4.892

a. Dependent Variable: IPSA Integrate Domain

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	BIOC I Grade	.183[b]	1.943	.055	.186	1.000
	BIOC II Grade	.152[b]	1.608	.111	.155	1.000
	Content Exam Score	.284[b]	2.869	.005	.270	.866
	CAT Score	.340[b]	3.680	.000	.338	.955
	CLASS-Bio Overall	.286[b]	3.079	.003	.288	.978
	Major - Non-Biochemistry	-.010[b]	-.108	.914	-.011	1.000
	Major - Graduate	.144[b]	1.524	.130	.147	1.000
	Gender	-.064[b]	-.667	.506	-.065	1.000
	Age Group	.000[b]	.000	1.000	.000	1.000
	Hispanic	-.042[b]	-.444	.658	-.043	1.000
	Asian	-.019[b]	-.195	.845	-.019	1.000
	American Indian	.020[b]	.212	.832	.021	1.000
	African American	.035[b]	.368	.713	.036	1.000
2	Major - Non-Biochemistry	.032[c]	.347	.729	.035	.899
	Major - Graduate	.073[c]	.798	.427	.080	.922
	Gender	-.039[c]	-.434	.665	-.043	.955
	Age Group	-.057[c]	-.630	.530	-.063	.947
	Hispanic	-.015[c]	-.164	.870	-.016	.958
	Asian	.034[c]	.377	.707	.038	.958
	American Indian	.014[c]	.158	.874	.016	.968
	African American	.070[c]	.783	.435	.078	.972
3	Gender	-.034[d]	-.367	.714	-.037	.925
	Age Group	-.067[d]	-.734	.465	-.074	.934
	Hispanic	-.020[d]	-.219	.827	-.022	.943
	Asian	.033[d]	.369	.713	.037	.952
	American Indian	.021[d]	.236	.814	.024	.960
	African American	.057[d]	.621	.536	.063	.919
4	Hispanic	-.031[e]	-.334	.739	-.034	.923
	Asian	.032[e]	.350	.727	.036	.951
	American Indian	.029[e]	.321	.749	.033	.940
	African American	.064[e]	.690	.492	.070	.911

a. Dependent Variable: IPSA Integrate Domain

b. Predictors in the Model: (Constant), Time (semesters)

c. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

d. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

e. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

Model fitting: CAT, CLASS-Bio

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	CLASS-Bio Overall, CAT Score[b]	.	Enter

a. Dependent Variable: IPSA Integrate Domain

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.432[a]	.186	.171	2.364	.186	12.012	2	105	.000

a. Predictors: (Constant), CLASS-Bio Overall, CAT Score

b. Dependent Variable: IPSA Integrate Domain

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	134.273	2	67.137	12.012	.000[b]
	Residual	586.857	105	5.589		
	Total	721.130	107			

a. Dependent Variable: IPSA Integrate Domain

b. Predictors: (Constant), CLASS-Bio Overall, CAT Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.492	1.274		.386	.700	-2.033	3.018
	CAT Score	.156	.045	.311	3.438	.001	.066	.246
	CLASS-Bio Overall	.038	.015	.237	2.624	.010	.009	.067

a. Dependent Variable: IPSA Integrate Domain

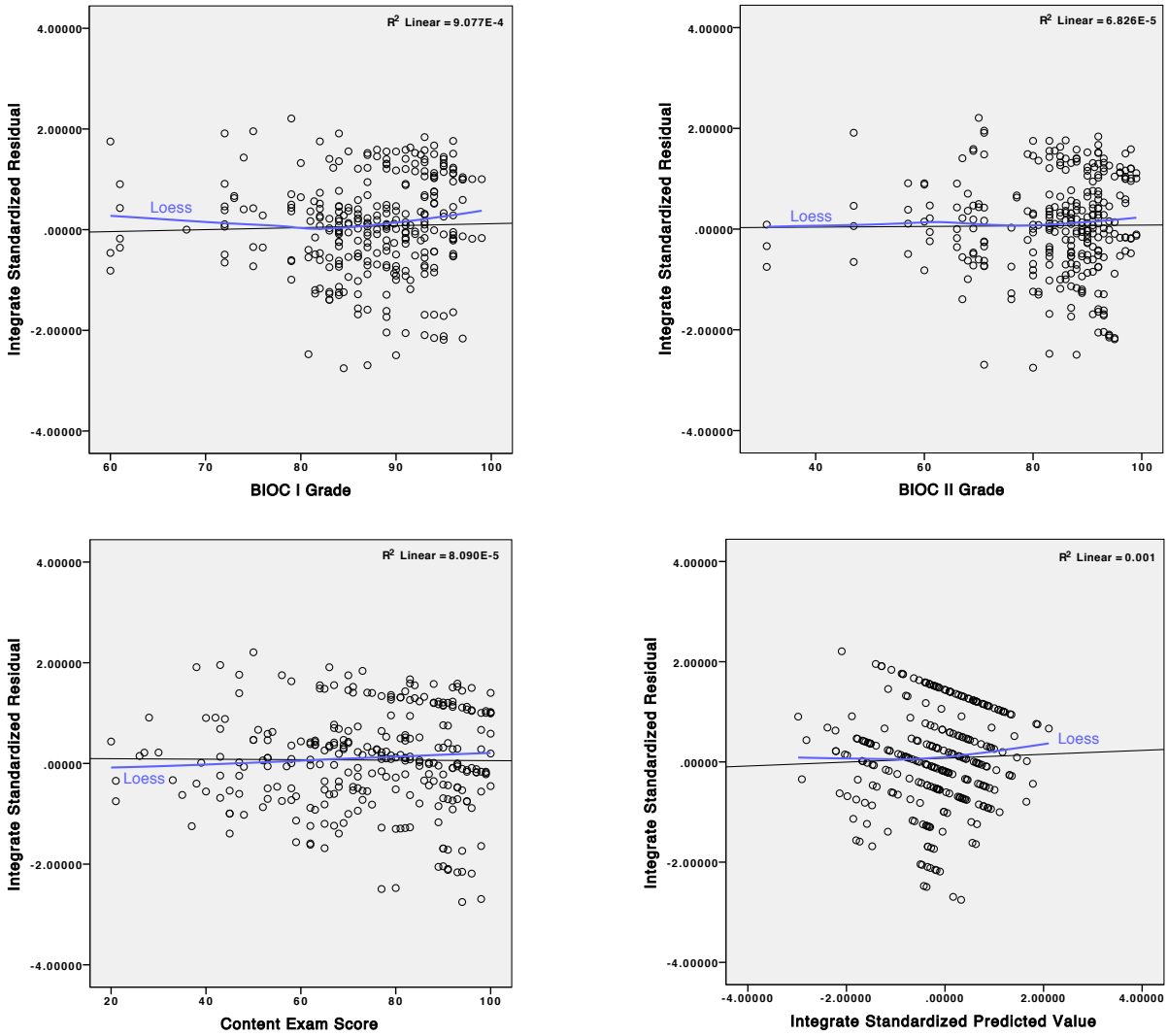
Integrate = 0.5 + (0.16)(CAT) + (0.04)(CLASS-Bio)

R2 = 0.19, Adj. R2 = 0.17, F = 12.012, $p < 0.001$
Intercept B = 0.5, p = 0.700, SE = 1.274, Lower = -2.0, Upper = 3.0
CAT B = 0.16, p = 0.001, SE = 0.045, Lower = 0.07, Upper = 0.25
CLASS-Bio B = 0.04, p = 0.010, SE = 0.015, Lower = 0.01, Upper = 0.07

Assumptions Testing for Integrate:

To test the assumption that the relationship between the IVs and DV has been correctly specified (*i.e.*, that a linear rather than non-linear model is appropriate), the following scatter plots may be evaluated:

- DV as a function of each IV
- Standardized residuals as a function of each IV
- Standardized residuals as a function of predicted Y values



If there is any relationship between predictor values and standardized residuals of the model, then it would indicate that the assumption has not been met. In this case, there is no evidence of a linear relationship between the standardized residuals and any of the three scaled predictors or predicted Y values (R^2 of the linear best fit lines < 0.01). The Loess best fit lines also show that there do not appear to be any non-linear relationships. Therefore,

without any type of relationship between these variables, the assumption holds that a linear rather than non-linear model is appropriate.

To test the assumption of constant variance of errors, or homoscedasticity, the same scatter plots may be evaluated:

- Standardized residuals as a function of each IV
- Standardized residuals as a function of predicted Y values

There are no relationships between standardized residuals and either the predictors or predicted values ($R^2 < 0.01$), which supports the assumption overall. However, all plots show slight wedge patterns, which indicates that the assumption of homoscedasticity may not have been met (heteroscedasticity may exist).

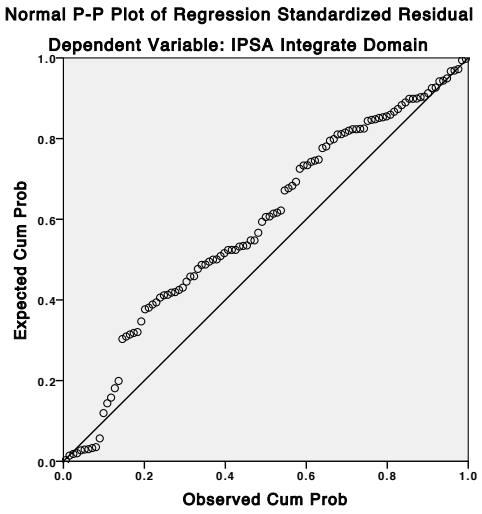
To evaluate the magnitude of nonconstant variance, residuals were divided at the median into two levels: low and high. Computing the ratio of the variances of the two groups determines the magnitude; more than a ten-fold difference calls for an alternate approach.

Descriptives

	Integrate (C) Standardized Residual (Binned)			Statistic	Std. Error
Integrate (C) Standardized Residual	<= .44525	Mean		-.3792484	.11304319
		95% Confidence Interval for Mean	Lower Bound	-.6060861	
			Upper Bound	-.1524106	
		5% Trimmed Mean		-.3214519	
		Median		-.1060665	
		Variance		.677	
		Std. Deviation		.82296687	
		Minimum		-2.60556	
		Maximum		.41660	
		Range		3.02216	
		Interquartile Range		1.13674	
		Skewness		-1.120	.327
		Kurtosis		.022	.644
	.44526+	Mean		1.1393581	.07792934
		95% Confidence Interval for Mean	Lower Bound	.9830516	
			Upper Bound	1.2956645	
		5% Trimmed Mean		1.1014336	
		Median		.9844933	
		Variance		.328	
		Std. Deviation		.57266137	
		Minimum		.44525	
		Maximum		2.59391	
		Range		2.14865	
		Interquartile Range		.80394	
		Skewness		.955	.325
		Kurtosis		.101	.639

With a ratio of 2.1 (0.677/0.328), a weighted least squares (WLS) regression is not necessary.

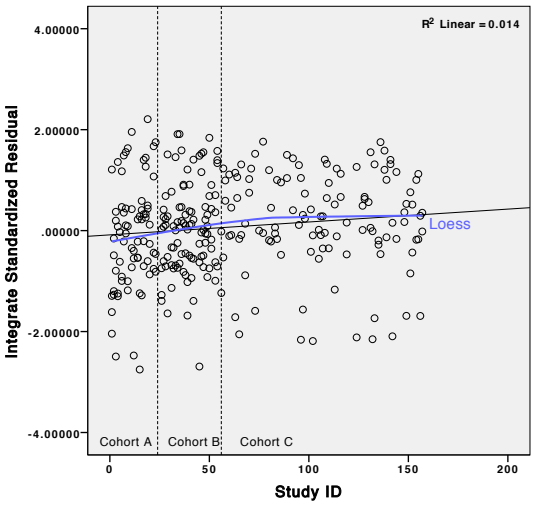
To test the assumption of normality of errors, a probability-probability (p-p) plot is evaluated.



Comparing the normal distribution (diagonal) with the residual distribution (circular markers), the p-p plot shows only slight departures from normality. Therefore, the assumption of normality of errors is met.

