1-2011

What Really Matters: Assessing Individual Problem-Solving Performance in the Context of Biological Sciences

Steven M. Mitchell
University of New Mexico, smmitchell@salud.unm.edu

William L. Anderson
University of New Mexico, wanderson@salud.unm.edu

Cheryl A. Sensibaugh
University of New Mexico, CSensibaugh@salud.unm.edu

Marcy Osgood
University of New Mexico, mosgood@salud.unm.edu

What Really Matters: Assessing Individual Problem-Solving Performance in the Context of Biological Sciences

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

What Really Matters: Assessing Individual Problem-Solving Performance in the Context of Biological Sciences

William L. Anderson
wanderson@salud.unm.edu

Cheryl A. Sensibaugh
CSensibaugh@salud.unm.edu

Marcy P. Osgood
mosgood@salud.unm.edu

Steven M. Mitchell
University of New Mexico School of Medicine
Albuquerque, New Mexico, USA
smmitchell@salud.unm.edu

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.

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.

Methods

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 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 1.
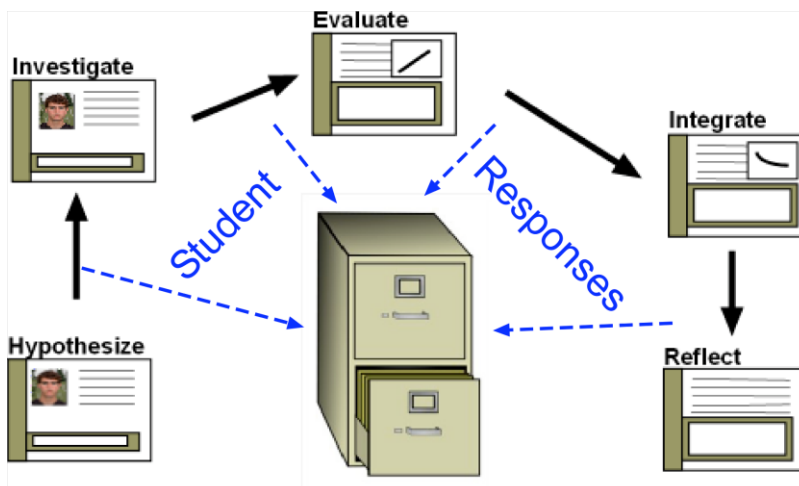
Figure 1. Individual Problem-Solving Assessment Structure. Each of the five domains of problem-solving are incorporated into the IPSA. To assess each domain on its own merit, student responses are collected in sequence and stored in a database. The software only allows forward progression through the 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.

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.

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. The two course instructors are responsible for grading the exams and providing feedback to students as necessary.
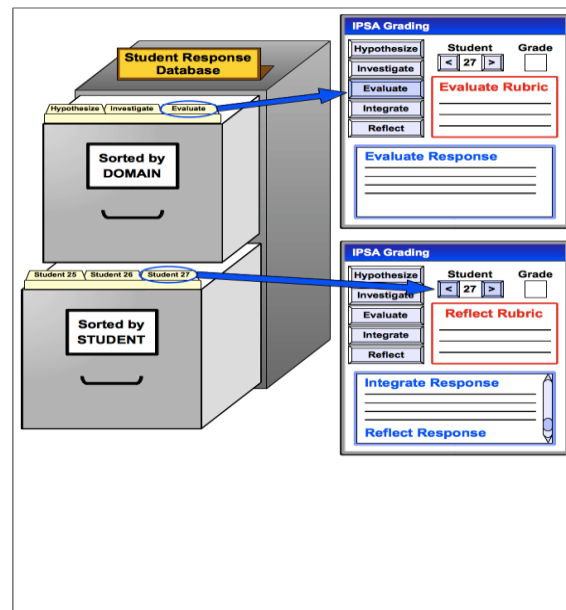

Figure 2. Database grading. Student responses to an IPSA may be retrieved from the database and sorted either by domain or by student. The grading rubric for each domain (red box) is also shown with the student records.

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). 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.

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.

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.

