Undergraduate Performance in Solving Ill-Defined Biochemistry Problems

Cheryl A. Sensibaugh,[†]* Nathaniel J. Madrid,[‡] Hye-Jeong Choi,[§] William L. Anderson,[‖¶] and Marcy P. Osgood[¶]

[†]Department of Biochemistry and Molecular Biology and [§]Department of Educational Psychology and Georgia Center for Assessment, University of Georgia, Athens, GA 30602; [‡]Department of Internal Medicine and [¶]Department of Biochemistry and Molecular Biology, University of New Mexico, Albuquerque, NM 87131

ABSTRACT

With growing interest in promoting skills related to the scientific process, we studied performance in solving ill-defined problems demonstrated by graduating biochemistry majors at a public, minority-serving university. As adoption of techniques for facilitating the attainment of higher-order learning objectives broadens, so too does the need to appropriately measure and understand student performance. We extended previous validation of the Individual Problem Solving Assessment (IPSA) and administered multiple versions of the IPSA across two semesters of biochemistry courses. A final version was taken by majors just before program exit, and student responses on that version were analyzed both quantitatively and qualitatively. This mixed-methods study quantifies student performance in scientific problem solving, while probing the qualitative nature of unsatisfactory solutions. Of the five domains measured by the IPSA, we found that average graduates were only successful in two areas: evaluating given experimental data to state results and reflecting on performance after the solution to the problem was provided. The primary difficulties in each domain were quite different. The most widespread challenge for students was to design an investigation that rationally aligned with a given hypothesis. We also extend the findings into pedagogical recommendations.

INTRODUCTION

As part of ongoing efforts to improve undergraduate biology education, facilitating the learning of core competencies and disciplinary practice requires research to answer many questions. The overall goal of our research is to help students learn the process of science, or to understand that "biology is evidence-based and grounded in the formal practices of observation, experimentation, and hypothesis testing" (American Association for the Advancement of Science, 2011, p. 14). To address this goal, the current work focuses on the summative assessment of students' abilities to apply the process of science, so that we may understand where to target future improvement efforts.

A Constructivist Framework of Learning

The theoretical framework of constructivism explains learning by taking the viewpoint that new knowledge is constructed using the building blocks of prior knowledge and experience (Bodner, 1986; Bodner and Orgill, 2007). Through decades of work in cognitive psychology, the constructivist theory of learning has developed several branches from this main trunk to further postulate ways in which knowledge building occurs, such as Kelly's theory of personal constructs, Piaget's personal constructivism, Solomon's social constructivism, and von Glasersfeld's radical constructivism (Bodner *et al*., 2001). Broadly speaking, personal constructs and personal constructivism highlight the role of the individual learner, while social constructivism highlights the role of the group(s) of which the learner is a part. A meta-analysis of recent advances in

Jennifer Momsen, *Monitoring Editor*

Submitted April 29, 2015; Revised August 29, 2017; Accepted September 7, 2017

CBE Life Sci Educ December 1, 2017 16:ar63

DOI:10.1187/cbe.15-04-0106

[‖]Professor emeritus.

*Address correspondence to: Cheryl A. Sensibaugh (csensiba@uga.edu).

science, technology, engineering, and mathematics education research implies that multiple forms of constructivism are applicable in the classroom, given the benefits of active learning in both individual and group settings (Freeman *et al.*, 2014). Therefore, we do not focus our theoretical framework on any particular form of constructivism. Instead, our efforts to promote problem-solving abilities in biochemistry rely on group activities as well as individual efforts. We emphasize the broader outcome of gaining knowledge by building upon what students have already learned and experienced. Furthermore, in operationalizing scientific problem solving, our research takes the stance that different facets of problem solving are constructed from identifiable and distinct building blocks. Those components take the form of measurement criteria during assessment of problem-solving ability.

Problem Solving

The manner in which people solve problems has been an area of inquiry for cognitive psychologists, educational psychologists, learning scientists, neuroscientists, and discipline-based education researchers. The nature of the problem has much to do with how problem solving is examined. A problem may be either domain general or domain specific. A domain-general problem, such as one encountered in everyday life, does not require any specialized knowledge. A domain-specific problem necessitates that particular knowledge be brought to bear to successfully solve the problem. Another characteristic to consider about the nature of the problem is its structure; that is, whether it is well defined or ill defined. Well-defined problems are constrained by multiple conditions and result in a limited number of solutions. In contrast, ill-defined problems are vague, present with relatively little information, and yield a greater number of solutions than well-defined problems. Both domain-general and domain-specific problems can be either well or ill defined.

Newell and Simon (1972) asserted a theory of human problem solving that accounted for the prior knowledge held by the solver, thus turning the corner from research of domain-general problem solving to domain-specific problem solving. Their theory proposed four elements that work in concert to reach a solution: human characteristics, the context of the problem, the structure of the problem, and potential paths to a solution (p. 789). Subsequent work by Chi and colleagues (1981) was important in differentiating how novices and experts (human characteristics) solved problems in physics (context) that were well defined (structure) by categorizing and representing the problems (solution paths). A main finding was that novices focused on literal surface features of the problem, while experts used an approach based on deep, abstract features. Differences between novice and expert problem solving have also been well documented in genetics (Smith and Good, 1984; Smith, 1988), evolution (Nehm and Ridgway, 2011), biology (Coley and Tanner, 2012), and chemistry (Bodner, 2015). Additionally, much of the research has focused upon well-structured problems in these contexts. Yet in biochemistry, relatively little work has been done to understand how ill-defined problems are successfully solved, or what might contribute to a lack of success.

Newell and Simon's theory of human problem solving suggests that it is impossible to fully extricate the ability of problem solving from the disciplinary context and specific content knowledge necessary to reach a solution. An especially useful framework for categorizing different types of domain-specific knowledge has been articulated by Alexander and colleagues (Alexander and Judy, 1988; Alexander *et al.*, 1989; Murphy and Alexander, 2002). First, declarative knowledge is "knowing what"; that is, having an understanding of factual content. Second, procedural knowledge is "knowing how" to apply declarative knowledge to carry out a strategy; that is, having the ability to solve problems. Third, conditional knowledge is "knowing when and where" to bring particular declarative and procedural knowledge to bear; that is, whether certain content and strategies are relevant to solving a given problem. Alexander's studies postulate that interactions exist between different types of knowledge. More recent discipline-based education research suggests the same (Prevost and Lemons, 2016).

Finally, metacognition also plays a role in problem solving. Metacognition, which loosely means "thinking about thinking," is formally defined as being both aware of thinking and able to control thinking (Cross and Paris, 1988). Thus, metacognition comprises both metacognitive knowledge and metacognitive regulation. Metacognitive knowledge allows one to identify what is known and what is unknown. Metacognitive regulation refers to one's control of thinking by taking action to learn. Past efforts in the field of metacognition were recently applied to biology education research (Stanton *et al.*, 2015; Dye and Stanton, 2017). Furthermore, a correlation between metacognition and the ability to solve problems has been demonstrated in chemistry (Rickey and Stacy, 2000; Sandi-Urena *et al.*, 2011). These efforts suggest that both components of metacognition are important when targeting potential causes of poor performance in problem solving.

This study seeks to address some of the recommendations in the National Research Council's report on discipline-based education research (NRC, 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). Specific areas now 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. In biochemistry courses at a large, public university in the southwestern United States, we operationalized the construct of scientific problem solving for pedagogical purposes (Anderson *et al.*, 2008; Mitchell *et al.*, 2011). We also discuss recommendations for pedagogical practice to maintain student-centered learning as a crucial underpinning for the research and to inform scholarly educators.

