CBE—Life Sciences Education
Vol. 13, 265–284, Summer 2014

Article

Development and Validation of a Rubric for Diagnosing Students' Experimental Design Knowledge and Difficulties

Annwesa P. Dasgupta,[*] Trevor R. Anderson,[†] and Nancy Pelaez[*]

[*]Department of Biological Sciences and [†]Divisions of Chemical Education and Biochemistry, Department of Chemistry, Purdue University, West Lafayette, IN 47907

Submitted September 24, 2013; Revised January 24, 2014; Accepted January 25, 2014
Monitoring Editor: Erin Dolan

It is essential to teach students about experimental design, as this facilitates their deeper understanding of how most biological knowledge was generated and gives them tools to perform their own investigations. Despite the importance of this area, surprisingly little is known about what students actually learn from designing biological experiments. In this paper, we describe a rubric for experimental design (RED) that can be used to measure knowledge of and diagnose difficulties with experimental design. The development and validation of the RED was informed by a literature review and empirical analysis of undergraduate biology students' responses to three published assessments. Five areas of difficulty with experimental design were identified: the variable properties of an experimental subject; the manipulated variables; measurement of outcomes; accounting for variability; and the scope of inference appropriate for experimental findings. Our findings revealed that some difficulties, documented some 50 yr ago, still exist among our undergraduate students, while others remain poorly investigated. The RED shows great promise for diagnosing students' experimental design knowledge in lecture settings, laboratory courses, research internships, and course-based undergraduate research experiences. It also shows potential for guiding the development and selection of assessment and instructional activities that foster experimental design.

INTRODUCTION

Undergraduate students are becoming increasingly engaged in biology research to meet more rigorous academic criteria, to gain a competitive employment edge upon graduation, or for various other reasons (Lopatto, 2003, 2008; Laursen *et al.*, 2010; Wei and Woodin, 2011). With many physical science and engineering subdisciplines focusing increasingly on problems related to living organisms, it is not surprising that more and more undergraduates are becoming engaged in biology research. Without biology experiments, there would be no way of investigating the nature of mechanisms in living systems; for example, how a firefly glows and how cells "know" when to divide. Designing experiments involves framing research questions to investigate observations; defining and understanding measurable variables; and processing, visualizing, and interpreting results.

Despite the obvious importance of experimental knowledge and numerous calls to involve undergraduate students in authentic research experiences (Wei and Woodin, 2011), surprisingly little is known about what students actually learn from designing experiments for biological research. What has been established, though, is that experimental design is challenging for many students from elementary school to the undergraduate level (Burns *et al.*, 1985; Bullock and Ziegler, 1999; Chen and Klahr, 1999; Fuller, 2002; Kuhn and Dean, 2005; Shi *et al.*, 2011; Sirum and Humburg, 2011). There is, therefore, increasing interest in helping biology students learn about the experimental research process in general, as supported by recommendations expressed in several recent reports (National Research Council, 2007; Association of American Medical Colleges and Howard Hughes Medical Institute, 2009; American Association for the Advancement of Science [AAAS], 2010; Association of American Colleges and Universities, 2013). These reports clearly emphasize

DOI: 10.1187/cbe.13-09-0192
Address correspondence to: Nancy Pelaez (npelaez@purdue.edu).