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 3 illustrates how student problem-solving domain patterns, or profiles, are depicted.
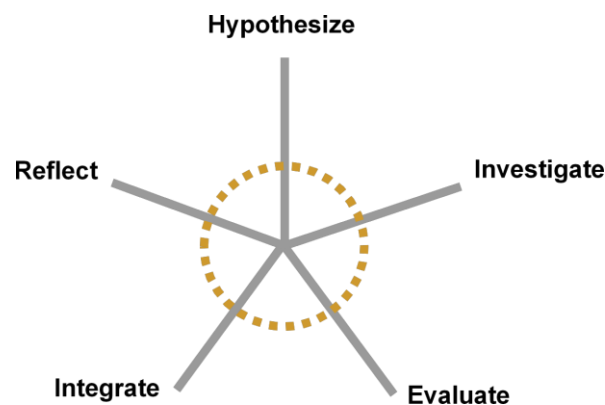

Figure 3: Radar Plot. A visual representation of the scoring ranges on an IPSA, with axes for each of the problem-solving domains (gray). Expected performance scores are indicated by the circular target (yellow dotted outline).

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. 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.

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.

Results and Discussion

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.

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 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 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.

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.

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 4 depicts the IPSA rubric-based scores assigned to two representative students by these three graders, with 4A and 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.
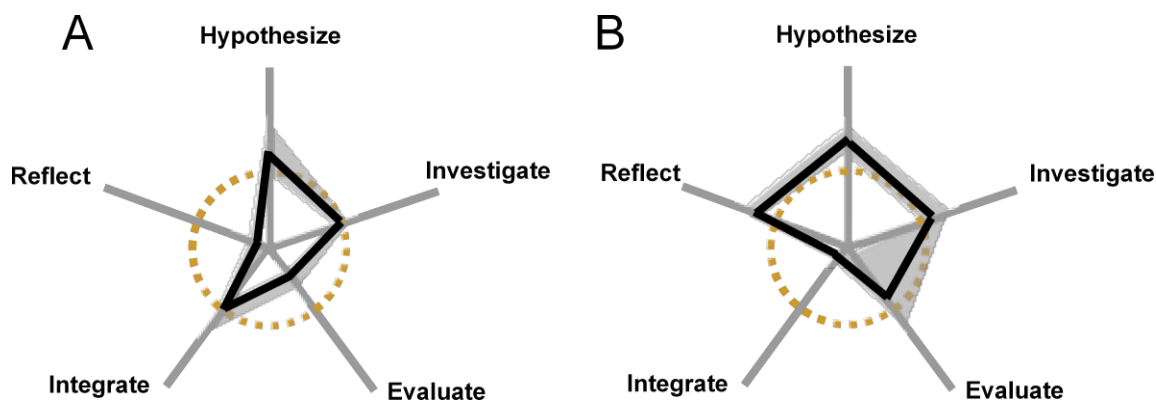
Figure 4. Inter-grader Reliability. Radar plots show mean scores assigned by three faculty members (black) and standard deviations (gray) for two representative students' IPSA results. The plot in **(A)** indicates standard deviations of less than 10%, while the plot in **(B)** indicates variability in grading the **Evaluate** domain.

Figure 4A illustrates an example of our typical grading consistency, with less than a 10% standard deviation between multiple graders. On the other hand, Figure 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. 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 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.
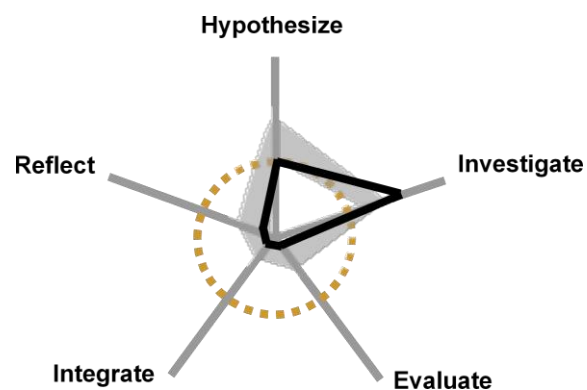
Figure 5. Graduate Student Grader Reliability. The radar plot of one student's IPSA grades, as assigned by a trained graduate student (black) and faculty (gray).