To address our goal of promoting the ability to solve ill-defined biochemistry problems, this work poses two research questions. First, "How do graduating biochemistry majors perform in scientific problem solving?" We describe performance quantitatively, in terms of scores derived from applying scoring rubrics. Before this study, in-depth qualitative analysis of student responses—beyond the rubric criteria—had not been performed. It is crucial to gain insight into students' solutions to

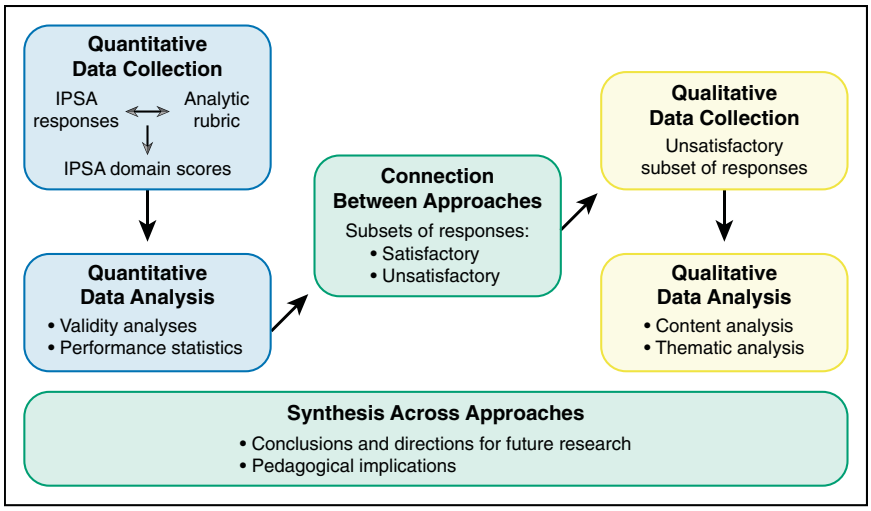

FIGURE 1. Mixed-methods study design. A sequential explanatory design was employed to generate evidence toward answering our research questions. Quantitative data collection and analysis (blue boxes) informed our first research question, while qualitative data collection and analysis (yellow boxes) addressed our second research question. Bridges connected and synthesized the two approaches (green boxes).

ill-defined problems and understand the diverse perspectives that unsuccessful students adopt in this process, so that we can better facilitate learning. We also ask, "What is the nature of unsatisfactory solutions to ill-defined biochemistry problems?" A mixed-methods approach (Figure 1) allowed us to both describe and begin explaining student performance (Ivankova *et al.*, 2006; Warfa, 2016).

On the basis of our prior work (Mitchell *et al.*, 2011) and interim preliminary analyses (unpublished data), we hypothesized that students would exhibit domain-specific difficulties. Various trends, or patterns of performance, had emerged. We suspected that additional patterns remained to be uncovered. The nature of unsatisfactory solutions had only previously been informed by experience and informal discussions with students, so we hypothesized that a wide range of possibilities would exist to explain the observed performance patterns.

METHODS

Educational Setting

The pedagogy for this study was carried out within two biochemistry courses: one semester on biomolecular structure and function (BIOC I) and a second semester on intermediary metabolism (BIOC II; Figure 2). These courses were required of biochemistry majors and were typically taken during the junior year. Table 1 summarizes the specifications and constructive alignment of course elements we developed for scientific problem solving (Biggs, 1999; Handelsman *et al.*, 2004, 2006). In our previous work, we defined problem solving as consisting of the scientific method along with metacognition (Anderson *et al.*, 2008). We stated learning objectives to align with that definition, so that each objective addressed one aspect, or domain, of scientific problem solving. All the learning objectives related to higher-order cognitive skills, such as applying and synthesizing information (Table 1), rather than lower-order cognitive skills such as remembering and understanding information (Bloom *et al.*, 1954; Anderson and Krathwohl, 2001).

The Individual Problem Solving Assessment (IPSA) was a computer-based summative measure of student performance for each objective (Mitchell *et al.*, 2011). An IPSA followed one biochemistry problem explicitly through each of the five domains. The mechanics of an IPSA involved progressively revealing each domain to students. Students could review—but not go back and alter—completed domains at any time (Figure 3). Each domain contained one item that prompted for an written response.

An IPSA opened with a scenario describing observations about a biochemical problem (Figure 3A). Only the Hypothesize domain was accessible to students at this point. After providing minimal information to supplement the observations, the IPSA prompted students to generate multiple hypotheses that explain the observed phenomenon. Once students entered their hypotheses, the Investigate domain became accessible, while subsequent domains remained inaccessible to students (Figure 3B). Here, students were prompted to design an experiment that would test a single given hypothesis, which was specified within the prompt. In the third section of an IPSA, the Evaluate domain, experimental results were provided in the form of figures, graphs, or tables, and students were prompted to evaluate the results (Figure 3C). Then, in the Integrate domain, the results were given and more data were provided. Students were prompted to integrate all available IPSA information into the original context of the problem, using course content knowledge, to come to a conclusion concerning the biochemical problem (Figure 3D). Finally, when the Reflect domain was reached, a plausible conclusion was provided, and students were asked to reflect on their responses (Figure 3E). Students typically completed an IPSA within 45 to 75 minutes.

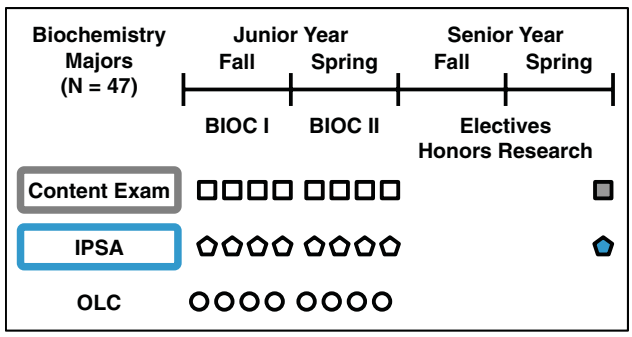

FIGURE 2. Educational setting and data collection. Two cohorts of biochemistry majors took nine content exams (squares) and nine IPSAs (pentagons) during their junior and senior years. Content exam scores and IPSA responses at program exit (filled polygons) were collected for analyses. The curriculum also included eight OLC group activities (circles).

TABLE 1. Table of specifications and constructive alignment of problem-solving course elements

Learning goal: Solve ill-defined biochemistry problems using the scientific method and reflect upon the process using metacognitive strategies				
Problem-solving domain	**Learning objectives**	**Number of assessment items by cognitive level**		**Learning activity**
		Lower	**Higher**	
Hypothesize	Given a set of observations, students should be able to generate hypotheses about potential biochemical mechanisms underlying biological phenomena.	0	1	Online case (OLC)
Investigate	Given a testable and falsifiable hypothesis regarding one distinct biochemical mechanism, students should be able to propose an experimental design to test that hypothesis.	0	1	
Evaluate	Given an experimental design and data, students should be able to deduce the experimental results.	0	1	
Integrate	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.	0	1	
Reflect	Given a conclusion, students should be able to critically evaluate their own performance.	0	1	
	Total number of items	0	5	1

Rubrics for instructors to grade IPSA responses contained specific criteria for scoring each domain on a scale of 0 to 10, with a score of 7 points defined as satisfactory performance (see the Supplemental Material). When determining whether expectations related to problem solving were met, the specific biochemistry content in a student response served as an identifiable marker of skill. Although the contextual milestones for each score differed across IPSA rubrics, the rubrics were designed so that score interpretation related to problem-solving ability was consistent across IPSAs and across domains (i.e., that 7 points is satisfactory, 10 points is exemplary). While the rubrics did contain content-specific markers, it is important to emphasize that IPSAs were not intended to measure content knowledge, nor have they been found to do so (Mitchell *et al.*, 2011). In this way, scores were generated for each student, in each domain, and on each IPSA that quantified performance in problem solving while recognizing the contextual cues within the problem.

Our learning activities, termed online cases (OLCs), were designed for student groups rather than individuals (Anderson *et al.*, 2008). Similar to an IPSA, the OLCs presented vaguely defined problems, situated within real-world contexts, and required application of the scientific method. To introduce students to the problem-solving process, instruction early in the first semester employed an example case. The instructor facilitated a class-wide discussion using Socratic questioning. Students then worked in groups on subsequent OLCs using a Web-based asynchronous discussion board. Each group was facilitated by either the instructor or a teaching assistant, who monitored the discussion board and guided students through the scientific ways of thinking about problem solving. An OLC was open to students for about 2 weeks. The scoring rubrics for OLCs generated one overall case score for all members of a group, rather than domain scores for each student. The scores were based on common milestones of progression through the case that were determined during development. When the discussion boards were closed, two forms of feedback were offered to students in addition to scores. Documents were posted online that addressed common difficulties and modeled successful strategies. Additionally, discussion time was devoted to the OLCs in class to allow students to ask specific follow-up questions.