To test the assumption of independent errors, the following may be evaluated:

- Index plots of standardized residuals for each member, with the index ordered by any potential nesting factor
- Intraclass correlation coefficient (ICC), as a measure of the proportion of between-group variance to total variance



Considering that Cohorts A and B have four time points plotted per participant, while Cohort C participants only have two data points, the plot shows reasonable dispersion with only a minimal wedge. The assumption of independent errors is met.

SPSS Output for Reflect:

Model Summary[f]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.418[a]	.175	.167	1.741	.175	22.506	1	106	.000
2	.451[b]	.204	.156	1.753	.029	.723	5	101	.608
3	.458[c]	.210	.146	1.764	.006	.381	2	99	.684
4	.477[d]	.227	.148	1.762	.018	1.115	2	97	.332
5	.497[e]	.247	.134	1.776	.020	.608	4	93	.658

a. Predictors: (Constant), Time (semesters)

b. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

c. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

d. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

e. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender, Asian, American Indian, African American, Hispanic

f. Dependent Variable: IPSA Reflect Domain

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	68.238	1	68.238	22.506	.000[b]
	Residual	321.394	106	3.032		
	Total	389.632	107			
2	Regression	79.344	6	13.224	4.304	.001[c]
	Residual	310.288	101	3.072		
	Total	389.632	107			
3	Regression	81.715	8	10.214	3.284	.002[d]
	Residual	307.917	99	3.110		
	Total	389.632	107			
4	Regression	88.636	10	8.864	2.856	.004[e]
	Residual	300.996	97	3.103		
	Total	389.632	107			
5	Regression	96.304	14	6.879	2.181	.014[f]
	Residual	293.328	93	3.154		
	Total	389.632	107			

a. Dependent Variable: IPSA Reflect Domain

b. Predictors: (Constant), Time (semesters)

c. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

d. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

e. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

f. Predictors: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender, Asian, American Indian, African American, Hispanic

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	5.848	.260		22.481	.000	5.333	6.364
	Time (semesters)	.540	.114	.418	4.744	.000	.314	.765
2	(Constant)	4.231	2.220		1.906	.059	-.173	8.635
	Time (semesters)	.440	.137	.342	3.208	.002	.168	.713
	BIOC I Grade	1.684E-5	.029	.000	.001	1.000	-.057	.057
	BIOC II Grade	.009	.021	.052	.416	.678	-.033	.050
	Content Exam Score	-.011	.014	-.108	-.808	.421	-.039	.016
	CAT Score	.026	.035	.069	.734	.465	-.044	.095
	CLASS-Bio Overall	.018	.012	.153	1.580	.117	-.005	.041
3	(Constant)	3.979	2.274		1.750	.083	-.532	8.491
	Time (semesters)	.444	.139	.344	3.199	.002	.169	.719
	BIOC I Grade	.001	.029	.002	.022	.983	-.057	.058
	BIOC II Grade	.011	.021	.066	.522	.603	-.031	.053
	Content Exam Score	-.011	.014	-.101	-.749	.455	-.038	.017
	CAT Score	.025	.035	.068	.711	.479	-.045	.094
	CLASS-Bio Overall	.018	.012	.154	1.496	.138	-.006	.043
	Major - Non-Biochemistry	-.345	.785	-.041	-.439	.662	-1.902	1.213
	Major - Graduate	-.498	.642	-.072	-.775	.440	-1.772	.776
4	(Constant)	4.494	2.300		1.954	.054	-.072	9.059
	Time (semesters)	.460	.139	.357	3.307	.001	.184	.736
	BIOC I Grade	-.003	.029	-.011	-.096	.924	-.061	.055
	BIOC II Grade	.008	.022	.046	.361	.719	-.035	.051
	Content Exam Score	-.007	.014	-.066	-.483	.630	-.035	.021
	CAT Score	.022	.035	.059	.617	.539	-.048	.092
	CLASS-Bio Overall	.020	.012	.167	1.608	.111	-.005	.044
	Major - Non-Biochemistry	-.231	.789	-.028	-.292	.771	-1.797	1.336
	Major - Graduate	-.498	.652	-.072	-.763	.447	-1.792	.797
	Gender	-.333	.359	-.086	-.927	.356	-1.045	.380
	Age Group	-.535	.473	-.105	-1.132	.260	-1.473	.403
5	(Constant)	5.114	2.395		2.135	.035	.358	9.871
	Time (semesters)	.448	.141	.348	3.183	.002	.169	.728
	BIOC I Grade	-.010	.030	-.039	-.335	.738	-.070	.050
	BIOC II Grade	.013	.022	.079	.609	.544	-.030	.057
	Content Exam Score	-.010	.015	-.093	-.670	.505	-.039	.019
	CAT Score	.021	.037	.056	.564	.574	-.052	.093
	CLASS-Bio Overall	.020	.012	.166	1.584	.117	-.005	.045
	Major - Non-Biochemistry	-.326	.802	-.039	-.406	.685	-1.918	1.267
	Major - Graduate	-.297	.685	-.043	-.433	.666	-1.657	1.064
	Gender	-.351	.363	-.091	-.968	.336	-1.071	.369
	Age Group	-.577	.487	-.113	-1.185	.239	-1.544	.390
	Hispanic	-.549	.401	-.133	-1.370	.174	-1.345	.247
	Asian	.047	.690	.006	.068	.946	-1.322	1.416
	American Indian	-.351	1.318	-.025	-.266	.791	-2.969	2.267
	African American	-1.065	1.348	-.075	-.790	.432	-3.743	1.612

a. Dependent Variable: IPSA Reflect Domain

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	BIOC I Grade	.004[b]	.041	.967	.004	1.000
	BIOC II Grade	.012[b]	.132	.895	.013	1.000
	Content Exam Score	-.010[b]	-.105	.916	-.010	.866
	CAT Score	.089[b]	.987	.326	.096	.955
	CLASS-Bio Overall	.141[b]	1.590	.115	.153	.978
	Major - Non-Biochemistry	-.076[b]	-.862	.391	-.084	1.000
	Major - Graduate	-.039[b]	-.439	.662	-.043	1.000
	Gender	-.103[b]	-1.175	.243	-.114	1.000
	Age Group	-.090[b]	-1.026	.307	-.100	1.000
	Hispanic	-.090[b]	-1.020	.310	-.099	1.000
	Asian	.021[b]	.238	.812	.023	1.000
	American Indian	-.022[b]	-.248	.805	-.024	1.000
	African American	-.079[b]	-.895	.373	-.087	1.000
2	Major - Non-Biochemistry	-.038[c]	-.402	.688	-.040	.899
	Major - Graduate	-.070[c]	-.758	.450	-.076	.922
	Gender	-.082[c]	-.903	.369	-.090	.955
	Age Group	-.116[c]	-1.273	.206	-.126	.947
	Hispanic	-.105[c]	-1.159	.249	-.115	.958
	Asian	.039[c]	.432	.667	.043	.958
	American Indian	-.020[c]	-.221	.825	-.022	.968
	African American	-.080[c]	-.885	.378	-.088	.972
3	Gender	-.090[d]	-.972	.333	-.098	.925
	Age Group	-.108[d]	-1.172	.244	-.118	.934
	Hispanic	-.102[d]	-1.114	.268	-.112	.943
	Asian	.040[d]	.432	.666	.044	.952
	American Indian	-.027[d]	-.298	.766	-.030	.960
	African American	-.069[d]	-.736	.464	-.074	.919
4	Hispanic	-.123[e]	-1.327	.188	-.134	.923
	Asian	.038[e]	.412	.681	.042	.951
	American Indian	-.017[e]	-.187	.852	-.019	.940
	African American	-.058[e]	-.622	.536	-.063	.911

a. Dependent Variable: IPSA Reflect Domain

b. Predictors in the Model: (Constant), Time (semesters)

c. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score

d. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry

e. Predictors in the Model: (Constant), Time (semesters), BIOC II Grade, CLASS-Bio Overall, CAT Score, BIOC I Grade, Content Exam Score, Major - Graduate, Major - Non-Biochemistry, Age Group, Gender

Model fitting: Time & CLASS-Bio only

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	CLASS-Bio Overall, Time (semesters)[b]	.	Enter

a. Dependent Variable: IPSA Reflect Domain

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.441[a]	.195	.179	1.729	.195	12.679	2	105	.000

a. Predictors: (Constant), CLASS-Bio Overall, Time (semesters)

b. Dependent Variable: IPSA Reflect Domain

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	75.795	2	37.898	12.679	.000[b]
	Residual	313.837	105	2.989		
	Total	389.632	107			

a. Dependent Variable: IPSA Reflect Domain

b. Predictors: (Constant), CLASS-Bio Overall, Time (semesters)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.633	.807		5.741	.000	3.033	6.233
	Time (semesters)	.513	.114	.398	4.489	.000	.286	.739
	CLASS-Bio Overall	.017	.011	.141	1.590	.115	-.004	.038

a. Dependent Variable: IPSA Reflect Domain

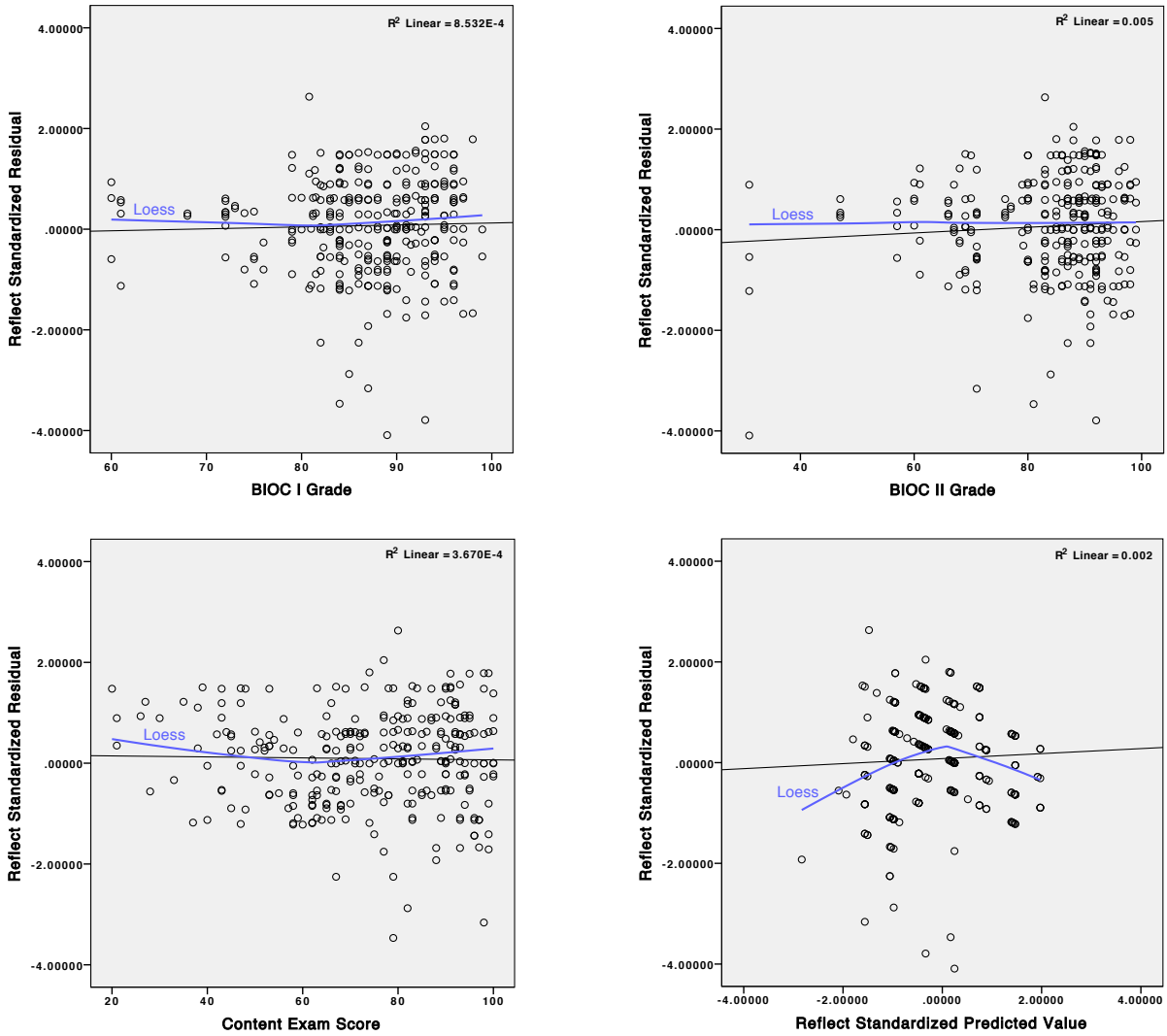
Reflect = 4.6 + (0.5)(Semester) + (0.02)(CLASS-Bio)