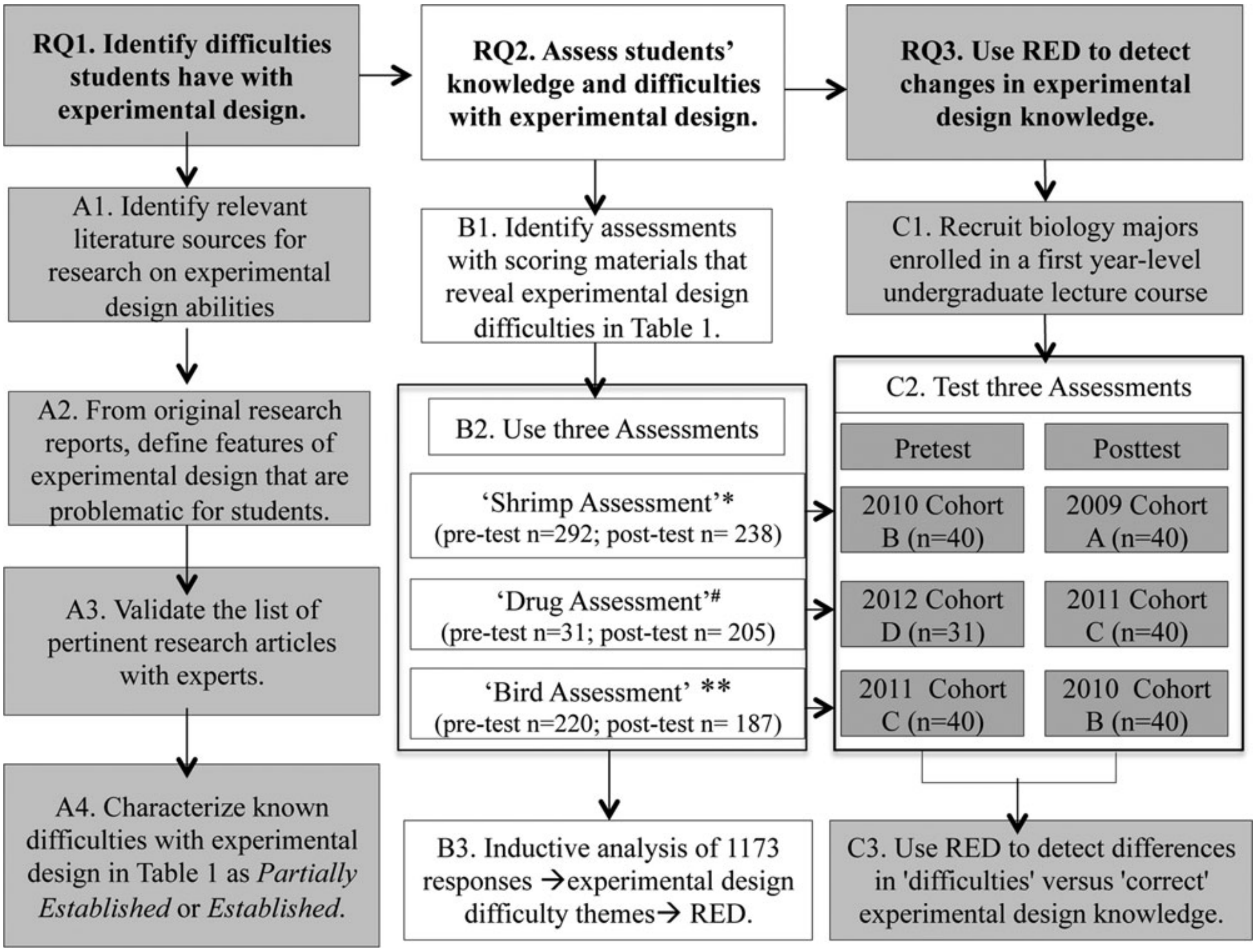
RQ1. Identify difficulties students have with experimental design.
RQ2. Assess students' knowledge and difficulties with experimental design.
RQ3. Use RED to detect changes in experimental design knowledge.
A1. Identify relevant literature sources for research on experimental design abilities
A2. From original research reports, define features of experimental design that are problematic for students.
A3. Validate the list of pertinent research articles with experts.
A4. Characterize known difficulties with experimental design in Table 1 as Partially Established or Established.
B1. Identify assessments with scoring materials that reveal experimental design difficulties in Table 1.
B2. Use three Assessments
'Shrimp Assessment'* (pre-test n=292; post-test n= 238)
'Drug Assessment'# (pre-test n=31; post-test n= 205)
'Bird Assessment' ** (pre-test n=220; post-test n= 187)
B3. Inductive analysis of 1173 responses → experimental design difficulty themes → RED.
C1. Recruit biology majors enrolled in a first year-level undergraduate lecture course
C2. Test three Assessments
Pretest
Posttest
2010 Cohort B (n=40)
2009 Cohort A (n=40)
2012 Cohort D (n=31)
2011 Cohort C (n=40)
2011 Cohort C (n=40)
2010 Cohort B (n=40)
C3. Use RED to detect differences in 'difficulties' versus 'correct' experimental design knowledge.