The biochemistry courses were each divided into four units (Figure 2). Students repeatedly practiced their problem-solving skills by completing one OLC in each unit, relevant to the current course topics. At the end of each unit, a content exam and IPSA were administered. Content exams, given during a class session, contained multiple-choice and short-answer items that primarily measured lower-order cognitive skills. Because IPSAs were computer-based, they were administered in a computer laboratory over a span of 3 days. Students scheduled a time outside class to complete each IPSA. Scores on the four OLCs and four IPSAs combined to account for 10% of a course grade. Ninety percent of course grades were determined by content exams, short quizzes, and content-oriented activities.

Before graduation, biochemistry majors were required to complete two program exit assessments: one that measured accumulated content knowledge, and one that measured accumulated problem-solving skill (Figure 2). To measure content knowledge at program exit, students completed the nationally standardized American Chemical Society (ACS) 2003 Biochemistry Exam. To assess graduating majors' ability to solve problems, we used a program exit IPSA titled "The Lorrat" (see the Supplemental Material). The problem presented in "The Lorrat" IPSA required application and synthesis of knowledge from both biochemistry courses in order to be solved. Although no score threshold was set in order to graduate, students were encouraged to do their best.

Data Collection

The study was conducted retrospectively 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. Two cohorts of students were pooled ($N = 55$); each entered the biochemistry curriculum in sequential academic years ($n_1 = 23$, $n_2 = 32$). After excluding six students who compressed the program timeline, and two students who transferred credit for BIOC I and II, the sample

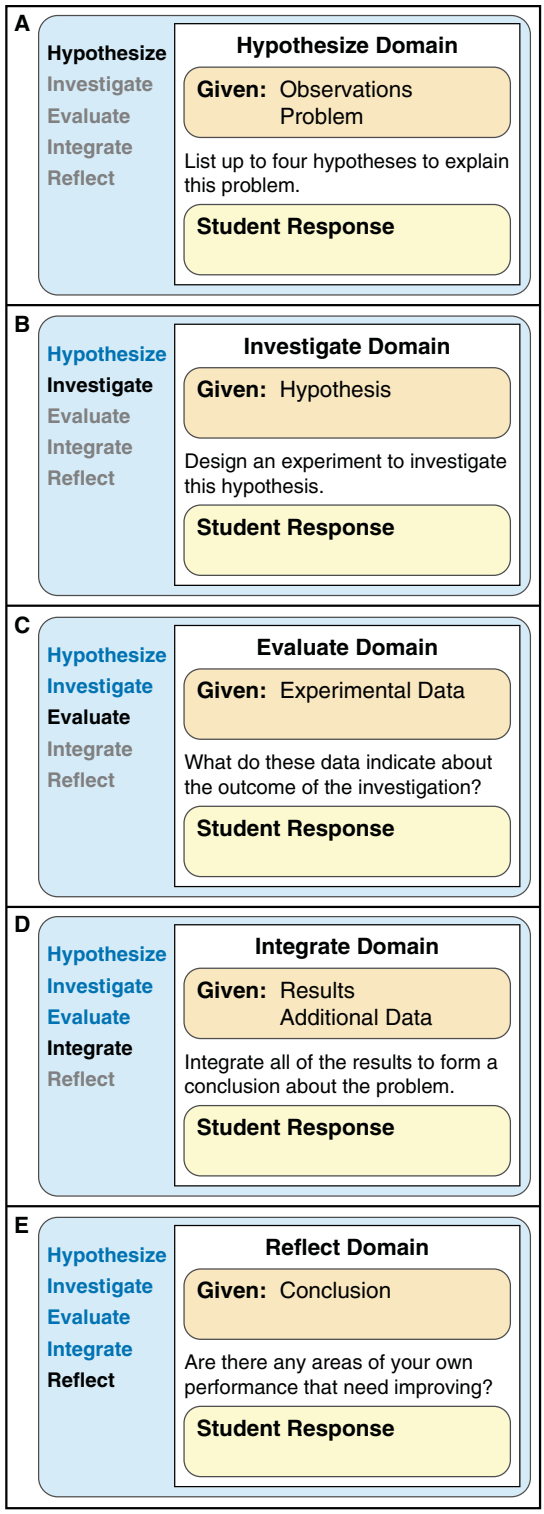

FIGURE 3. IPSA mechanics. The progressive-reveal nature of an IPSA is captured in simplified versions of screen shots from each domain during computer administration. (A) Hypothesize, (B) Investigate, (C) Evaluate, (D) Integrate, and (E) Reflect. Black domain text on the left indicates the currently 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.

included 47 participants. Scores on the ACS exam and responses on "The Lorrat" IPSA were analyzed (Figure 2). Our rationale for focusing solely upon the IPSA at program exit, rather than across multiple IPSAs, was twofold. First, doing so allowed concurrent consideration of content knowledge as determined by a nationally standardized assessment, instead of by assessments that were locally generated. Second, our research question guided us to determine how students performed at graduation before considering an investigation within the curriculum. That is, if the evidence showed that most students performed well after completing the program, there would be less concern about longitudinal specifics.

Student Backgrounds

All study participants were biochemistry majors. One-third enrolled in honors research courses and presented a thesis. Other research experience was not measured, because laboratory experiences outside of our department could not be controlled for biochemistry content. All students completed prerequisite courses, which included laboratory components, yet those experiences were guided by step-by-step protocols rather than by employing the process of scientific inquiry or requiring experimental design.

Regarding demographics, most students were traditionally aged Caucasian males. However, 13% of the sample was composed of returning students, and 36% of all students were female. The Hispanic or Latino/a population was represented by 34% of students, 11% were Asian, and 2% were American Indian.

Statistical Analyses

SPSS (version 23, IBM) was used for all analyses. Descriptive statistics summarized student backgrounds and performance. Means of IPSA domain scores were calculated with 95% confidence intervals. Inferential statistics with correlation analyses allowed testing of the null hypothesis that Pearson's correlation coefficients (r) were not significantly different from zero, with alpha set to 0.05. 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).

IPSA Score Validity

Table 2 summarizes the validity argument and approach to validating the IPSA (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014; Reeves and Marbach-Ad, 2016). Given that the IPSA was designed to measure knowledge of solving ill-defined problems, several methods were used to generate evidence toward supporting claims about the items and scores. We also compared our results from this study with those of our previous work with the IPSA (Mitchell *et al.*, 2011).

A table of specifications, or test blueprint, formalized prior definitions of the procedural knowledge concepts each item was intended to assess (Table 1). The competencies were explicitly aligned with higher-order cognitive levels. Test content lends support to the claim that IPSA items represent a variety of domains of scientific problem solving.

Sample responses to each IPSA prompt, which were representative of typical responses, were compiled. Light edits to

TABLE 2. IPSA validity argument and approach

Intended use of the IPSA: Support inferences from domain scores about a student's procedural knowledge of solving ill-defined problems			
Claims	**Categories of validity evidence**	**Methods of determination**	**Studies[a]**
Items represent a variety of domains of scientific problem solving.	Test content	Align items with concepts assessed Table of specifications	2011: pp. 16–20 This study: Table 1
Items engage students in the domains of problem solving.	Response processes	Sample responses	This study: Table 3
Domain scores are distinct from one another.	Internal structure	Align domains with steps of the scientific method and metacognition Correlation analysis	2011: pp. 4, 9 This study: Table 4
Domain scores are somewhat related to—yet distinct from—scores of content knowledge and research experience.	Relations with other variables[b]	Correlation analysis	2011: p. 9 This study: Table 4

[a]The 2011 study (Mitchell *et al.*, 2011) sampled medical students, while this study sampled biochemistry students.
[b]The measures of content knowledge were the Comprehensive Basic Science Exam (2011 study) and the ACS Biochemistry Exam (this study).

punctuation were made to improve readability. Student responses were reviewed with an eye toward whether students were attempting to display the intended procedural knowledge for each domain. The response processes, or ways that students respond to the prompts, would support the claim that IPSA items engage students in the domains of problem solving.

Correlation analyses were performed to determine the relationships among IPSA domain scores, content exam scores, and whether or not students engaged in undergraduate honors research experience. Quantifying relatedness of the five domain scores in terms of correlations was an examination of the IPSA's internal structure. Such evidence would lend support to the claim that domain scores are distinct from one another. Additionally, determining the relations that domain scores had with content exam scores and research experience would support the claim that those factors are somewhat related, yet remain distinct.