R2 = 0.20, Adj. R2 = 0.18, F = 12.679, $p < 0.001$
Intercept B = 4.6, $p < 0.001$, SE = 0.807, Lower = 3.0, Upper = 6.2
Time B = 0.5, $p < 0.001$, SE = 0.114, Lower = 0.3, Upper = 0.7
CLASS-Bio B = 0.02, p = 0.115, SE = 0.011, Lower = 0.00, Upper = 0.04

Assumptions Testing for Reflect:

To test the assumption that the relationship between the IVs and DV has been correctly specified (*i.e.*, that a linear rather than non-linear model is appropriate), the following scatter plots may be evaluated:

- DV as a function of each IV
- Standardized residuals as a function of each IV
- Standardized residuals as a function of predicted Y values



If there is any relationship between predictor values and standardized residuals of the model, then it would indicate that the assumption has not been met. In this case, there is no evidence of a linear relationship between the standardized residuals and any of the three scaled predictors or predicted Y values (R^2 of the linear best fit lines < 0.01). The Loess best fit lines of the IV plots also show that there do not appear to be any non-linear relationships. However, the Loess line of the predicted value plot indicates that a quadratic model may be a

better fit. Overall though, the assumption holds that a linear rather than non-linear model is appropriate.

To test the assumption of constant variance of errors, or homoscedasticity, the same scatter plots may be evaluated:

- Standardized residuals as a function of each IV
- Standardized residuals as a function of predicted Y values

There are no relationships between standardized residuals and either the predictors or predicted values ($R^2 < 0.01$), which supports the assumption overall. However, all plots show slight wedge patterns, which indicates that the assumption of homoscedasticity may not have been met (heteroscedasticity may exist).

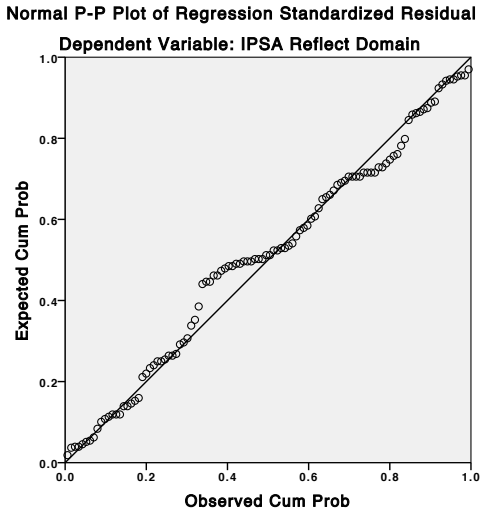
To evaluate the magnitude of nonconstant variance, residuals were divided at the median into two levels: low and high. Computing the ratio of the variances of the two groups determines the magnitude; more than a ten-fold difference calls for an alternate approach.

Descriptives

	Reflect (C) Standardized Residual (Binned)			Statistic	Std. Error
Reflect (C) Standardized Residual	<= .58216	Mean		-.2556345	.08401262
		95% Confidence Interval for Mean	Lower Bound	-.4241425	
			Upper Bound	-.0871265	
		5% Trimmed Mean		-.2171228	
		Median		-.1882649	
		Variance		.381	
		Std. Deviation		.61736414	
		Minimum		-2.22869	
		Maximum		.57994	
		Range		2.80864	
		Interquartile Range		.86440	
		Skewness		-.924	.325
		Kurtosis		.723	.639
	.58217+	Mean		1.4558611	.08280990
		95% Confidence Interval for Mean	Lower Bound	1.2897655	
			Upper Bound	1.6219568	
		5% Trimmed Mean		1.4383674	
		Median		1.3335663	
		Variance		.370	
		Std. Deviation		.60852602	
		Minimum		.58437	
		Maximum		2.74966	
		Range		2.16529	
		Interquartile Range		1.08099	
		Skewness		.385	.325
		Kurtosis		-1.063	.639

With a ratio of 1.0 (0.381/0.370), a weighted least squares (WLS) regression is not necessary.

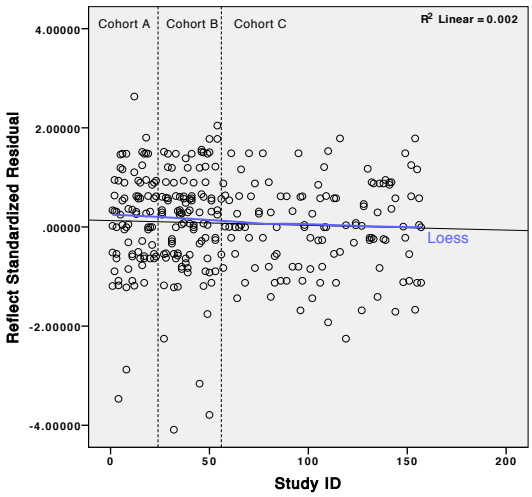
To test the assumption of normality of errors, a probability-probability (p-p) plot is evaluated.



Comparing the normal distribution (diagonal) with the residual distribution (circular markers), the p-p plot shows only slight departures from normality. Therefore, the assumption of normality of errors is met.

To test the assumption of independent errors, the following may be evaluated:

- Index plots of standardized residuals for each member, with the index ordered by any potential nesting factor
- Intraclass correlation coefficient (ICC), as a measure of the proportion of between-group variance to total variance



Considering that Cohorts A and B have four time points plotted per participant, while Cohort C participants only have two data points, the plot shows reasonable dispersion, with only a slight wedge pattern. The assumption of independent errors is met.

Appendix B
Supplementary Information

What really matters: Assessing individual problem-solving performance in the context of biological sciences

Steven M. Mitchell[1], William L. Anderson[2], Cheryl A. Sensibaugh[2], and Marcy Osgood[2]

[1] School of Medicine, University of New Mexico, Albuquerque, NM, USA

[2] Department of Biochemistry and Molecular Biology, University of New Mexico, Albuquerque, NM, USA

Computer screen captures of an Individual Problem-Solving Assessment (IPSA) which was used in 2008 with a class of 70 students in an advanced intermediary metabolism class.

This appendix presents a more detailed introduction to the computer-based Individual Problem-Solving Assessment (IPSA) and how we use a database for grading student responses. It should be stresses that computer administration of the exam is not necessary as we have also used these exams in a paper and pencil format.

This case, evaluating problems surrounding the catabolism of phenylalanine, is from an advanced intermediary metabolism course. Following the initial screens that require students to log into the testing system, students are presented with a short incomplete case scenario and then asked to list their hypotheses to explain the nature of the problem in the case scenario. This hypothesize screen is shown in Fig. B.1. Note that the initial case presentation is in a scrolling box to permit the possibility of using large or small case presentations.

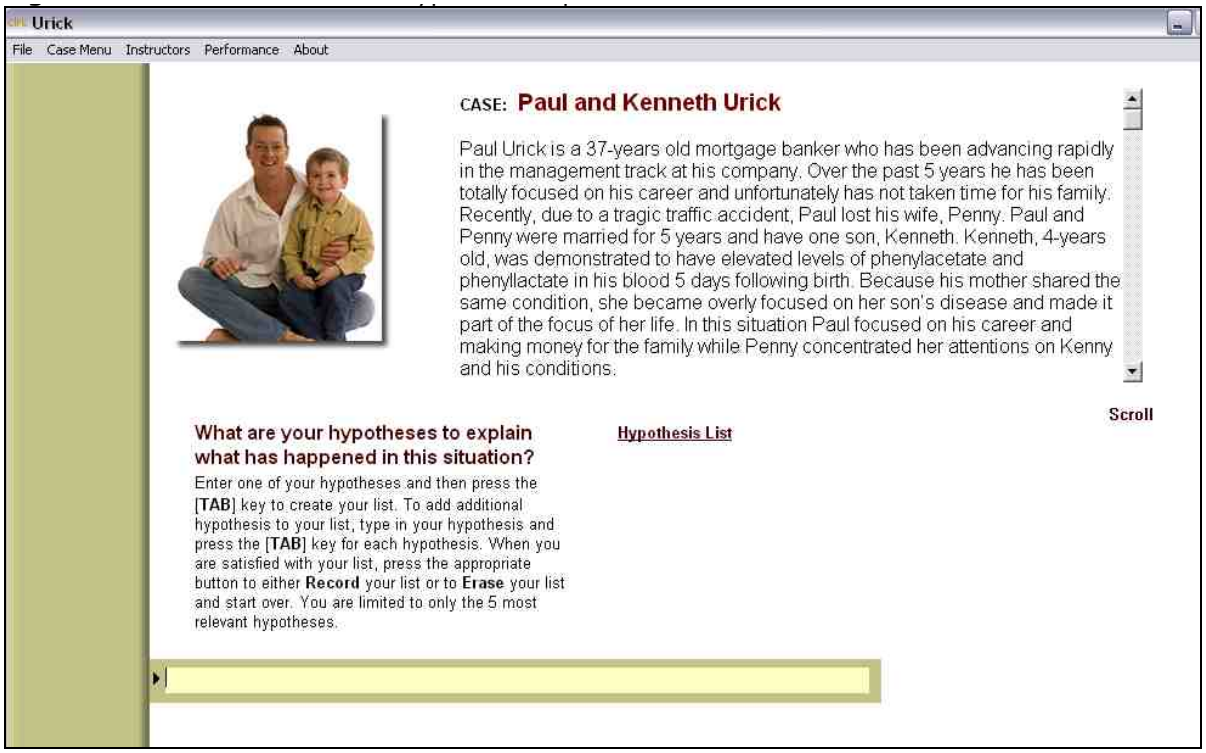


Figure B.1: Initial case scenario and hypothesize question.

Students are then given a more detail case history and are asked to begin investigating their leading hypothesis by identifying the key words they will use in their literature search. Once these key words are entered, the students are presented with the results of a literature search (Fig. B.2). The electronic case format allows students to be given learning materials during the test and prohibits them from going back and changing a previous answer.

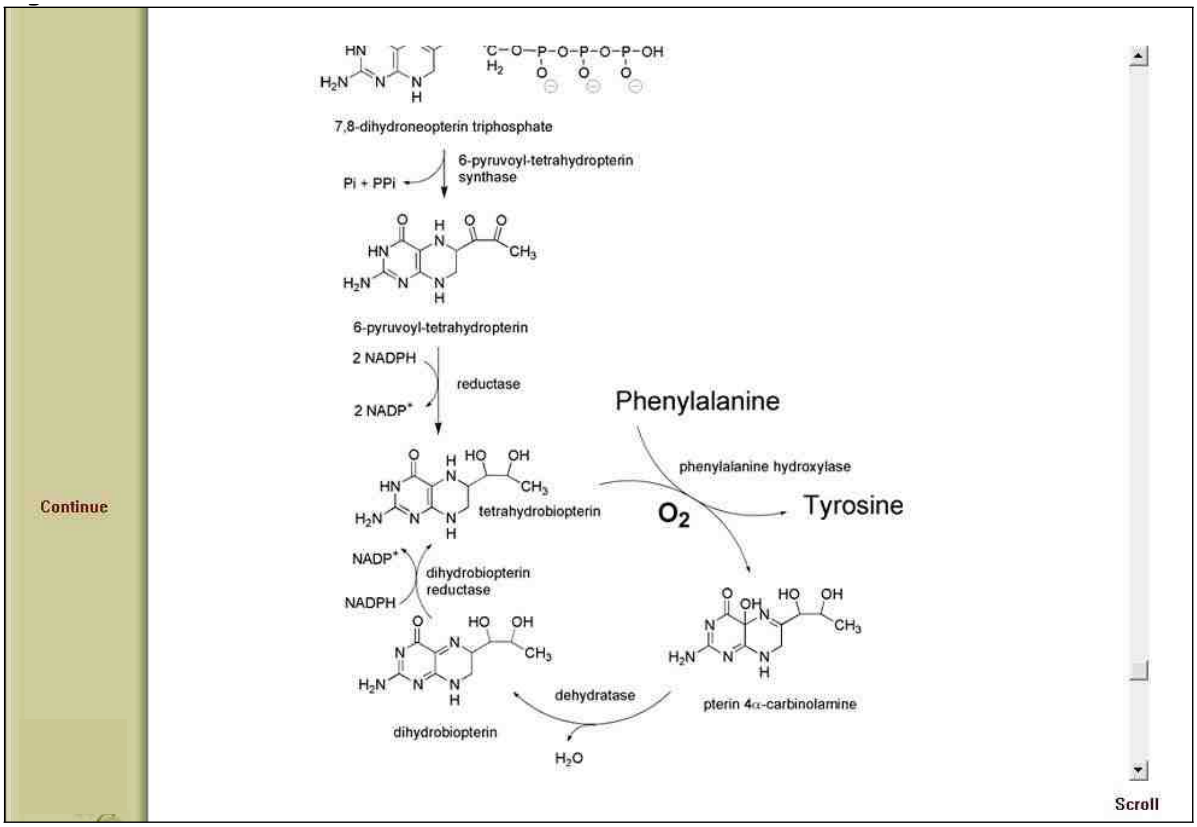


Figure B.2: Results of a literature search.