Figure 1. The process for developing and validating the RED involved (A) a systematic review of the literature to identify experimental design difficulties documented by research, (B) testing three published assessments by looking at more than 1100 responses to see how well they probe for difficulties consistent with research on experimental design difficulties from the literature, and (C) recruiting four cohorts of students to take the assessments to develop a RED based on their responses to published assessments collected before and after an introductory biology course. The assessments are used with permission from: #, SRI International (2003) and the College Board (*, 2006; **, 2009).

"experimental design" as a core scientific ability. But what does it mean to acquire knowledge about experiments? How can we best determine whether students are learning about experimental design and what difficulties they might be encountering?

It is important that all undergraduate biology students experience the process of biological research as a key component of their biology curricula. This is strongly supported by a wide range of studies in the literature reporting numerous benefits to students from doing research, including a more positive attitude toward research and plans for postgraduate education in the sciences (AAAS, 2010). Most of the studies rely on rubrics (Dolan and Grady, 2010; Feldon *et al*., 2010; Timmerman *et al*., 2011), surveys (Kardash, 2000; Lopatto, 2004, 2007; Laursen *et al*., 2010; Thiry *et al*., 2012; Kloser *et al*., 2013), and interviews (Gutwill-Wise, 2001; Thiry *et al*., 2012) to evaluate student learning about research. However, few of these directly measure what undergraduate students actually learned from such research experiences. There is, therefore, a gap in our knowledge in this area. In this paper, we propose to address this gap through the development of a rubric for experimental design (RED) that can be used to diagnose undergraduate biology students' experimental design knowledge and difficulties. Toward achieving this goal, we addressed the following three research questions:

1. What types of difficulties do students have with experimental design?
2. To what extent do published assessments reveal evidence of first-year undergraduate biology students' knowledge and difficulties with experimental design?
3. Can a RED be usefully deployed to detect changes in undergraduate students' experimental design knowledge during a first-year biology course?

An overview of the research process deployed for developing and validating the RED is given in Figure 1. To address research question 1 (RQ1), we performed a multistep literature review (Figure 1A) to identify, characterize, and classify known experimental design difficulties. To address research question 2 (RQ2), we deployed a process (Figure 1B) that identified three published assessment instruments, which were tested for their ability to detect difficulties in first-year undergraduate biology students. Data from addressing RQ1 and RQ2, namely published data about difficulties from the literature as well as data from student responses to the three published assessment instruments, were used to inform the development of the RED. The RED was then tested in a pre/posttest experimental design (Figure 1C) to address research question 3 (RQ3).

LITERATURE REVIEW

To learn about the difficulties undergraduate biology students have with experimental design (RQ1), as per Figure 1A, our first step was to review the literature. This would also enable us to define the abilities necessary for competent experimental design, including identifying a problem; generating hypotheses; planning experimental procedures with treatment, control, and outcome variables; and interpreting findings to make inferences (AAAS, 2010). For the literature review, we first tracked down original research from two reports from the National Academies (Singer *et al.*, 2006; Duschl *et al.*, 2007). This helped us to identify key peer-reviewed journals from disciplines ranging from psychology and cognition to discipline-based education research journals, including those used by cell biologists, physiologists, and ecologists. Original research on difficulties was also found in articles from peer-reviewed journals in the areas of teacher education and undergraduate education (e.g., *Journal of College Science Teaching* and *American Biology Teacher*) and in dissertations. We did not use any secondary sources, except to identify references to primary sources we might have missed. Although our main interest is in undergraduate difficulties, we included studies from child development, due to the possibility that our undergraduate students might still demonstrate difficulties documented by research studies on experimental design abilities with children. Within each area, we identified research articles that address student difficulties or abilities related to one or more aspect of experimental design. This process helped us compile an initial list of findings from research, which was reviewed by a scientist, a cognitive scientist, and a science teacher educator, and checked against references presented at a symposium on psychological sciences, Psychology of Science: Implicit and Explicit Processes (Purdue University, 2010).