Content Analysis

To gain insight into the nature of unsatisfactory responses, we used qualitative content analysis (Patton, 2015) to identify common elements of student writing on "The Lorrat" exit IPSA. Responses were transferred from Excel into MAXQDA (version 12, VERBI GmbH). The first iteration of the list of codes was established using rubric criteria (see the Supplemental Material). Authors C.A.S. and N.J.M. independently coded unsatisfactory responses. Codes were added as necessary to identify elements unaccounted for by the rubrics. Various segments of a response, ranging from a phrase, to a sentence or two, to the entire response, could be tagged with either a single or multiple code(s). Codings were discussed until consensus was reached.

Thematic Analysis

Within each IPSA domain, codes were organized by considering both the declarative (content) knowledge and procedural (process) knowledge brought to bear. Because the IPSA was intended to measure the core competency of problem solving, we arranged the codes in a hierarchy consisting of groups of procedural knowledge (i.e., what students were doing), and then specific declarative knowledge codes within each procedural group. This hierarchical structure embraced the interactions thought to occur between procedural and declarative knowledge (Alexander and Judy, 1988). In other words, our thematic grouping of codes was guided by the process of problem solving, and we saw that multiple areas of specific biochemistry content could be applied to a single procedural theme.

RESULTS

IPSA Score Validity

Table 3 summarizes representative student responses across a range of scores, as an indicator that IPSA items engaged students in the domains of problem solving. This sampling is from multiple students. While the selected responses are not comprehensive and wide variability was observed, the prompts for all of the domain items elicited attempts to take appropriate steps within each domain toward solving the problem.

Our claim that IPSA domain scores are distinct from one another was supported by an overall lack of correlation between scores (Table 4). However, in contrast to prior findings (Mitchell *et al.*, 2011), the Investigate domain scores moderately correlated with Evaluate ($r = 0.31$, $p = 0.032$) and Integrate ($r = 0.41$, $p = 0.004$). As expected based on previous work, a moderate correlation was also demonstrated between Evaluate and Integrate domain scores ($r = 0.33$, $p = 0.025$).

Our claim that domain scores are somewhat related to—yet distinct from—scores of content knowledge was also supported by a general absence of correlations. Only the Evaluate domain scores moderately correlated with content exam scores ($r = 0.36$, $p = 0.014$). In our past study, that correlation was strong ($r = 0.53$, $p < 0.02$), and a moderate correlation was found between Integrate domain scores and content exam scores ($r = 0.44$, $p < 0.02$). Differences between findings were likely due to studying different participants (i.e., medical students vs. biochemistry students) and employing different assessments of content knowledge (i.e., the Comprehensive Basic Science Exam vs. the ACS Biochemistry Exam).

Our claim that domain scores are somewhat related to—yet distinct from—research experience was demonstrated by a moderate correlation with only the Investigate domain scores ($r = 0.31$, $p = 0.032$). In other words, students who engaged in two semesters of honors research and presented a thesis just before graduation earned higher scores in the Investigate domain than students without research experience.

TABLE 3. Representative IPSA responses

Domain	Performance level and response
Hypothesize	Unsatisfactory
	"Hypothesis 1: The lorrat has an active metabolism, even when resting. Hypothesis 2: The constant breakdown of fatty acids could contribute to the reduction in adipose tissue. Hypothesis 3: A highly active metabolic state is exothermic, which would keep the lorrat constantly warm."
	Satisfactory
	"Hypothesis 1: High oxygen affinity in lorrat hemoglobin adjusted for elevation. Hypothesis 2: The lorrat could have an overexpressed metabolic enzyme. Hypothesis 3: The lorrat may have a diet high in lipids and carbohydrates. Hypothesis 4: The lorrat lacks certain anabolism enzymatic activity."
Investigate	Unsatisfactory
	"I would use primary cells cultured from the stock lorrat tissue and culture two types of cells. I would use the normal, wild type, cells just as they grow from the little lorrat and then culture a cell knocking out the mechanism to create PEPCK. I would run metabolic analysis experiments on an extracellular flux analyzer (called the Seahorse XF Analyzer). This would show me the difference in both oxygen consumption rate and extracellular acidification rates (ECAR) simultaneously, which is an indirect method of measuring glycolysis. I would expect the PEPCK knockout to have a lower ECAR than the wild type."
	Satisfactory
	"We could look for the RNA corresponding to the PEPCK gene as a marker of upregulation of PEPCK transcription. To do this we could design an RNA segment complementary to the PEPCK mRNA and then attach a fluorescent reporter to this complementary segment. When the complementary segment is bound to the target mRNA the fluorescent reporter will be activated. Testing of several different tissue samples collected from different lorrats as well as the testing of tissue samples from similar species of animals."
Evaluate	Unsatisfactory
	"The aldolase stuff was similar for both the rat and the lorrat, which was expected. The concentration of PEPCK was substantially increased as well as the activity. The Km is roughly the same so it has roughly the same affinity meaning the enzyme is probably not mutated. There could be several reasons for this: the transcription could be increased because a repressor protein is mutated, or an activator is mutated forcing the gene to be on all the time."
	Satisfactory
	"The aldolase in both the lab rat and the lorrat are similar with hardly any change. However, the PEPCK activity and [PEPCK] are doubled while the Km remains the same. This tells me that the lorrat has twice as much PEPCK enzyme thus able to find OAA molecules in the body twice as fast and the PEPCK activity would be able to process OAA twice as much on top of that."
Integrate	Unsatisfactory
	"The results for creatine, glucose, and glycogen metabolites were unremarkable. The results for lactate and TAG's indicate that the lorrat muscle tissue is breaking down the lactate (via PEPCK) and not utilizing fatty acid catabolism via the TCA. The rat is catabolizing fatty acids, and is not breaking down the lactate (the first few steps of gluconeogenesis). It's basically a difference in pathways being used for energy production; the lorrat prefers to use excess lactate to produce PEPCK and glucose through gluconeogenesis, while the rat is breaking down fatty acids to enter into the TCA."
	Satisfactory
	"PEPCK converts oxaloacetate to phosphoenolpyruvate. Phosphoenolpyruvate can then be converted to pyruvate which will be used by the CAC or it can convert to 3-phosphoglycerate which may eventually lead to glucose, glycogen, or triacylglycerols. We see that with an increase in PEPCK activity comes an increase in [triacylglycerol] and a decrease in post-exercise blood [lactate] but no significant increase in [glucose] or [glycogen]. It appears that the increase activity of PEPCK leads to oxaloacetate being converted to phosphoenolpyruvate which is then being converted to 3-phosphoglycerate and then on to dihydroxyacetone phosphate and then triacylglycerols. Instead of making sugars the lorrat is making fat which undergoes oxidation providing energy for the lorrat with less anaerobic metabolism."
Reflect	Unsatisfactory
	"Part 1: Not to my standards. Part 2: I believe Biochemistry 445 and 446 definitely helped me most."
	Satisfactory
	"Part 1: I believe that I was able to provide at least a minimum amount of correct and relevant information in my answers, considering that it has been two years since I have taken a similar exam. Part 2: I would have to say that the extensive education that I received in my biochemistry classes has definitely helped to develop my critical thinking skills, as well as much of the basic and most important topics of biochemistry. Part 3: It reinforced to me that when presented with any unfamiliar circumstance or problem, the key is to not get discouraged, but to take a step back and critically analyze and engage in the situation."

TABLE 4. Correlations at biochemistry program graduation[a]

Score	1	2	3	4	5	6
1. IPSA Hypothesize	1.00					
2. IPSA Investigate	–0.06	1.00				
3. IPSA Evaluate	0.04	**0.31***	1.00			
4. IPSA Integrate	0.10	**0.41****	**0.33***	1.00		
5. IPSA Reflect	0.20	0.21	0.11	0.12	1.00	
6. Content exam	–0.17	0.08	**0.36***	0.26	–0.14	1.00
7. Research experience	0.05	**0.31***	0.23	0.28	–0.07	0.08

[a]Plain text indicates correlations that were not statistically different from zero. Bold indicates moderate correlations ($r \geq 0.3$). $N = 47$.
*$p < 0.5$.
**$p < 0.01$.