As the case progresses, students are sequentially asked to investigate a specific hypothesis by designing an experiment, to evaluate data that results from an experiment, and eventually to solve a dilemma related to the experimental data that requires the student to integrate the basic science knowledge about the topic in order to argue in support on one side of the dilemma or the other side. Figure B.3 illustrates how graphical data is presented to the student. It should be noted that in addition to tables and graphical data, this format is capable of presenting photographs, video or audio data for the student's analysis. For example medical school cases have used video tapes of simulated patient encounters and presents data in the form lung and cardiac sounds. It should be noted that there are problems with the experimental design described below and it will be the student's responsibility to point out the design flaws in the presented experiments.

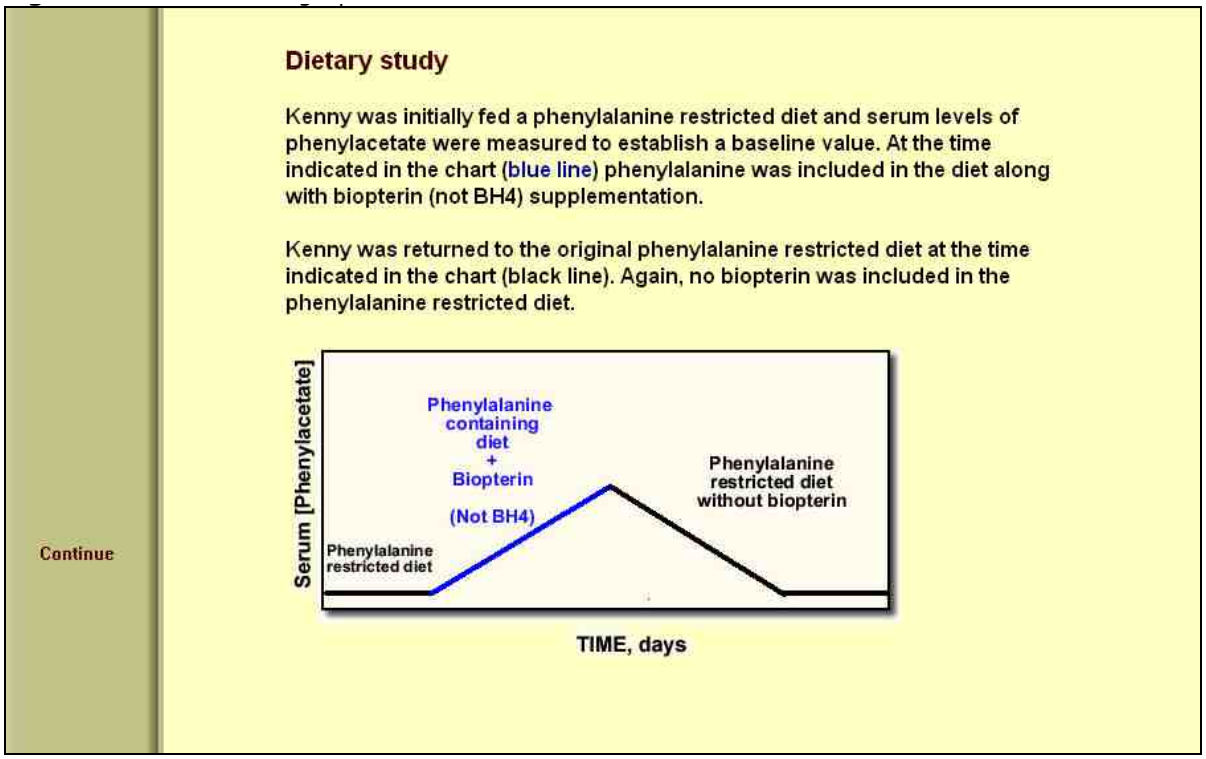


Figure B.3: Presentation of graphical data.

The student's responses to these questions are entered into textboxes, as illustrated in Fig. B.4. These text boxes can be set to limit the number of words available to the student. This has been found to be very effective in preventing students from writing everything they know about a topic in a "shotgun" type of answer and forces them to focus on answering a specific question.

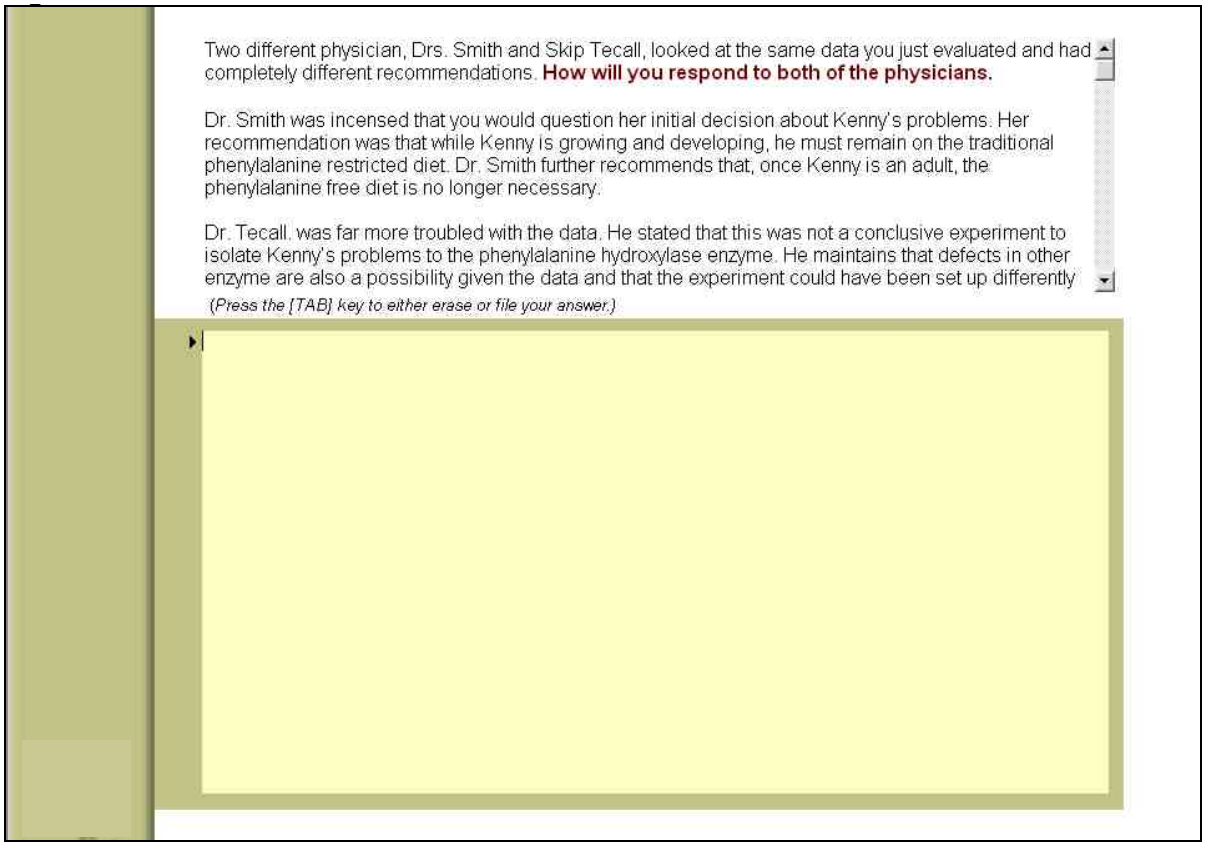


Figure B.4: Student answers entered into a text box.

Once the students have completed the examination, their responses are automatically saved to a database for grading. Figure B.5 shows an example of the database screen for grading the **Integrate** question. At the right of the screen, the grading rubrics are provided for the faculty-grader. It should be noted that there are two different approaches to grading. One approach is to set the database tab on one domain and then grade the entire class on that domain. The second approach is to select one student and sequentially follows a single student's responses through all five problem-solving domains. The first approach appears results in the most consistent grading while the second approach is preferable for grading the Reflect domain.

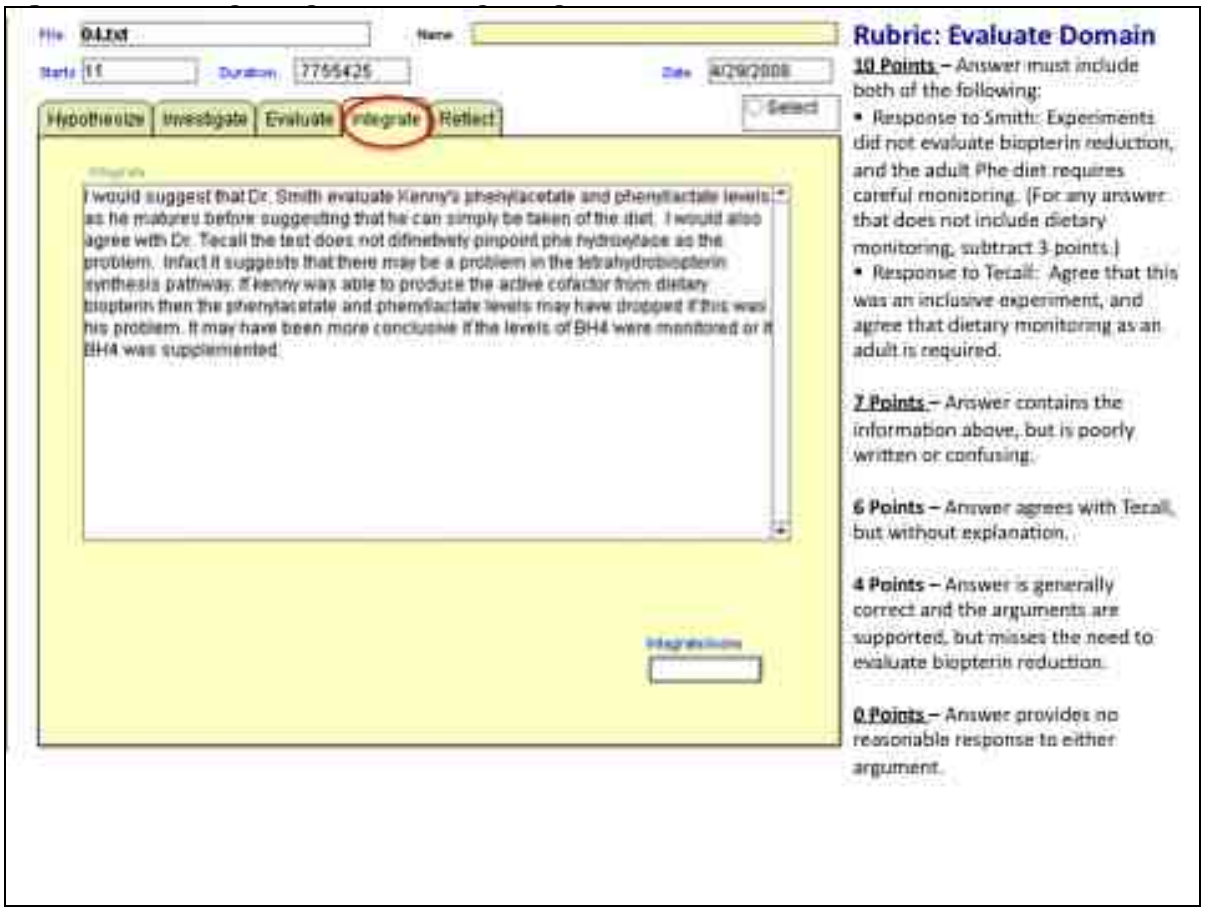


Figure B.5: Database grading screen with grading rubrics.