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 4 and 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.

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. 6. A full library of archetypal performance patterns has not yet been defined, and is under investigation.
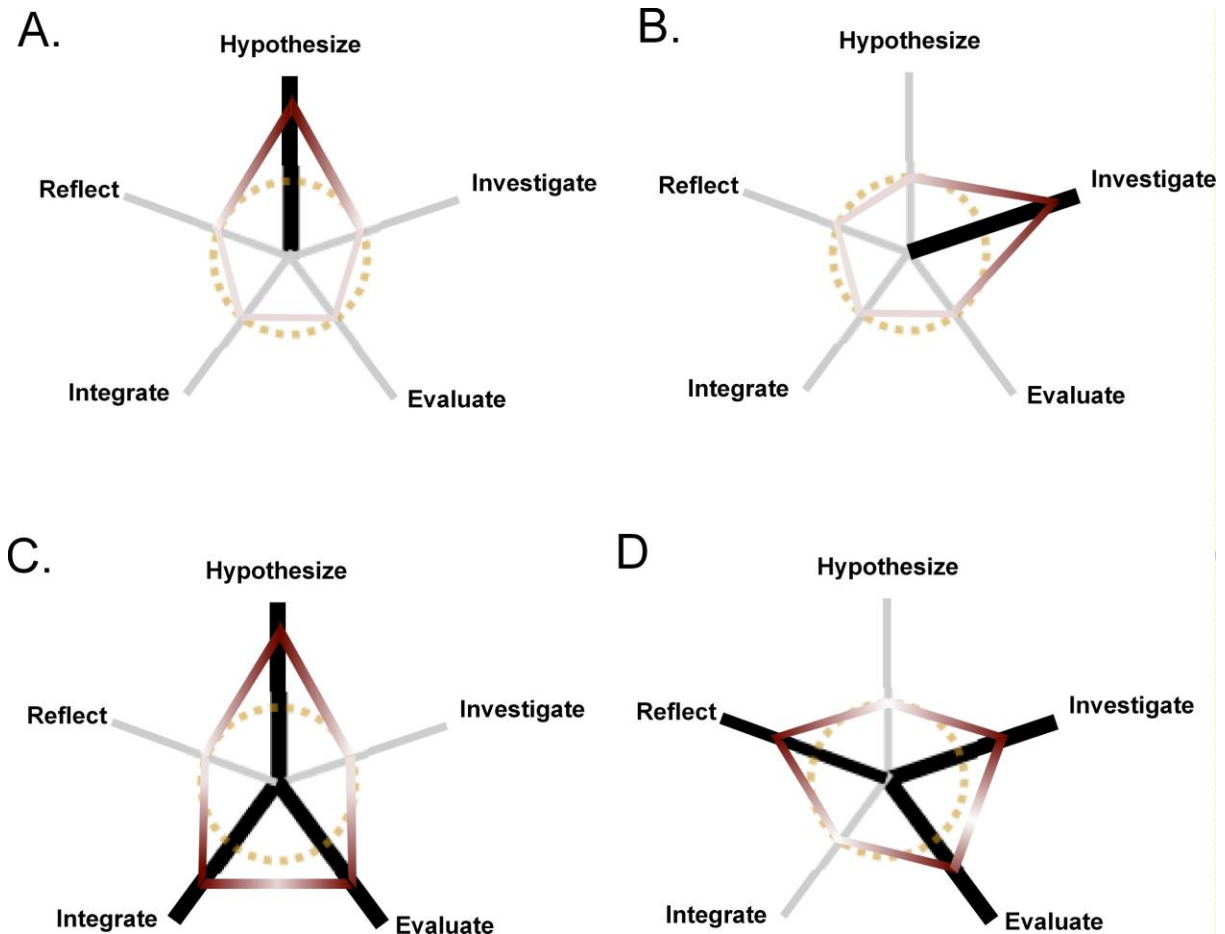
Figure 6. Four Common Student Performance Patterns on IPSAs. In **(A)**, students exhibit difficulty in the **Hypothesize** domain. In **(B)**, the low **Investigate** domain score indicates a challenge with contextualizing hypotheses within the scenario. In **(C)**, the low scores for both the **Evaluate** and **Integrate** domains correlate with a lack of content knowledge. In **(D)**, difficulty in the **Reflect** domain reflects poor metacognition.