Student Performance

Mean IPSA domain scores summarize performance in ill-defined problem solving by biochemistry majors at graduation (Figure 4). Course grades and content exam scores were consistent with historical trends (unpublished data). The average student in this sample performed satisfactorily only in the Evaluate and Reflect domains. Considering the learning objectives addressed in these domains (Table 1), average participants were able to do the following:

- state experimental results when an experimental design and data were provided, and
- critically evaluate their own performance when a conclusion, or final solution to the problem, was provided.

To further probe this performance phenomenon beyond aggregate means, we quantified the prevalence of all possible domain combinations of satisfactory domain performance (Table 5). Graduating biochemistry majors most commonly exhibited only three different patterns. Indeed, the average pattern (#15) occurred in 13% of cases. The other two patterns were similarly prevalent as variations of the average pattern. Scores were frequently either satisfactory in the Integrate domain as well as in Evaluate and Reflect (#26), or they were only satisfactory in the Reflect domain (#6). Taken together, these three patterns accounted for 41% of the students in this sample.

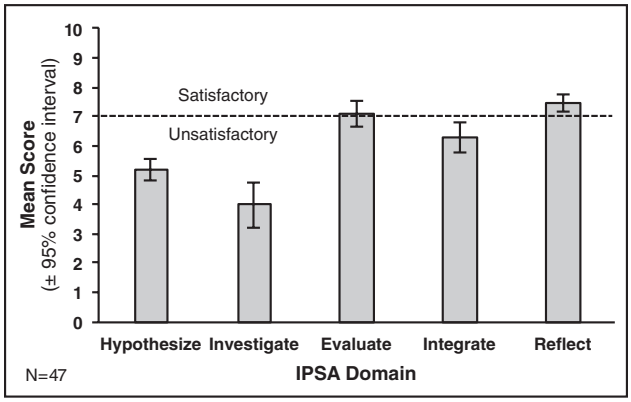

FIGURE 4. IPSA performance. Mean IPSA domain scores with 95% confidence intervals are reported. Scores of seven or greater are considered satisfactory (dashed line).

Considering the other end of the performance spectrum, 6% of students exhibited unsatisfactory performance in all five domains (#1). No graduating biochemistry major was able to achieve satisfactory performance across all domains (#32).

Because successfully solving ill-defined problems requires proficiency in all domains, we do not weight the importance of domains. Regarding general groups of domains, 21% of participants earned satisfactory scores in only one domain, 28% in two domains, 27% in three domains, and 18% in four domains. Overall, this lack of success confirmed the need for deeper understanding of students' solutions.

The Nature of Unsatisfactory Solutions

After quantitatively scoring responses using the rubrics, we anticipated the potential for unsatisfactory responses to contain more unacceptable than acceptable statements. However, analyzing distributions of unacceptable segments within responses revealed two primary types of unsatisfactory responses (Figure 5A). One group—the majority of responses—contained four or fewer statements that were coded as unacceptable. The second group of unsatisfactory responses did, indeed, contain many unacceptable statements. This trend was also apparent within domains (Figure 5, B–F). The following sections examine the responses for each domain in more detail.

Hypothesize Domain

In the Hypothesize domain, most responses (39/47) were unsatisfactory. Content analysis of those responses resulted in 114 coded segments (Supplemental Table S1). Many hypotheses were not mechanistic, failing to explain *how* the observations might have arisen. Nearly a third of the coded segments were hypotheses that the lorrat simply had an increased metabolism. One-fifth of the segments narrowed down hypotheses to a particular area of metabolism (i.e., carbohydrate, citric acid cycle, lipid), yet the mechanism remained vague. Taken together, the unmechanistic hypotheses accounted for 49% of all the coded segments (Table 6). Surprisingly, another 9% of segments were inconsistent with given information.

Teleological thinking was recently characterized in biology by Coley and Tanner (2012) as "causal reasoning based on the assumption of a goal, purpose, or function." Such thinking appeared in 7% of segments (Supplemental Table S1). In these cases, students hypothesized that the observations were somehow due to the lorrat needing to adapt to its environment, use energy efficiently, or proliferate (which are all outcomes, rather than underlying causes or mechanisms).

TABLE 5. Prevalence of IPSA performance patterns

Satisfactory domains		Pattern[a]	Percent of students
None	1	Hypothesize–Investigate–Evaluate–Integrate–Reflect	6
One	2	**Hypothesize**–Investigate–Evaluate–Integrate–Reflect	—
	3	Hypothesize–**Investigate**–Evaluate–Integrate–Reflect	—
	4	Hypothesize–Investigate–**Evaluate**–Integrate–Reflect	2
	5	Hypothesize–Investigate–Evaluate–**Integrate**–Reflect	4
	6	Hypothesize–Investigate–Evaluate–Integrate–**Reflect**	15
Two	7	**Hypothesize–Investigate**–Evaluate–Integrate–Reflect	—
	8	**Hypothesize**–Investigate–**Evaluate**–Integrate–Reflect	—
	9	**Hypothesize**–Investigate–Evaluate–**Integrate**–Reflect	—
	10	**Hypothesize**–Investigate–Evaluate–Integrate–**Reflect**	2
	11	Hypothesize–**Investigate–Evaluate**–Integrate–Reflect	—
	12	Hypothesize–**Investigate**–Evaluate–**Integrate**–Reflect	—
	13	Hypothesize–**Investigate**–Evaluate–Integrate–**Reflect**	—
	14	Hypothesize–Investigate–**Evaluate–Integrate**–Reflect	4
	15	Hypothesize–Investigate–**Evaluate**–Integrate–**Reflect**	13
	16	Hypothesize–Investigate–Evaluate–**Integrate–Reflect**	9
Three	17	**Hypothesize–Investigate–Evaluate**–Integrate–Reflect	—
	18	**Hypothesize–Investigate**–Evaluate–**Integrate**–Reflect	—
	19	**Hypothesize–Investigate**–Evaluate–Integrate–**Reflect**	—
	20	**Hypothesize**–Investigate–**Evaluate–Integrate**–Reflect	—
	21	**Hypothesize**–Investigate–**Evaluate**–Integrate–**Reflect**	4
	22	**Hypothesize**–Investigate–Evaluate–**Integrate–Reflect**	2
	23	Hypothesize–**Investigate–Evaluate–Integrate**–Reflect	4
	24	Hypothesize–**Investigate–Evaluate**–Integrate–**Reflect**	2
	25	Hypothesize–**Investigate**–Evaluate–**Integrate–Reflect**	2
	26	Hypothesize–Investigate–**Evaluate–Integrate–Reflect**	13
Four	27	**Hypothesize–Investigate–Evaluate–Integrate**–Reflect	—
	28	Hypothesize–**Investigate–Evaluate–Integrate–Reflect**	9
	29	**Hypothesize**–Investigate–**Evaluate–Integrate–Reflect**	9
	30	**Hypothesize–Investigate**–Evaluate–**Integrate–Reflect**	—
	31	**Hypothesize–Investigate–Evaluate**–Integrate–**Reflect**	—
All	32	**Hypothesize–Investigate–Evaluate–Integrate–Reflect**	—

[a]Bold domains are those in which satisfactory scores were earned. $N = 47$.

Investigate Domain

The unsatisfactory responses in the Investigate domain (39/47) resulted in 108 coded segments (Supplemental Table S2). While a third of the segments were acceptable statements regarding experimental design, 66% of the segments proposed designs that were not aligned with the given hypothesis (Table 6). Although the task was to investigate possible up-regulation of the PEPCK enzyme at the transcriptional level (i.e., to measure levels of mRNA), nearly all unsatisfactory responses (36/39) proposed one or more methods that were not aligned with the hypothesis (Figure 6A).

Evaluate Domain

Less than half of the Evaluate domain responses were unsatisfactory (19/47; Supplemental Table S3). Uniquely in this domain, many unsatisfactory responses (13/19) included statements that extended into other domains. This accounted for 23% of all coded segments (Table 6). While some responses veered off-track into both the Hypothesize and Integrate domains, most only addressed one of those domains (Figure 6B). Additionally, incorrect statements of results appeared in 20% of segments.

Integrate Domain

Unsatisfactory scores were earned for nearly half of the responses in the Integrate domain (21/47). Content analysis produced 98 coded segments (Supplemental Table S4). Unsubstantiated or incorrect conclusions accounted for 17% of all segments (Table 6). Further correlation analysis revealed that including unsubstantiated or incorrect conclusions within a response correlated moderately and negatively with IPSA scores in this domain ($r = -0.48$, $p < 0.001$). As IPSA scores decreased, it was more likely that an unsubstantiated conclusion was part of the response.