Appendix C
Supplementary Information

Scientific problem solving within an undergraduate biochemistry and molecular biology curriculum

Cheryl A. Sensibaugh, William L. Anderson[1], Marcy Osgood

Department of Biochemistry and Molecular Biology, University of New Mexico, Albuquerque, NM, USA

[1] Professor emeritus

Appendix C.I

Critical thinking Assessment Test (CAT)

The CAT measures four critical thinking skills, defined by developers. This is a secure instrument; therefore, it is not made publicly available. It is administered in hard copy, with testing time limited to 60 minutes. There are fifteen questions total, some with multiple parts.

Critical Thinking Skills

1. Evaluating Information
   - Separate factual information from inferences.
   - Interpret numerical relationships in graphs.
   - Understand the limitations of correlational data.
   - Evaluate evidence and identify inappropriate conclusions.

2. Creative Thinking
   - Identify alternative interpretations for data or observations.
   - Identify new information that might support or contradict a hypothesis.
   - Explain how new information can change a problem.

3. Learning and Problem Solving
   - Separate relevant from irrelevant information.
   - Integrate information to solve problems.
   - Learn and apply new information.
   - Use mathematical skills to solve real-world problems.

4. Communication
   - Communicate ideas effectively

Sample Question

A scientist working at a government agency believes that an ingredient commonly used in bread causes criminal behavior. To support his theory, the scientist notes the following evidence.

- 99.9% of the people who committed crimes consumed bread prior to committing crimes.
- Crime rates are extremely low in areas where bread is not consumed.

Do the data presented by the scientist strongly support their theory? Yes ____ No ____

Are there other explanations for the data besides the scientist's theory? If so, describe.

_____

What kind of additional information or evidence would support the scientist's theory?

_____

Appendix C.II

Colorado Learning Attitudes about Science Survey for Biology (CLASS-Bio)

The CLASS-Bio measures eight factors, determined by statistical factor analysis, that contribute to perceptions about learning biology. Students select their degree of agreement with 31 statements on a Likert scale with five levels: strongly agree, agree, neutral, disagree, and strongly disagree. Testing time is typically 10 minutes.

Table C.II.1. CLASS-Bio Statements by Category

	Statement	Real World Connection	Enjoyment (Personal Interest)	Conceptual Connections/Memorization	Problem Solving: Reasoning	Problem Solving: Synthesis & Application	Problem Solving: Strategies	Problem Solving: Effort	Uncategorized
1	My curiosity about the living world led me to study biology.		✔						
2	I think about the biology I experience in everyday life.	✔	✔						
3	After I study a topic in biology and feel that I understand it, I have difficulty applying that information to answer questions on the same topic.					✔			
4	Knowledge in biology consists of many disconnected topics.								✔
5	When I am answering a biology question, I find it difficult to put what I know into my own words.					✔			
6	I do not expect the rules of biological principles to help my understanding of the ideas.			✔		✔			
7	To understand biology, I sometimes think about my personal experiences and relate them to the topic being analyzed.						✔		
8	If I get stuck on answering a biology question on my first try, I usually try to figure out a different way that works.			✔	✔		✔	✔	
9	I want to study biology because I want to make a contribution to society.		✔						
10	If I don't remember a particular approach needed for a question on an exam, there's nothing much I can do (legally!) to come up with it.					✔			
11	If I want to apply a method or idea used for understanding one biological problem to another problem, the problems must involve very similar situations.			✔		✔			
12	I enjoy figuring out answers to biology questions.	✔	✔					✔	
13	It is important for the government to approve new scientific ideas before they can be widely accepted.								✔

Table C.II.1. CLASS-Bio Statements by Category

	Statement	Real World Connection	Enjoyment (Personal Interest)	Conceptual Connections/Memorization	Problem Solving: Reasoning	Problem Solving: Synthesis & Application	Problem Solving: Strategies	Problem Solving: Effort	Uncategorized
14	Learning biology changes my ideas about how the natural world works.	✔			✔				
15	To learn biology, I only need to memorize facts and definitions.			✔					
16	Reasoning skills used to understand biology can be helpful to my everyday life.	✔			✔				
17	It is a valuable use of my time to study the fundamental experiments behind biological ideas.	✔			✔				
18	If I had plenty of time, I would take a biology class outside of my major requirements just for fun.		✔						
19	The subject of biology has little relation to what I experience in the real world.	✔		✔					
20	There are times I think about or solve a biology question in more than one way to help my understanding.						✔	✔	
21	If I get stuck on a biology question, there is no chance I'll figure it out on my own.					✔			
22	When studying biology, I relate the important information to what I already know rather than just memorizing it the way it is presented.						✔	✔	
23	There is usually only one correct approach to solving a biology problem.			✔					
24	When I am not pressed for time, I will continue to work on a biology problem until I understand why something works the way it does.				✔			✔	
25	Learning biology that is not directly relevant to or applicable to human health is not worth my time.	✔							
26	Mathematical skills are important for understanding biology.								✔
27	I enjoy explaining biological ideas that I learn about to my friends.		✔					✔	
28	We use this statement to discard the surveys of people who are not reading the statements, so select agree only, not strongly agree, for this statement.	N/A							
29	The general public misunderstands many biological ideas.								✔
30	I do not spend more than a few minutes stuck on a biology question before giving up or seeking help from someone else.					✔		✔	
31	Biological principles are just to be memorized.			✔					
32	For me, biology is primarily about learning known facts as opposed to investigating the unknown.			✔					

Appendix C.III

IPSA Scoring Rubrics and Inter-Rater Reliability

Hypothesize Domain at 0 Semesters (Protein purification)
What are your top five hypotheses for the function of the protein?

10: Three hypotheses with rationales
9: Three hypotheses
8: Two hypotheses with rationales
7: Two hypotheses about the following functions:
    muscle contraction or repair
    catabolism, oxidation, or energy production
    protein synthesis or translation of mRNA
    oxygen binding or transport
6: One hypothesis with rationale
5: One hypothesis or all are part of the same function
4: Unacceptable hypotheses:
    DNA replication or transcription
    ion binding or transport (calcium, iron, etc.)
    same function as the rRNA to which it binds
    signaling
    localization of the protein
3: Pattern-matching (the protein is hemoglobin)
2: Restating the problem:
    structural properties (soluble, polar, not membrane-bound)
    interacts with ribosomes
    reacts with oxygen
1: Off-topic
0: No response

Hypothesize Domain at 1 Semester (Carbohydrate metabolism)

What are your top four hypotheses to explain this situation?

10: Three hypotheses with rationales
9: Three hypotheses
8: Two hypotheses with rationales
7: Two hypotheses about the following:
    Genetics (metabolism, energy, diabetes type I)
    Diet (nutrition, co-factors, vitamins, etc.)
    Signaling (hormones, neurotransmitters, diabetes type II)
    Oxygen transport or delivery (RBCs, hemoglobin)
    Environment (infection or toxin)
    Trauma
    Cancer
    Autoimmune
    Psychiatric disorder (bi-polar, schizophrenia)
6: One hypothesis with rationale
5: One hypothesis or all are part of the same function
4: Unacceptable hypotheses:
    abuse, neglect, sleep/social/learning disorder
3: Pattern-matching
2: Restating the problem
1: Off-topic
0: No response

Hypothesize Domain at 2 Semesters (Amino acid metabolism)

List your top four mechanistic hypotheses that Paul needs to consider as an underlying cause for Kenny's elevated metabolites.

10: Three hypotheses with rationales
9: Three hypotheses
8: Two hypotheses with rationales
7: Two hypotheses about the following:
    (need to be specific about areas of metabolism; given a lot)
    Phe processing deficiency: phenylalanine hydroxlase/mfo
    Cofactor deficiency: BH4 synthesis or oxygen for mfo
    environment (infection/toxin/medication)
    cancer
    autoimmune
6: One hypothesis with rationale
5: One hypothesis
4: Unacceptable hypotheses:
    transaminase deficiency (2014-01:ok if upregulated)
    focus on nitrogen disposal
3: Pattern-matching, OR no specific hypothesis
    enzyme/co-factor deficiency
    genetics
    diet
2: Restating the problem (Kenny has PKU)
1: Off-topic
0: No response

Hypothesize Domain at 4 Semesters (Carbohydrate and Lipid Metabolism)

What are your top four biochemical hypotheses to explain the Lorrat's unique physiological abilities?

10: Three hypotheses with rationales
9: Three hypotheses
8: Two hypotheses with rationales
7: Two hypotheses about the following:
    Altered genetics/genetic processing of metabolic proteins
    Altered regulation of body temperature
    Altered cellular structure (more mitochondria)
    Oxygen transport/delivery (lung capacity, Hb, Mb, BPG)
    Nutritional deficiency
    Environment (infection or toxin)
    Trauma
    Cancer
    Autoimmune
6: One hypothesis with rationale
5: One hypothesis
4: Unacceptable hypotheses:
    Teleological conceptions (outcomes)
    Increased/fast metabolism
    Increased energy needed for proliferation
3: Pattern-matching
2: Restating the case/problem (something functions differently)
1: Off-topic
0: No response

Investigate Domain at 0 Semesters (Protein Purification)

Based upon the information you currently possess about the protein, outline your proposed first TWO steps to achieve the best separation of the crude muscle homogenate.

Not scored; prompt is not an experimental design task.

Investigate Domain at 1 Semester (Carbohydrate Metabolism)

Your mentor wants you to evaluate Brian's fructose-bisphosphate phosphatase enzyme for a possible enzyme defect, without purifying and characterizing the enzyme, which may take years. Your task is to design an experimental approach to elucidate the molecular basis for a putative enzyme defect in fructose-bisphosphate phosphatase. In your experiment you need to clearly identify both your dependent and independent variables.

10: As for 7, with three of the below
9: As for 7, with two of the below
8: As for 7, with one of the below:
    rationale = to determine enzyme functionality
    expected results
    interpretation of expected results
7: Kinetics analysis with:
    positive control (reference sample)
    IV = Δ[S]
    DV = rate of product formed
6: One of the above missing/incorrect
5: Two of the above missing/incorrect
4: Three of the above missing/incorrect
3: Proposal doesn't test enzyme function (metabolite levels and in vivo approaches)
2: Proposal will not yield useful information in this case
1: Off-topic
0: No response

Investigate Domain at 2 Semesters (Amino Acid Metabolism)

What experiment do you recommend that Kenny's physician carry out, so that the physician can help Paul understand how to best care for Kenny?

10: As for 7, with three of the below
9: As for 7, with two of the below
8: As for 7, with one of the below:
    rationale = to distinguish between putative pathways
    expected results
    interpretation of expected results
7: EITHER a Dietary study:
    Neg. Control = low phe (without BH4)
    Pos. Control = increased phe (without BH4)
    IV = supplemental BH4
    DV = serum levels of phe metabolites
  OR Kinetics analyses:
    Run for phe hydroxylase or BH2 reductase
    Positive Control = certified reference enzymes
    IV = Δ[S]
    DV = rate of products formed
6: One parameter missing/incorrect
5: Two parameters missing/incorrect
4: Three parameters missing/incorrect
3: Four parameters missing/incorrect
2: Proposal will not yield useful information in this case OR measure [metabolite]
1: Off-topic; no experiment proposed
0: No response

Investigate Domain at 4 Semesters (Carbohydrate and Lipid Metabolism)

Briefly describe your proposed experimental design, with appropriate controls, to test this hypothesis. DO NOT simply name a technique, but rather explain the reasoning for your design and how the methods will provide supportive evidence.