Some difficulties with experimental design had rich descriptions and solid evidence, while we found limited evidence for others. For this research study, we elaborated on Grayson *et al.*'s (2001) framework to characterize and classify these experimental design difficulties as follows (Figure 1A4). Difficulties were classified as *established* if they met the following criteria: 1) identified in at least three studies, 2) were found in two or more different populations, 3) showed evidence that the difficulty was more than just the direct result of a single assessment, and 4) appeared with reasonable prevalence in data that supported a stable description of the difficulty. In contrast, difficulties were classified as *partially established* if they had been: 1) documented only in one or two studies and 2) could have been the result of a single assessment or the way those students were taught. With limited evidence, a partially established difficulty merits further research. But with increasing triangulation of data and multiple observations in different contexts, it was determined that the identified difficulty was an authentic part of student thinking rather than a function of how a particular textbook presented material, how a particular teacher taught, or the nature of a particular question. By classifying the difficulties in this manner, we would know which partially established and established difficulties we could confidently use to inform the development of the rubric. Any remediation of such difficulties would, therefore, be based on sound knowledge of the nature of the difficulty. Of course, some of the difficulties were later classified at a higher level based on our own data generated while addressing RQ1.

As summarized in Table 1, we found that most of the reported difficulties with experimental design could be classified as established, while only a few met our criteria of partially established, due to limited evidence. The difficulties we found fell into five categories as listed in Table 1: the experimental subject itself (difficulty I), variables (difficulty II, A–F), measures of experimental outcomes (difficulty III), dealing with variability (difficulty IV, A–E), and interpreting experimental conclusions (difficulty V, A–B). As shown in Table 1, difficulties were found across different populations of students at multiple educational levels, including elementary, middle, and high school students, undergraduates who were not science majors, and undergraduate science students.

A surprising finding by Salangam (2007) is that some students do not know how to identify the experimental subject (difficulty I). This difficulty is classified as partially established, because it was found in only one quasi-experimental study with undergraduate students who were not science majors. Further research is needed to establish to what extent this difficulty is found across different populations of students.

Thinking about and working with different variables presents students with a variety of difficulties (Table 1, difficulty II, A–F). Elementary school students are known to struggle with experimental controls, and they are more competent in recognizing than designing such controls (Bullock and Ziegler, 1999). Manipulation of experimental variables is difficult for middle and high school students. This fact has been known for 50 yr, since Karplus first demonstrated that students have problems with formal operational reasoning patterns like combinatorial reasoning, or the simultaneous manipulation of two independent variables in a study (Fuller, 2002). Middle and high school students also have trouble identifying treatment, outcome, and control variables (Burns *et al.*, 1985; Dolan and Grady, 2010). Gormally *et al.* (2012) recently reported that biology undergraduate students in a general education course still have difficulties with quantitative variables. Another problem undergraduate students have with treatment and outcome variables is inappropriately associating these variables in constructing a testable hypothesis (Griffith, 2007; Salangam, 2007; Harker, 2009; Beck and Blumer, 2012; Libarkin and Ording, 2012; D'Costa and Schlueter, 2013). These problems, associating treatment and outcome variables, have also been reported among undergraduates outside the biology major, for example, in psychology (Koehler, 1994). Even undergraduate biology majors have trouble understanding quantitative variable concepts such as probability distributions, statistical *p* values,

Table 1. Experimental design difficulties classified on the four-level framework and how they relate to what the three published assessments measure