Reflect Domain

According to our scoring rubric (see the Supplemental Material), slightly more than one-fifth of Reflect domain responses were unsatisfactory (10/47). Nearly all 24 segments were acceptable (Supplemental Table S5), yet in those cases, the response did not address all three parts of the prompt. Consequently, 96% of segments were classified as incomplete responses (Table 6). Only one segment was a thoughtless self-assessment, stating that the student "hoped" all the tasks had been met.

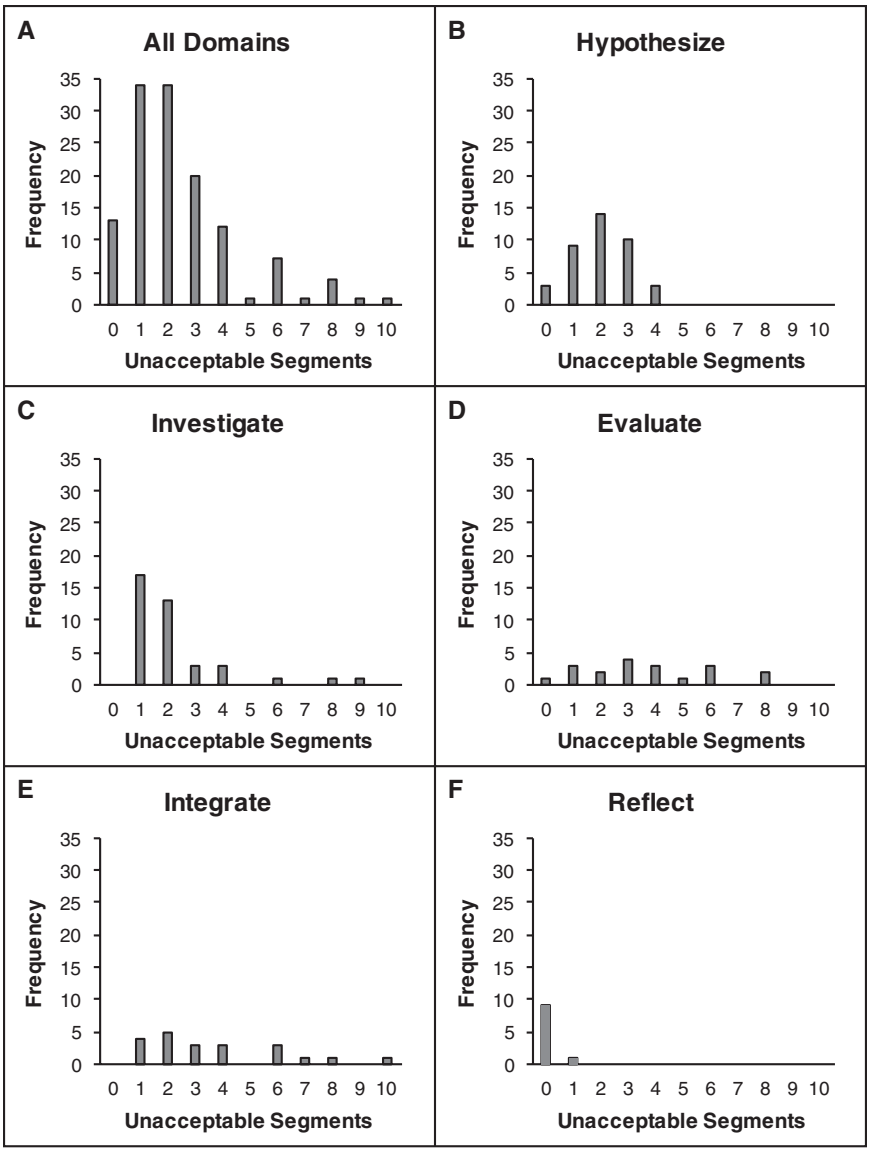

FIGURE 5. Distributions of unacceptable segments in unsatisfactory responses. Histograms show the frequencies of unsatisfactory responses that contained particular numbers of unacceptable segments for all domains combined (A) and by domain (B–F). The number of unsatisfactory responses and unacceptable segments within those responses varied by domain, yet the sample size of all responses was consistent for each domain ($N = 47$).

DISCUSSION

Our overall goal was to promote students' ability to solve ill-defined biochemistry problems. First, we described the performance of biochemistry majors just before graduation, in terms of both average domain scores (Figure 4) and patterns of performance across domains (Table 5). While some students were successful in some domains, the widespread occurrence of unsatisfactory performance indicates the need for developing additional ways to facilitate student construction of procedural knowledge related to scientific problem solving.

One limitation of this retrospective study was that the portion of course points reserved for IPSAs within the two biochemistry courses could not be altered. We suspect that the low-stakes approach (see *Methods*) may not have fully incentivized attainment of the learning objectives. A second limitation imposed restrictions on examining the validity and reliability of IPSA scores. Because the instrument was designed with only one item in each domain, there were no degrees of freedom with which to carry out exploratory or confirmatory factor analyses. Similarly, Cronbach's alpha values, or indicators of internal reliability, could not be computed for domains (which we suspect would be psychometric dimensions), because there were no other items with which consistency of scores could be compared. The retrospective design also precluded determining test–retest reliability.

Given the current lack of validated assessments that measure the multifaceted process of solving ill-defined problems that are specific to any life sciences discipline, it is still valuable to examine IPSA outcomes. Tracking performance patterns is an alternative to analyzing means for targeting and prioritizing domains in which performance is weakest. Our prior efforts elucidated four common patterns of performance among both biochemistry majors as well as medical students (Mitchell *et al.*, 2011). The first two patterns, struggling in the Hypothesize or Investigate domains (regardless of performance in other domains), each remained consistent for 83% of graduating majors. A third pattern, simultaneous difficulty with both the Evaluate and Integrate domains, occurred in 23% of seniors. The previous study identified those domains as moderately correlating with content exam scores (for medical students). Yet in the current study, only the Evaluate domain showed a statistically significant correlation with content knowledge. Using that criterion suggests that 40% of graduating biochemistry majors lacked the declarative knowledge necessary to be successful in the Evaluate domain. The final pattern that emerged from prior work, unsatisfactory scores in the Reflect domain, was demonstrated in our current investigation by only 20% of seniors. This indicates that most program graduates were able to critically evaluate their own IPSA performance. However, according to content analysis of unsatisfactory responses, *accurate* self-evaluations accounted for only 8% of the coded segments (Supplemental Table S5). This finding is consistent with the work of Ziegler and Montplaisir (2014), who showed that undergraduate biology students' perceptions of their own knowledge do not always concur with measurements of that knowledge. While performance in the Reflect domain is one indicator of metacognitive ability, we stress that the Reflect

TABLE 6. Characterization of primary difficulties within unsatisfactory responses

Domain	Procedural knowledge code and example	Percent of coded segments
Hypothesize	Unmechanistic hypotheses	49
	"The lorrat has a high basal metabolic rate compared to other mammals."	
Investigate	Experimental design does not align with hypothesis	66
	"You can also test the enzyme activity using a spectrometry, using coupled enzymatic assay, comparing PEPCK from muscle to PEPCK in other organs."	
	"Since the hypothesis is focusing on the upregulation of PEPCK at the transcription level, I would therefore find it most appropriate to begin investigating the DNA sequence of PEPCK."	
Evaluate	Extending response beyond Evaluate	23
	"The increased quantity of enzyme is responsible for the difference seen in the metabolic pathway of the lorrat."	
	Incorrect results	20
	"Km value for PEPCK inhibition was higher in the lorrat than in the rat."	
Integrate	Unsubstantiated or incorrect conclusions	17
	"Exhibit E shows us how the lorrat is better suited at clearing out lactate build up during exercising."	
Reflect	Incomplete response	96
	"1. I think I came up very short in designing an experiment; I totally got side tracked and over thought it.	
	2. I think the material and the case studies during both 445 and 446 helped me the most on this case." (3. Not addressed)	

domain is not a metacognition inventory and is thus incapable of fully measuring either metacognitive knowledge or metacognitive regulation. Our analysis of performance patterns suggests that the Hypothesize and Investigate domains take immediate priority.