10: As for 7, with three of the below
9: As for 7, with two of the below
8: As for 7, with one of the below:
    rationale = to detect transcription
    expected results
    interpretation of expected results
7: All four of the following:
    Method
        Quantitative RT-PCR, luciferase/beta-gal reporter assay,
        electrophoresis/northern blot, hybridization techniques
        (cDNA microarray measuring hybridization of mRNA is theoretically logical)
    (Negative) Control – small mammal reference sample
    IV - differences in transcription
    DV – PEPCK mRNA in muscle tissue
6: One of the above missing/incorrect
5: Two of the above missing/incorrect
4: Three of the above missing/incorrect
3: Four of the above incorrect
2: Proposal is not aligned with hypothesis (kinetics)
1: Off-topic
0: No response

Evaluate Domain at 0 Semesters (Protein Purification)

In the space below, write a brief report to your advisor in which you critically evaluate this purification protocol. You must support your recommendations with data.

10: As for 9, with explanation of reasoning behind SA (normalization)
9: Recommend eliminating two steps based on SA
8: Recommend eliminating two steps based on SA, but wrong math
7: Recommend eliminating one step based on SA:

Procedure	Specific Activity (SA; units•mg^{-1}•ml^{-1})
Initial homogenate	200
Differential centrifugation	600
**Salt precipitation*	*250*
Ion exchange chromatography	4,000
Size exclusion chromatography	15,000
**Affinity chromatography*	*14,444*

6: Report SA but do not make recommendations
OR recommend eliminating one step based on SA, but wrong math
5: Make recommendations without SA (raw data only)
4: Claim that SA increased without quantifying
3: Need an improved protocol or another purification step
OR no consideration of SA, re-stating raw data
2: Explain methods
1: Off-topic
0: No response

Evaluate Domain at 1 Semester (Carbohydrate Metabolism)

What conclusions do you draw from these data? Your answer must describe, using proper biochemistry terminology, the characteristics of Brian's pyruvate carboxylase enzyme in as much detail as is justified by the data.

10: As in 7, with three of the below
9: As in 7, with two of the below
8: As in 7, with one of the below:
    Explain there is no evidence to claim inhibition
    Describe inhibition analysis (add putative inhibitor)
    Vary inhibitor concentration
7: BOTH of the following:
    1-Increased Km (lower affinity/need more S)
    2-Same Vmax
    ok - "consistent with" competitive inhibition*
6: One of the above missing/incorrect
5: Both of the above missing/incorrect
4: *Claim that a competitive inhibitor is present (even if both parameters correct)
3:
2: Explain methods
1: Off-topic
0: No response

Evaluate Domain at 2 Semesters (Amino Acid Metabolism)

What is the result of the dietary study? In other words, what evidence have you acquired?

10: As for 9, and allude to or describe proper investigation
9: As for 7, with explanation that mechanism has not been attributed to either BH4 synthesis or phe hydroxylase
8: As for 7, and pinpoint that BH2 reductase not tested
7: BOTH of the following:
    1-Flawed experiment
    2-Evidence only that serum phenylacetate increased with increased phe intake, which was already known.
6: Only describe flawed experiment (but correctly)
5: Only describe evidence (but correctly)
4: Describe both incorrectly
3: Describe one incorrectly
2: Summarize and assume methods suffice as given
1: Off-topic
0: No response

Evaluate Domain at 4 Semesters (Carbohydrate and Lipid Metabolism)

How do the parameters of interest compare across the two species?

10: As in 7, with specific activity AND value of aldolase control
9: As in 7, with specific activity OR value of aldolase
8: As in 7, with specific activity (same)
OR value of aldolase (validity of result), but incorrect or vague
7: All of the following:
    For Lorrat compared to control
    Increased [PEPCK]
    Increased PEPCK activity
    Equivalent Km
6: One of the above missing/incorrect
5: Two of the above missing/incorrect
4: Three of the above missing/incorrect
3: Four of the above incorrect
2: Explain methods
1: Off-topic
0: No response

Integrate Domain at 0 Semesters (Protein Purification)

How can you explain these molecular weight data, going on the assumption that you do actually have a pure protein after the size exclusion chromatography step?

10: Chromatography shows dimer (2x38) and SDS shows one subunit (38) comprised of two peptides joined by disulfide bonds (15+23)
9: Chromatography shows dimer and SDS shows one subunit comprised of two peptides joined by disulfide bonds
8: As for 7, with calculations (15+23=38, 38x2 = 76)
7: Protein has two subunits, and subunits each contain two peptides
6: The protein is comprised of multiple peptides
5: Different weights are due to different protein conformations
OR protein with at least 4 disulfide bonds
4: Disulfide bonds get trapped in the gel
OR sds (a detergent) broke disulfide bonds
3: The weight of the SDS adds to the protein weight
2: Performic acid hydrolyzed peptide bonds or disulfide bonds within a single peptide
OR single peptide split by diff techniques
OR single peptide + other molecules
1: Off-topic, or the assumption of purity is invalid (multiple impure proteins)
0: No response

Integrate Domain at 1 Semester (Carbohydrate Metabolism)

Even given the normal kinetics, your mentor is unwilling to give up on the defective enzyme hypothesis. He claims that it is still a possibility that one of the enzymes of gluconeogenesis could be defective leading to Brian's condition. With supporting arguments, explain why you either agree or disagree.

10: Agree and provide four alternatives
9: Agree and provide three alternatives
8: Agree and provide two alternatives
7: Agree and provide one possible alternative:
    gene expression
    protein synthesis
    protein stability
    protein targeting
6: Agree but no other mechanistic possibilities specified
5: Agree but incorrect reasoning:
    regulatory mechanisms
    co-factor deficiencies
    inhibitor
    disregard assumption
4: Disagree and/or argue that nothing else could cause the gluconeogenic enzymes to be defective
3: Disagree and/or discuss alternatives outside of the gluconeogenic enzymes
2: Disagree
1: Off-topic
0: No response

Integrate Domain at 2 Semesters (Amino Acid Metabolism)

(1) What additional evidence have you acquired from the kinetic analysis?

(2) Based on all of the information available in this case, what is your recommendation to Paul for Kenny's future diet and care?

10: BOTH of the following:
(1) Evidence that BH2 reductase is dysfunctional, and cite increased Km/lower affinity for BH2 substrate
(2) Treat per Dr. Tecall

9: Both addressed but Vmax is lower

8: Both addressed but without mention of higher Km

7: Both addressed - one incorrectly

6: Both addressed - both incorrectly

5: Only one addressed - correctly

4: Only one addressed – incorrectly

3:

2: Main interpretation is that Kenny has PKU; not focused on determining mechanism (even if correct treatment)

1: Off-topic

0: No response

Integrate Domain at 4 Semesters (Carbohydrate and Lipid Metabolism)

How do you interpret the data collected throughout this investigation, to explain any role that PEPCK might have in the lorrat achieving its unique physiological abilities? Be sure to address the results of the previous protein assays, these new metabolic assays, and any other relevant information.

10: All eight of the following:
Increased muscle PEPCK expression/concentration
→ high PEPCK activity
→ high DHAP (some reference to glycerol)
→ high TAGs
→ aerobic catabolism
→ (a) low lactate
→ (b) high ATP yield
→ unique abilities

9: Seven of the above
8: Six of the above
7: Five of the above
6: Four of the above
5: Three of the above
4: Two of the above
3: One of the above
2: Incorrect interpretation
1: Off-topic
0: No response

Reflect Domain at 0 Semesters (Protein Purification)

Cohorts A & B

Part 1: What functional aspects of the protein led your advisor to this hypothesis?
Part 2: And, what further experiments will you carry out to test that hypothesis?

Sum of Parts 1 & 2

Part 1:

5: Interacts with negatively-charged nucleic acid
4: Nucleic acid AND reacts with oxygen
3:
2: Unacceptable aspects:
hydrophilic, soluble, found in muscle, results from Integrate, ion exchange chrom., reacts with oxygen
1: Off-topic; no functional aspects
0: No response

Part 2:

5: Two methods with rationale
4: Two methods, or one with rationale
3: One method:
ion exchange, sequencing, or isoelectric focusing
2: Irrelevant methods
1: Off-topic; no method
0: No response

Cohort C

Part 1: What functional aspects of the protein led your advisor to this hypothesis?
Part 2: And, what further experiments will you carry out to test that hypothesis?
Part 3: Critically evaluate your performance on this Individual Problem Solving Assessment.

Sum of Parts 1, 2, & 3

Part 1:

3: Interacts with negatively-charged nucleic acid
2: Nucleic acid AND reacts with oxygen
1: Unacceptable aspects:
hydrophilic, soluble, found in muscle, results from Integrate, reacts with oxygen
0: No response or off-topic

Part 2:

3: Two methods, or one with rationale
2: One method:
ion exchange, sequencing, or IEF
1: Irrelevant or no method
0: No response

Part 3:

4: Cognitive dissonance resolved
(Thought x, but now y.)
3: Plans for next time
(preparation, not rush, fully explain)
2: Self-assessment (accurate or not)
1: Discuss IPSA structure, problem solving, scientific method
0: No response, questions weren't specific, IPSAs are invalid, etc.

Reflect Domain at 1 Semester (Carbohydrate Metabolism)

Since you now know the metabolic basis of Brian's problem, how do you evaluate your performance on this exam? Have you identified any areas of your current knowledge that need refreshing?

10: Four aspects and assessment is accurate
9: Four aspects
8: Three aspects
7: Two of the following aspects:
    self-assessment
    area of good performance
    area of improvement
    method for improving
6: One of the above
5:
4: Unsupported self-assessment (Only, "I did okay.")
3:
2:
1: Off-topic
0: No response

Reflect Domain at 2 Semesters (Amino Acid Metabolism)

Critically evaluate your performance on this IPSA.

10: Four aspects and assessment is accurate
9: Four aspects
8: Three aspects
7: Two of the following aspects:
    self-assessment
    area of good performance
    area of improvement
    method for improving
6: One of the above
5:
4: Unsupported self-assessment (Only, "I did okay.")
3:
2:
1: Off-topic
0: No response

Reflect Domain at 4 Semesters (Carbohydrate and Lipid Metabolism)

1) Were you able to meet each of the tasks required in this case?

2) What aspects of your undergraduate education helped you the most for solving this case?

3) In one sentence or less, describe any personal relevance of working through this case study.

10: As for 7, with three of the below

9: As for 7, with two of the below

8: As for 7, with one of the below:

Self-assessment is accurate

Describe method for improvement

Helped learn process not just facts

7: Addressed all three of the following:

Self-assessment (do not accept "I hope so.")

Most helpful program aspect

Personal relevance is helped learn content, saw improvement over time, need for future profession, etc. (it counts as long as it's addressed)

6: Two of the above

5: One of the above

4:

3:

2:

1: Off-topic

0: No response

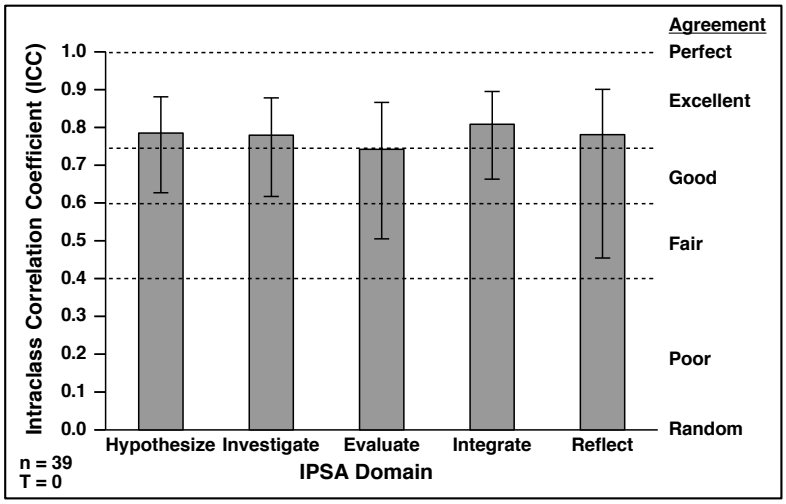


Figure C.III.1: Inter-Rater Reliability. The intraclass correlation coefficient (ICC) estimated inter-rater reliability for IPSA domain scores generated by two raters. Error bars indicate 95% confidence intervals. Cut-off values for interpreting the ICC with regard to rater agreement are depicted by dashed lines (Cicchetti, 1994; Hallgren, 2012).