				Published assessments[d]		
	Difficulty[a]	Level[b]	Demographic population[c]	Shrimp	Drug	Bird
I	Identifying the experimental subject (Salangam, 2007)	Partially established	UN	x	x	x
II.	Variables: a variable property of an experimental subject					
A	Categorical (discrete) variable (Picone *et al.*, 2007)	Partially established	UN			
B	Quantitative (continuous) variable (Colon-Berlingeri and Burrowes, 2011; Gormally *et al.*, 2012; Harker, 2009; Hiebert, 2007; Picone *et al.*, 2007)	Established	UB			
C	Treatment (independent) variable (Beck and Blumer, 2012; Burns *et al.*, 1985; D'Costa and Schlueter, 2013; Dolan and Grady, 2010; Griffith, 2007; Harker, 2009; Hiebert, 2007; Koehler, 1994; Libarkin and Ording, 2012; Picone *et al.*, 2007; Salangam, 2007; Tobin and Capie, 1982)	Established	MS, HS, UN, UB	x	x	x
D	Outcome (dependent) variable (Beck and Blumer, 2012; Burns *et al.*, 1985; D'Costa and Schlueter, 2013; Dolan and Grady, 2010; Griffith, 2007; Harker, 2009; Koehler, 1994; Libarkin and Ording, 2012; Picone *et al.*, 2007; Salangam, 2007; Tobin and Capie, 1982)	Established	MS, UN, UB	x	x	
E	Control (comparison) group (Bullock and Ziegler, 1999; D'Costa and Schlueter, 2013; Dolan and Grady, 2010; Gormally *et al.*, 2012; Harker, 2009; Hiebert, 2007; Shi *et al.*, 2011)	Established	ES, MS, U		x	
F	Combinatorial reasoning (Karplus by Fuller, 2002; Lawson and Snitgen, 1982; Lawson *et al.*, 2000; Tobin and Capie, 1981)	Established	MS, HS, U	x	x	x
III	Measurement of results (Dolan and Grady, 2010; Harker, 2009; Hiebert, 2007; Salangam, 2007; Tobin and Capie, 1982)	Established	MS, UB	x	x	x
IV.	How to deal with variability					
A	Recognition of natural variation within a biological sample (Kanari and Millar, 2004; Picone *et al.*, 2007)	Established	MS, UB		x	
B	Random (representative) sample (Colon-Berlingeri and Burrowes, 2011; Gormally *et al.*, 2012; Metz, 2008)	Established	UB		x	
C	Randomization of treatments (Colon-Berlingeri and Burrowes, 2011; Gormally *et al.*, 2012; Hiebert, 2007)	Established	UB	x	x	x
D	Replication of treatments (Harker, 2009; Kanari and Millar, 2004)	Established	MS, UB	x	x	x
E	Reducing effect of unrelated variables (Chen and Klahr, 1999; D'Costa and Schlueter, 2013; Kuhn and Dean, 2005; Tobin and Capie, 1982)	Established	ES, MS, UB	x	x	x
V.	Interpretation of experimental conclusions					
A	Scope of inference/generalizability of results (Chen and Klahr, 1999; Colon-Berlingeri and Burrowes, 2011; Lawson *et al.*, 2000; Metz, 2008; Tobin and Capie, 1982)	Established	ES, MS, U	x	x	x
B	Cause and effect conclusions (Dolan and Grady, 2010; Griffith, 2007; Gormally *et al.*, 2012; Grunwald and Hartman, 2010; Harker, 2009; Hiebert, 2007; Klahr *et al.*, 1993; Kuhn and Pearsall 2000; Kuhn *et al.*, 1992; Libarkin and Ording, 2012; Metz, 2008; Park and Pak, 1997; Roth *et al.*, 1998; Schauble, 1990, 1996; Schauble and Glaser, 1990)	Established	ES, MS, U	x	x	

[a] A review of the literature revealed that student difficulties with experimental design knowledge could be organized into five categories I–V. For definitions of the terms under I–V refer to the glossary of terms in the Supplemental Material (p. 20).

[b] Based on the four-level framework (Grayson *et al.*, 2001), "Level" refers to how much insight there is about a particular difficulty. Difficulties found across different populations of students at multiple educational levels are classified as *established;* others that require further research are classified as *partially established*.