Figure 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. 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. 6B was to increase the student's sensitivity to the environment of the problem.

Students exhibiting the pattern illustrated in Fig. 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 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 6A and 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 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 (Ash and Clayton, 2009). 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.

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 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. 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. 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.
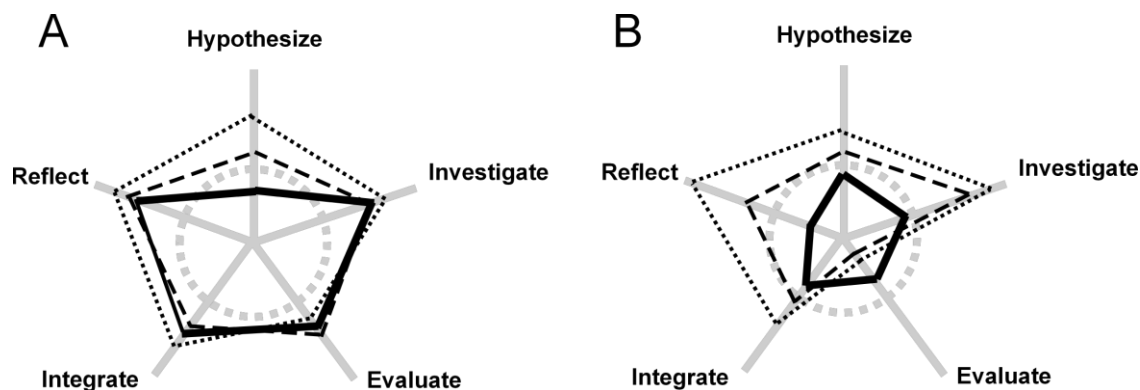
Figure 7: Longitudinal Performance Patterns. The change in two students' IPSA performance patterns over two semesters, at three points in time: initial (fine dashed line), midway (broad dashed line), and final (continuous line). Student **(A)** was only able to significantly improve in the **Hypothesize** domain, while student **(B)** made substantial strides and eventually exceeded z expectations in all the domain scores.

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.

References

Allen, S. and 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

Appendix

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. 1. Note that the initial case presentation is in a scrolling box to permit the possibility of using large or small case presentations.