Regarding the qualitative nature of unsatisfactory solutions, we used the model of scientific process defined by Wilson and Rigakos (2016) to generate several explanations for our results. The scientific process consists of overlapping ideas that combine to provide a holistic perspective on procedural knowledge in science: the scientific method, experimental design, and the nature of science. The scientific method does not contain any independent elements; rather, the elements we have defined as domains overlap with either experimental design or with both experimental design and the nature of science. Even apart from metacognitive skills (i.e., the Reflect domain), this suggests that teasing apart the underlying mechanisms that explain performance in domains of the scientific method is a complicated and difficult undertaking.

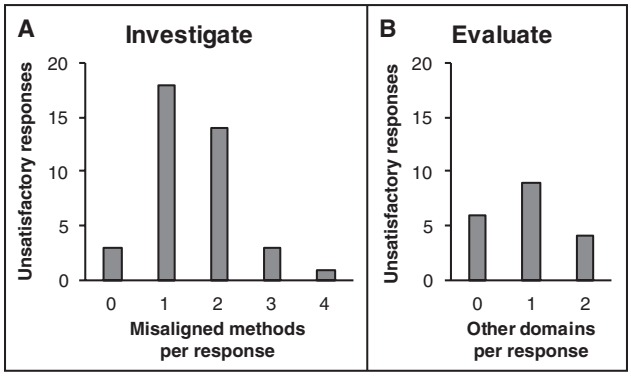

FIGURE 6. Distributions of primary difficulties. Histograms show the frequencies of unsatisfactory responses that contained particular numbers of each type of unacceptable segment. (A) In the Investigate domain, unsatisfactory scores ($n = 39$) primarily stemmed from proposing the use of methods that did not align with the given hypothesis. Some of those responses proposed multiple misaligned methods. (B) In the Evaluate domain, unsatisfactory scores ($n = 19$) commonly resulted from addressing other domains (i.e., Hypothesize, Integrate, or both), at the expense of fully evaluating the given data.

Broadly speaking, communication skills are key to the nature of science and may impact performance, because the IPSA requires written responses. A known limitation of qualitative content analysis is that only what is expressed can be analyzed. Students may, in truth, understand more than they write. The retrospective nature of this study was also a limitation that prevented us from interviewing students to probe their knowledge more deeply. Even so, within unsatisfactory responses, we found that many segments conveyed acceptable ideas (Supplemental Tables S1–S5), and numbers of unacceptable statements within single responses were low (Figure 5). For example, a response could contain only acceptable ideas, yet lack sufficient detail, and thus be scored as unsatisfactory (see Table 3, Evaluate domain). We conclude that one contributor to poor performance is a failure to express the necessary acceptable ideas, rather than revealing a preponderance of unacceptable ideas. This is likely tied to communication skills.

Of the primary difficulties we identified (Table 6), some are consistent with Wilson and Rigakos's model of the scientific process (2016), while others are new aspects to consider. In the Hypothesize domain, our requirement for mechanistic hypotheses stems from the fact that the discipline of biochemistry largely concerns itself with questions of *how* observed phenomena arise. Because the model was developed for use across multiple disciplines, we merely point out that some of the necessary elements of generating hypotheses (e.g., testable ideas) could be insufficient, depending on intended learning outcomes.

In the Investigate domain, we identified a critical component of the scientific process that is not explicitly stated within the model. Alignment of experimental designs with hypotheses is such a fundamental notion, and it was the most impactful upon IPSA scores. We speculate that the reason for its prevalence is due to students inappropriately transferring methods about which they are most knowledgeable to settings where those methods will not be able to provide evidence about the hypothesis. For example, enzyme kinetics assays and protein purification were emphasized in several earlier IPSAs, as well as within a biochemistry laboratory course completed by students in this study. Methods and rationales for quantifying mRNA levels, as "The Lorrat" IPSA required (see the Supplemental Material), were not treated as extensively. Additionally, the language of other misaligned proposals suggested that students were familiar with those methods from other research settings outside biochemistry courses (Table 3).

In the Evaluate domain, when students extended their responses into other domains, this contributed to unsatisfactory scores due to our rubric criterion for remaining focused on stating results (see the Supplemental Material). It was encouraging that students were naturally using the scientific method, immediately interpreting the results they stated, or proposing alternate experiments. Yet the prevalence within unsatisfactory responses indicated that extending responses was clearly at the cost of thoroughly stating the results. Another explanation of unsatisfactory scores in the Evaluate domain is that results were stated incorrectly. Likewise, in the Integrate domain, unsubstantiated conclusions were drawn. Stating results and drawing conclusions are both competencies that are consistent with the model of scientific process. Understanding which parts of such a large model more frequently present difficulties for students has important pedagogical implications.

Now that understanding of procedural knowledge difficulties is beginning to emerge, an important next step is to standardize IPSA prompts and scoring rubrics based on process rather than specific content, so that the same rubric can be applied to any IPSA. This would allow future research on validation by enabling multiple versions to be administered simultaneously within a course. Even more exciting is the potential to compare performance across time, either with or without educational interventions, to further illuminate how to promote the successful solving of ill-defined problems in biochemistry.

PEDAGOGICAL IMPLICATIONS

Despite constructive alignment of learning objectives, assessments, and activities related to scientific problem solving (Table 1), explicit instruction, and repeated practice during two semesters of biochemistry courses, this study reveals that graduating biochemistry majors still struggle to solve ill-defined problems. The process of problem solving includes a range of domains, and the reasons underlying poor performance vary by domain (Table 6). Therefore, we suggest that a multifaceted approach that combines the following strategies may help students realize more gains in solving ill-defined biochemistry problems than were observed in this study.

- *Incentivize learning with a sufficient portion of course points.* In keeping with evidence summarized in a practice-oriented volume by Felder and Brent (2016), 10–20% of assessment points should address higher-order learning objectives (p. 167). The minimum value indicates to students that those learning objectives are important, while the maximum value prevents masking the attainment of foundational objectives.
- *Define performance criteria clearly.* Educators need to be clear about expectations, by making rubric criteria transparent and modeling what those criteria mean. For example, generating hypotheses and then classifying them as either mechanistic or unmechanistic could improve performance in the Hypothesize domain. For the Evaluate domain, help students focus on ensuring that results statements are complete and accurate, rather than spending their time extending responses into other domains. Likewise, for the Integrate domain, develop activities that explicitly require students to state conclusions in terms that relate all the results to the observed phenomenon. Rubric criteria can also be used as a springboard to promote scientific writing skills by organizing thoughts in a stepwise manner.
- *Facilitate scientific communication skills.* Because the IPSA is a written assessment, clarity and completeness of ideas are crucial to satisfactory performance. As just mentioned, rubric criteria can guide determinations of which ideas to express in each domain. Graduating biochemistry majors who performed well demonstrated understanding of the importance of organization when communicating scientific thoughts. Although students in this study were familiar with the IPSA format, they were much more experienced with the objective and short-answer assessments seen throughout their educational training. Success could be achieved on those assessments by using key words and phrases. Yet that communication style is incongruent with the nature of science. Just as pieces of conceptual knowledge must be connected when learning a discipline, words and phrases must be connected in writing to clearly and thoroughly express solutions to problems. Many active, student-centered learning techniques are amenable to facilitating scientific writing skills, from minute papers and jigsaws to reflection journals and larger writing projects. Peer-review exercises could also be incorporated with any of these formats. At the time of this retrospective study, our biochemistry courses had not yet been transformed to a student-centered focus. It would be interesting to study dosage effects of activities that are designed to enhance communication skills, to determine whether additional practice improves performance in problem solving, and if so, how much practice is necessary.
- *Facilitate alignment between hypotheses and experimental designs.* This study indicates that the most troublesome aspect of the entire process of solving ill-defined biochemistry problems was understanding the kind of evidence that would be necessary to appropriately address a given hypothesis (Table 6). Students became entrenched in familiar experimental designs, regardless of whether the results would yield fruitful information. We urge educators to take an approach that draws explicit connections between hypotheses and investigations. Assure students that they are

not expected to demonstrate methodological expertise (i.e., which buffer and how many microliters to use). Instead, emphasize development of the reasoning behind measuring what should be measured, while identifying appropriate controls and variables.