Appendix C.IV
Preliminary Study on IPSA Performance

Methods

A small subset of biochemistry majors (N = 11) was randomly selected from Cohort A. Means and standard deviations were generated for the first and last IPSAs taken during the second semester of the junior year, at T=1 and T=2, respectively. Achievement rates were also calculated, along with rates for corresponding content exams for comparison.

Results

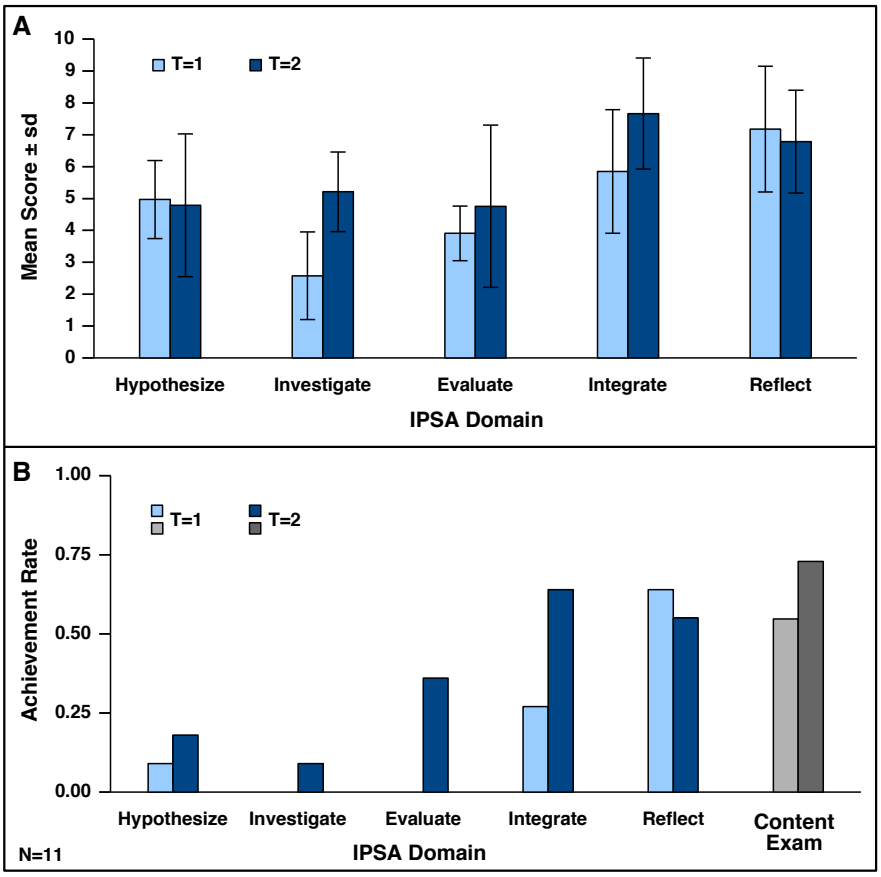


Figure C.IV.1: Preliminary Description of Problem Solving Performance. A pilot sample informed hypotheses of **(A)** the average student and **(B)** most students. Error bars indicate s.d.

Appendix C.V

Academic and Demographic Student Backgrounds

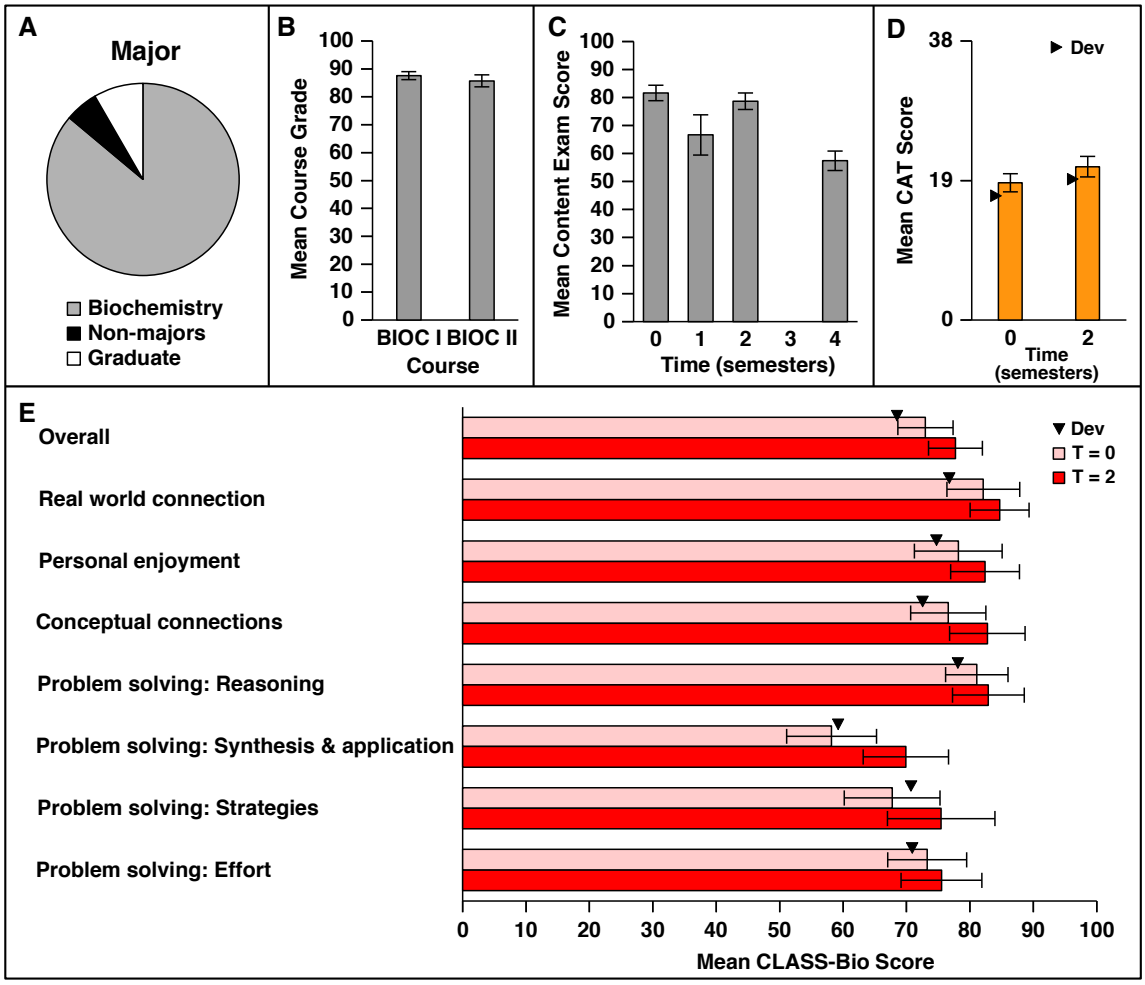


Figure C.V.1: Academic Backgrounds. The following aspects were evaluated as academic factors that may impact performance in problem solving: **(A)** academic major, **(B)** biochemistry course grades, **(C)** scores on biochemistry content exams, **(D)** scores on the critical thinking instrument, and **(E)** scores on the learning attitudes instrument. In addition to overall attitude scores, those on subsets of the assessment are also included. **(D-E)** The charts include mean scores reported during development (Dev) of the CAT (Stein *et al.*, 2010) and CLASS-Bio (Semsar *et al.*, 2011). **(B-E)** Error bars indicate 95% confidence intervals.

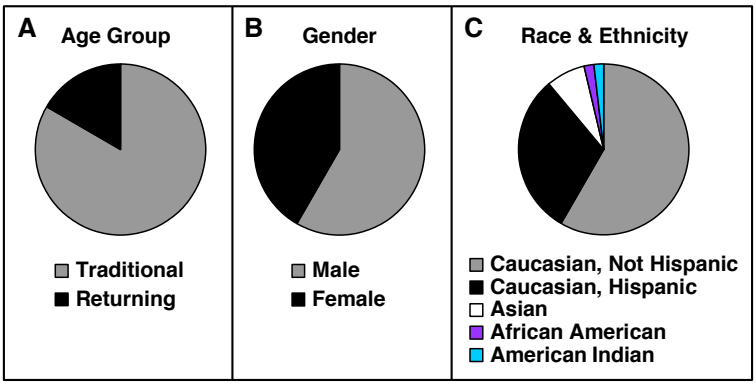


Figure C.V.2: Demographic Backgrounds. The following aspects were evaluated as demographic factors that may impact performance in problem solving: **(A)** age, **(B)** gender, and **(C)** race and ethnicity.

Appendix C.VI
Score Distributions

Methods

Descriptive statistics – To summarize the spread of scores across pooled students, distributions were represented as boxplots, with the following reference points: minimum score; the 25th, 50th, and 75th percentiles; and maximum score. Measures included course grades, content exam scores, CAT scores, CLASS-Bio scores, and IPSA domain scores.

Results

Distributions of course grades and content exam scores were as expected, based on prior experience (Fig. C.VI.1A-B). CAT score distributions shifted upward during the first year of the biochemistry program (Fig. C.VI.1C). Distributions of CLASS-Bio scores also shifted upward across time, for most sub-scores as well as for the overall score (Fig. C.VI.1D). These outcomes were consistent with our emphasis on problem-based learning.

IPSA domain score distributions were inconsistent across both time and domains (Fig. C.VI.2). In many cases, the spread spanned most of the scale. These distributions will also be informative during future efforts to standardize the IPSA prompts and rubrics.

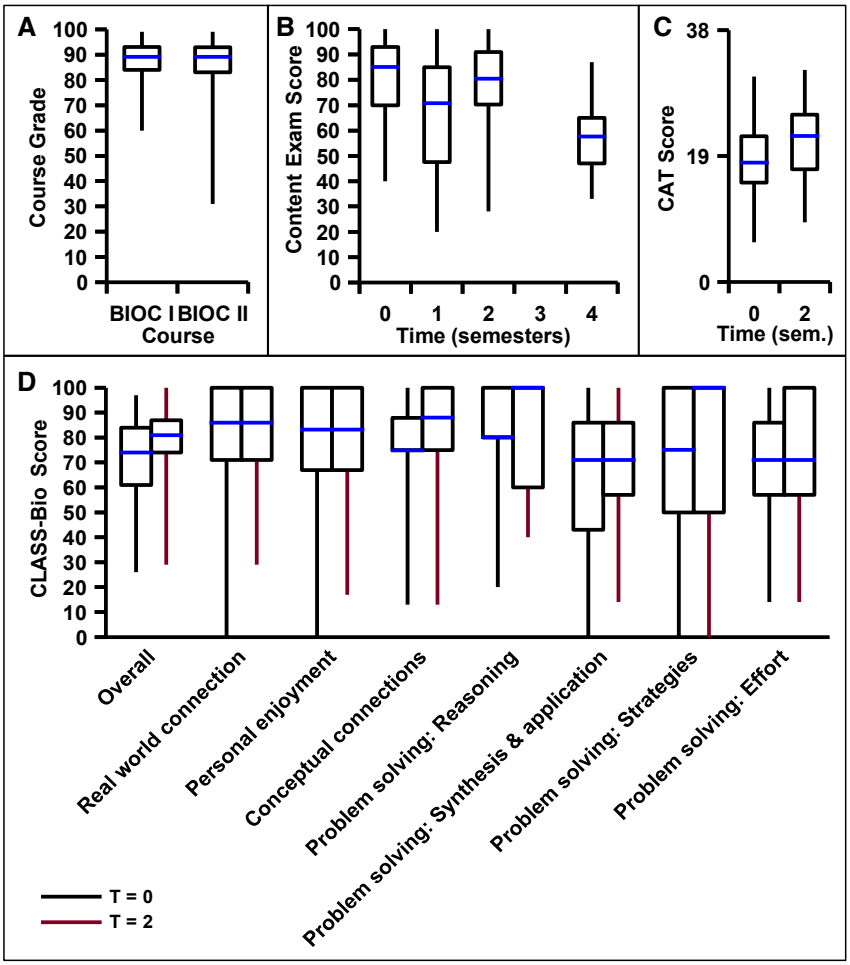


Figure C.VI.1: Academic Score Distributions. Boxplots summarize the score distributions of **(A)** course grades, **(B)** content exams, **(C)** CAT, and **(D)** CLASS-Bio. The bottom and top of the whiskers indicate minimum and maximum scores, respectively. The 25th percentile is at the bottom of the box, the 50th percentile is in or on the box (blue line), and the 75th percentile is at the top of the box.