**Fig.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. 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.

Fig. 2: Results of a literature search
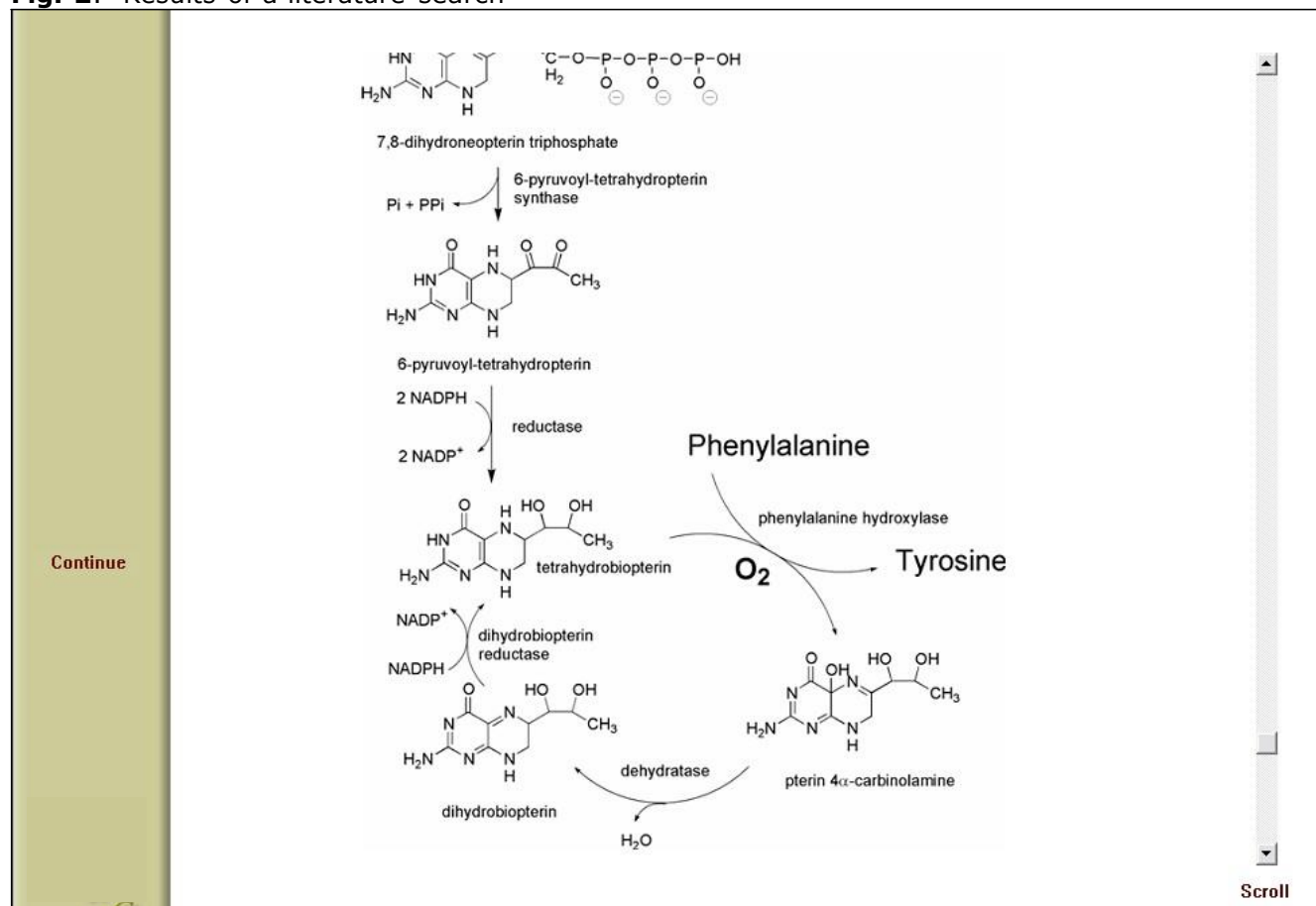
As the case progresses, students are sequentially asked to investigate a specific hypothesis by designing an experiment, to evaluate data that results from an experiment, and eventually to solve a dilemma related to the experimental data that requires the student to integrate the basic science knowledge about the topic in order to argue in support on one side of the dilemma or the other side. Figure 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.