Research experience is known to aid the development of scientific process skills (Elgin *et al.*, 2016, and references therein). Indeed, we found a moderate and positive relationship between participating in honors research and scores in the IPSA Investigate domain, but not in other domains (Table 4). This is likely due to the experience being heavily mentored and directed. While honors research students worked on a project across two semesters, wrote a thesis, and gave a formal presentation, the experience did not require development of a hypothesis. By contrast, the IPSAs started with observations and generating hypotheses. Given the wide variability of research experiences across institutions, it cannot be assumed that the experience provides practice aligning hypotheses and experimental designs. Students also inappropriately transferred methods from laboratory courses to the IPSAs, simply because superficial features of the problem were similar (Table 6 and Supplemental Table S2). Taken together, these results imply that, while research experience is valuable to development of many scientific-thinking processes, such experience may not directly support skills measured by the IPSA.

- *Enhance feedback on problem solving.* During group OLC discussions, students were continuously monitored, and cases were reviewed during class (Anderson *et al.*, 2008). After individual assessment with the IPSAs, we used radar diagrams to visually represent scores, but only when students specifically requested assistance (Mitchell *et al.*, 2011). It might be more widely beneficial to automate this type of output and provide it along with scores to all students. Of course, students would need to be trained upfront on interpreting the diagrams. Rubric transparency is another way to enhance feedback, as discussed earlier. When students are armed with explicit criteria for performing well, they know exactly what their scores mean, and where they need to improve. Standardizing the rubrics so that they relate only to problem solving and can be applied to any IPSA (see *Discussion*) will also enhance the feedback process. Finally, gathering students' perspectives about feedback would further empower them during learning. For example, anonymously poll students to determine whether various forms of feedback were helpful and to elicit suggestions for other feedback.

ACKNOWLEDGMENTS

We thank the Crosslake Learning Resource Center (CLRC) for the use of previously developed IPSAs and administration software. This paper was improved based on thoughtful feedback from anonymous reviewers and from biology education research colleagues at the University of Georgia, especially Peggy Brickman and Tessa Andrews. We are also grateful to Paula Lemons for her advice about qualitative aspects of this study. This work was supported in part by the National Science Foundation under Grant No. DUE 1043079.

REFERENCES

Alexander, P. A., & Judy, J. E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research*, *58*(4), 375–404.

Alexander, P. A., Pate, P. E., Kulikowich, J. M., Farrell, D. M., & Wright, N. L. (1989). Domain-specific and strategic knowledge: Effects of training on students of differing ages or competence levels. *Learning and Individual Differences*, *1*(3), 283–325.

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC. Retrieved November 15, 2017, from www.visionandchange.org/

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

Anderson, L. W., & Krathwohl, D. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

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 online discussions. *CBE—Life Sciences Education*, *7*(2), 254–262. doi: 10.1187/cbe.07-06-0037

Biggs, J. (1999). *Teaching for quality learning at university: What the student does* (1st ed.). Buckingham, UK: Society for Research into Higher Education and Open University Press.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1954). *Taxonomy of educational objectives*. New York: Addison Wesley.

Bodner, G. M. (1986). Constructivism: A theory of knowledge. *Journal of Chemical Education*, *63*(10), 873–878.

Bodner, G. M. (2015). Research on problem solving in chemistry. In Garcia-Martinez, J., & Serrano-Torregrosa, E. (Eds.), *Chemistry education: Best practices, opportunities and trends* (pp. 181–201). Weinheim, Germany: Wiley-VCH.

Bodner, G., Klobuchar, M., & Geelan, D. (2001). The many forms of constructivism. *Journal of Chemical Education*, *78,* 1107.

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

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5,* 121–152.

Coley, J. D., & Tanner, K. D. (2012). Common origins of diverse misconceptions: Cognitive principles and the development of biology thinking. *CBE—Life Sciences Education*, *11*(3), 209–215. doi: 10.1187/cbe.12-06-0074

Cross, D. R., & Paris, S. G. (1988). Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology*, *80*(2), 131–142.

Dye, K. M., & Stanton, J. D. (2017). Metacognition in upper-division biology students: Awareness does not always lead to control. *CBE—Life Sciences Education*, *16*(2), ar31. doi: 10.1187/cbe.16-09-0286

Elgin, S. C., Bangera, G., Decatur, S. M., Dolan, E. L., Guertin, L., Newstetter, W. C., … Labov, J. B. (2016). Insights from a convocation: Integrating discovery-based research into the undergraduate curriculum. *CBE—Life Sciences Education*, *15*(2), fe2. doi: 10.1187/cbe.16-03-0118

Felder, R. M., & Brent, R. (2016). *Teaching and learning STEM: A practical guide*. San Francisco: Jossey-Bass.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences USA*, *111*(23), 8410–8415. doi: 10.1073/pnas.1319030111

Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., DeHaan, R., … Wood, W. B. (2004). Scientific teaching. *Science*, *304,* 521–522.

Handelsman, J., Miller, S., & Pfund, C. (2006). *Scientific teaching*. New York: Freeman.

Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field Methods*, *18*(1), 3–20. doi: 10.1177/1525822x05282260

Maher, J. M., Markey, J. C., & Ebert-May, D. (2013). The other half of the story: Effect size analysis in quantitative research. *CBE—Life Sciences Education*, *12*(3), 345–351. doi: 10.1187/cbe.13-04-0082

Mitchell, S. M., Anderson, W. L., Sensibaugh, C. A., & Osgood, M. P. (2011). What really matters: Assessing individual problem-solving performance in the context of biological sciences. *International Journal for the Scholarship of Teaching and Learning*, *5*(1), ar17.

Murphy, P. K., & Alexander, P. A. (2002). What counts? The predictive powers of subject-matter knowledge, strategic processing, and interest in domain-specific performance. *Journal of Experimental Education*, *70*(3), 197–214.

National Research Council. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. Singer, S. R., Nielsen, N. R., & Schweingruber, H. A., (Eds.). Washington, DC: National Academies Press.

Nehm, R. H., & Ridgway, J. (2011). What do experts and novices "see" in evolutionary problems? *Evolution: Education and Outreach*, *4*(4), 666–679. doi: 10.1007/s12052-011-0369-7

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Patton, M. Q. (2015). *Qualitative research and evaluation methods: Integrating theory and practice* (4th ed.). Los Angeles: Sage.

Prevost, L. B., & Lemons, P. P. (2016). Step by step: Biology undergraduates' problem-solving procedures during multiple-choice assessment. *CBE—Life Sciences Education*, *15*(4), ar71. doi: 10.1187/cbe.15-12-0255

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. *CBE—Life Sciences Education*, *15*(1), rm1. doi: 10.1187/cbe.15-08-0183

Rickey, D., & Stacy, A. M. (2000). The role of metacognition in learning chemistry. *Journal of Chemical Education*, *77*(7), 915–920.

Sandi-Urena, S., Cooper, M. M., & Stevens, R. H. (2011). Enhancement of metacognition use and awareness by means of a collaborative intervention. *International Journal of Science Education*, *33*(3), 323–340. doi: 10.1080/09500690903452922

Smith, M. U. (1988). *Toward a unified theory of problem solving: Views from the content domains*. Hillsdale, NJ: Erlbaum.

Smith, M. U., & Good, R. (1984). Problem solving and classical genetics: Successful versus unsuccessful performance. *Journal of Research in Science Teaching*, *21*(9), 895–912.

Stanton, J. D., Neider, X. N., Gallegos, I. J., & Clark, N. C. (2015). Differences in metacognitive regulation in introductory biology students: When prompts are not enough. *CBE—Life Sciences Education*, *14*(2), ar15. doi: 10.1187/cbe.14-08-0135

Warfa, A. M. (2016). Mixed-methods design in biology education research: Approach and uses. *CBE—Life Sciences Education*, *15*(4), rm5. doi: 10.1187/cbe.16-01-0022

Wilson, K. J., & Rigakos, B. (2016). Scientific process flowchart assessment (SPFA): A method for evaluating changes in understanding and visualization of the scientific process in a multidisciplinary student population. *CBE—Life Sciences Education*, *15*(4), ar63. doi: 10.1187/cbe.15-10-0212

Ziegler, B., & Montplaisir, L. (2014). Student perceived and determined knowledge of biology concepts in an upper-level biology course. *CBE—Life Sciences Education*, *13*(2), 322–330. doi: 10.1187/cbe.13-09-0175