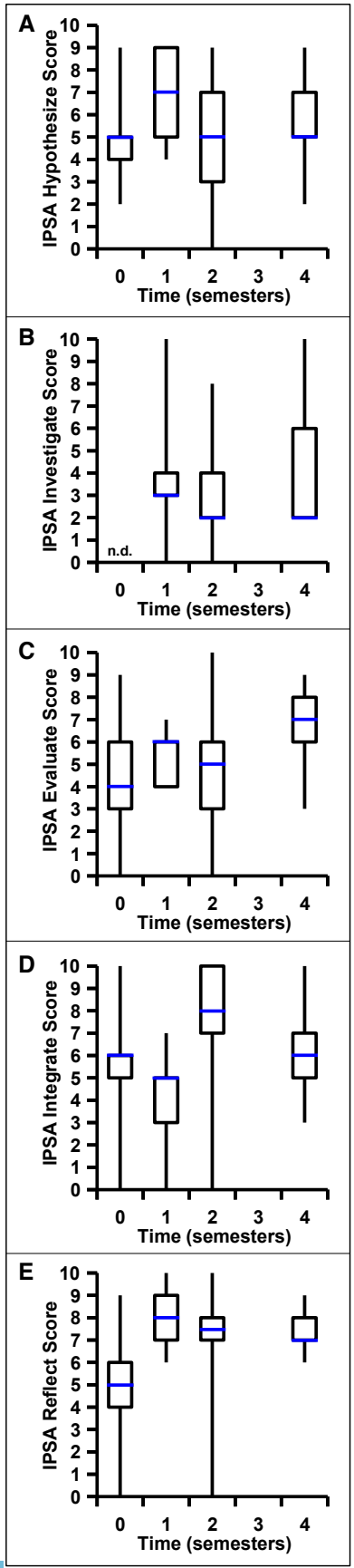


Figure C.VI.2: IPSA Score Distributions. Boxplots depict the spreads of IPSA scores across four semesters in each domain: **(A)** Hypothesize, **(B)** Investigate, **(C)** Evaluate, **(D)** Integrate, and **(E)** Reflect. n.d., no data.

Appendix C.VII
Cohort Differences

Methods

Analyses of variance (ANOVAs) – Differences between the cohorts occurred during the three years of this study. Regarding instructors, WA taught the BIOC II course for Cohorts A and B, while MO taught the course for Cohort C. Assessment differences also occurred. Content exams were slightly modified from year to year, since past exams were made available to students. One IPSA prompt was modified in the third year, the Reflect prompt of the IPSA at T=0, to ask students to evaluate their performance (Appendix III). Finally, active learning was incorporated into both biochemistry courses for all cohorts, yet to an increasing degree across time.

ANOVAs determined whether the cohorts exhibited statistically significant differences in mean scores, for the scores analyzed. Course grades, content exam scores, and IPSA domain scores were dependent variables. The independent variable was the cohort. The assumption of normality was tested by visual inspection of distribution histograms, and the assumption of homogeneity was tested using Levene's statistic. Fisher's LSD pairwise comparisons maintained the significance level at 0.05. Cohen's *d* values were calculated to estimate the effect size of statistically significant differences. Values of at least 0.2 are considered small differences, at least 0.5 are medium differences, and at least 0.8 are large differences.

Results

Among the three student cohorts, the assumption of normality was reasonably met, yet a few measures did not meet the assumption of homogeneity of variance, specifically: the BIOC II course grade (Levene's = 18.876, $p < 0.001$), the IPSA Integrate score at T=1 (Levene's = 15.427, $p < 0.001$), and the IPSA Reflect score at T=2 (Levene's = 3.559, $p = 0.032$). To assume homogeneity of variance is to say that the scores of each cohort have equal variances. Since course grades were expected to be different due to a change of instructor, the assumption's violation is not a serious practical concern. Also, the cohorts were pooled since we expected some differences across cohorts, and took an approach to

include more broadly diverse students. Therefore, the apparent violation in two IPSA domain scores is not of practical significance.

ANOVAs showed that statistically significant differences existed among cohort means for all measures except the IPSA Investigate domain score. Pairwise comparisons of mean scores revealed among which cohorts the differences existed (Table C.VII.1). Most prominently, BIOC II course grade means differed among all three cohorts (Fig. C.VII.1A). Fewer differences existed across cohorts for content exam scores (Fig. C.VII.1B) and IPSA domain scores (Fig. C.VII.2).

Table C.VII.1. Effect Sizes of Cohort Differences

Measure	Cohorts	*F*	*p*	*d*	Size
BIOC I	B/C	5.931	0.017	0.58	Medium
	A/C	6.248	0.014	0.62	Medium
BIOC II	A/B	14.322	< 0.001	1.14	Large
	B/C	58.874	< 0.001	1.82	Large
	A/C	7.104	0.009	0.69	Medium
Content Exam, T=0	B/C	23.334	< 0.001	1.15	Large
	A/C	7.603	0.007	0.69	Medium
Content Exam, T=2	B/C	19.536	< 0.001	1.10	Large
IPSA Hypothesize, T=0	A/C	7.292	0.008	0.80	Large
IPSA Hypothesize, T=1	A/B	9.747	0.003	0.60	Medium
IPSA Evaluate, T=0	B/C	4.806	0.031	0.52	Medium
IPSA Evaluate, T=4	A/B	7.931	0.007	0.65	Medium
IPSA Integrate, T=2	B/C	8.440	0.004	0.69	Medium
IPSA Reflect, T=2	B/C	8.392	0.005	0.69	Medium

Despite these findings, we do not consider the differences to be practically significant in terms of this study. Some differences would be expected due to student variability in relatively small cohorts (under 100 students). Yet the disciplinary focus remained upon a biochemistry curriculum across multiple semesters. Cohorts A and B were comprised entirely of two-year biochemistry majors, and 75 percent of one-year students in Cohort C were biochemistry majors.

When researching a convenience sample of biochemistry majors at our institution, we argue that the variability seen within this study's pooled group would also be detected in a single cohort of more than 100 students, if it existed. By pooling the cohorts, our analytical

results are more broadly generalized to our population of biochemistry students than results for any single cohort of the study. We concluded that the statistical differences found when comparing cohorts were not meaningful in terms of research implications.

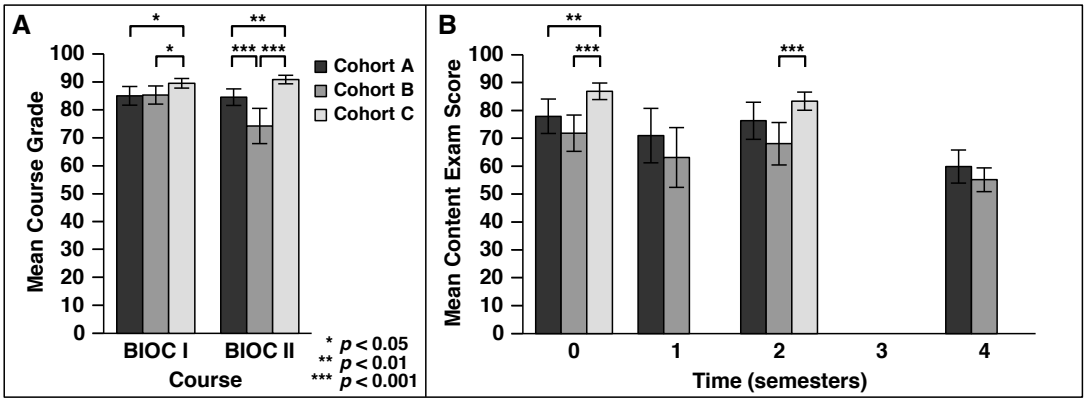


Figure C.VII.1: Academic Scores Across Cohorts. Mean scores for each cohort in the study were compared by ANOVA. Cohorts that exhibited statistically significant differences on a particular measure are marked with a bracket. The measures included **(A)** course grade and **(B)** content exam score. Error bars indicate 95% confidence intervals for the mean.

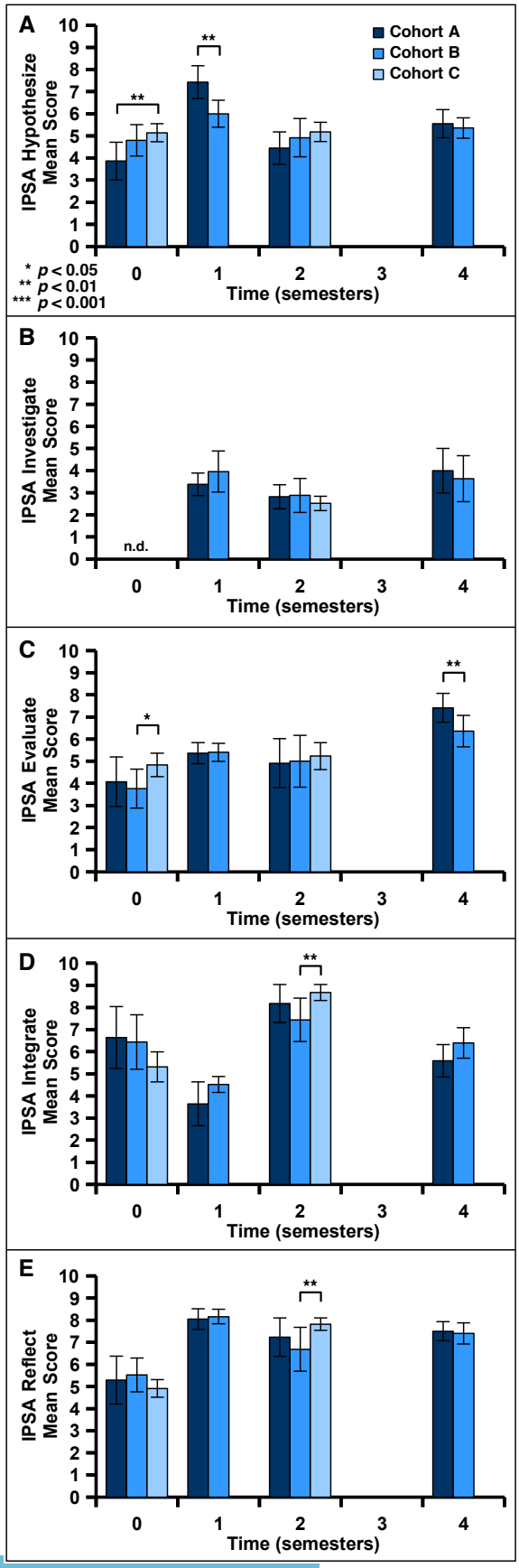


Figure C.VII.2: IPSA Scores Across Cohorts. Mean scores for each cohort in the study were compared by ANOVA. Cohorts that exhibited statistically significant differences are marked with a bracket. Error bars indicate 95% confidence intervals for the mean. n.d., no data.

References

American Association for the Advancement of Science. (2011). Vision and Change in Undergraduate Biology Education: A Call to Action.

American Society for Biochemistry and Molecular Biology. (2012). Policies and procedures for certifying bachelor's degrees in biochemistry and molecular biology.

Association of American Medical Colleges, AAMC-HHMI Committee. (2009). Scientific Foundations for Future Physicians. Washington, D.C.

National Research Council (2012). *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering* (S. R. Singer, N. R. Nielsen & H. A. Schweingruber Eds.). Washington, DC: The National Academies Press.