Fig. 3: Presentation of graphical data
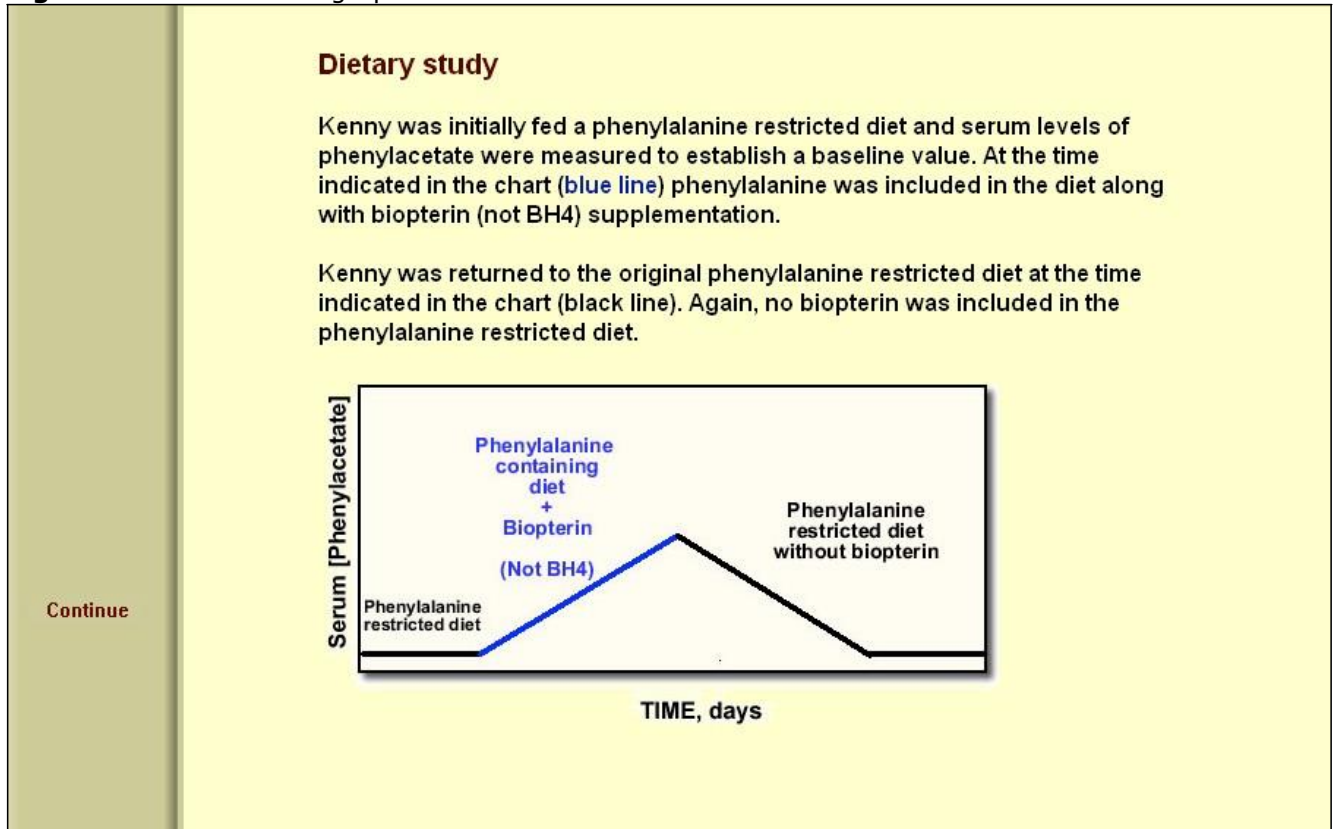
The student's responses to these questions are entered into textboxes, as illustrated in Fig.4. These text boxes can be set to limit the number of words available to the student. This has been found to be very effective in preventing students from writing everything they know about a topic in a "shotgun" type of answer and forces them to focus on answering a specific question.

Fig. 4: Student answers entered into a text box

Two different physician, Drs. Smith and Skip Tecall, looked at the same data you just evaluated and had completely different recommendations. **How will you respond to both of the physicians.**

Dr. Smith was incensed that you would question her initial decision about Kenny's problems. Her recommendation was that while Kenny is growing and developing, he must remain on the traditional phenylalanine restricted diet. Dr. Smith further recommends that, once Kenny is an adult, the phenylalanine free diet is no longer necessary.

Dr. Tecall. was far more troubled with the data. He stated that this was not a conclusive experiment to isolate Kenny's problems to the phenylalanine hydroxylase enzyme. He maintains that defects in other enzyme are also a possibility given the data and that the experiment could have been set up differently

(Press the [TAB] key to either erase or file your answer.)

Once the students have completed the examination, their responses are automatically saved to a database for grading. Figure 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.

Fig. 5: Database grading screen with grading rubrics

The text box in the figure reads:

> I would suggest that Dr. Smith evaluate Kenny's phenylacetate and phenyllactate levels as he matures before suggesting that he can simply be taken of the diet. I would also agree with Dr. Tecall the test does not difinetively pinpoint phe hydroxylase as the problem. Infact it suggests that there may be a problem in the tetrahydrobiopterin synthesis pathway. If kenny was able to produce the active cofactor from dietary biopterin then the phenylacetate and phenyllactate levels may have dropped if this was his problem. It may have been more conclusive if the levels of BH4 were monitored or if BH4 was supplemented.

Rubric: Evaluate Domain