[c] U: undergraduate students; UN: undergraduate science nonmajors; UB: undergraduate biology students; ES: elementary school students; MS: middle school students; HS: high school students.

[d] x represents cases in which scoring materials from the publishers claim the assessment measures knowledge consistent with the difficulty documented by past research.

and regression analysis (Hiebert, 2007; Harker, 2009; Colon-Berlingeri and Burrowes, 2011). They also have problems creating graphs from raw quantitative data (Picone *et al*., 2007), and with treatment and outcome (Picone *et al*., 2007; D'Costa and Schlueter, 2013) and control variables (Hiebert, 2007; Harker, 2009; Shi *et al*., 2011; D'Costa and Schlueter, 2013). While we classified these as established difficulties, we found only one study that exposed difficulties science nonmajors' have graphically representing categorical variable data (Table 1, difficulty IIA). This single report about categorical variable difficulties (Picone *et al*., 2007) was classified as partially established, because further investigations are required to determine whether the difficulty is limited to graphs or whether students also struggle with the concept of categorical variables in general. Moreover, research is needed to test for this difficulty with other relevant populations, such as biology majors.

Several studies have established that, from middle school to biology undergraduate levels, students often fail to state their findings accurately in a way that relates to the actual measures used in an experiment (difficulty III). Making decisions about what variables to measure at various stages of an experiment is also poorly understood by many students (Tobin and Capie, 1982; Hiebert, 2007; Harker, 2009; Dolan and Grady, 2010). Biology students who are not science majors have difficulty distinguishing between the relevant and unrelated variables that they need to measure to address a given experimental goal (Salangam, 2007).

Student difficulties with natural variability have been well documented in multiple studies examining students doing experiments (Table 1, difficulty IV). For example, some elementary and middle grade students do not understand how variability might be controlled by reducing effects of unrelated variables (difficulty IVE; Chen and Klahr, 1999; Kuhn and Dean, 2005), while middle school students have trouble interpreting findings when faced with natural variation (difficulty IVA; Kanari and Millar, 2004). Dealing with natural variation (difficulty IVA) is also a difficult task for undergraduate biology majors and nonmajors (Picone *et al*., 2007). Biology students have difficulty reducing the effect of unrelated variables in their experiments (difficulty IVE; D'Costa and Schlueter, 2013). Few undergraduate students know that random assignment of treatments to samples of experimental subjects (difficulty IVC) provides a way to measure and minimize the effect of natural variation in samples (Hiebert, 2007). Studies show that some middle school students fail to see the need to replicate treatments as a way to deal with variability (difficulty IVD) (Kanari and Millar, 2004), while biology undergraduates show a similar problem (Harker, 2009). Undergraduate biology students also have trouble with randomization of treatments (difficulty IVC) and the idea of having a representative sample of experimental subjects (difficulty IVB; Gormally *et al*., 2012). Colon-Berlingeri and Burrowes (2011) and Metz (2008) demonstrated that biology undergraduates have difficulty summarizing trends from data with probability distributions and fail to use distributions to provide information about variation and representativeness of an experimental sample (difficulty IVB). In summary, students of all ages clearly struggle to deal with variability in an experiment.

Problems with interpreting experimental findings are another well-documented difficulty. Students from elementary school (Chen and Klahr, 1999), middle school (Tobin and Capie, 1982), and undergraduate levels (Tobin and Capie, 1981; Lawson *et al*., 2000) struggle with estimating the extent of inferences made from experimental findings (Table 1, difficulty V). Another extensively reported issue (difficulty V B) is making claims about cause-and-effect relationships in experiments. This problem is prevalent among students from the elementary school to the undergraduate level (Schauble, 1996; Libarkin and Ording, 2012).

It is surprising to note that experimental design difficulties have met our established or partially established criteria as long as 50 yr ago, and yet these difficulties persist with a range of students from elementary school to undergraduate levels. Undergraduate biology instructors may be unaware that these well-documented difficulties may be a challenge for their own students. Using the previously identified difficulties, we set out to find tools for diagnosing these problems in our own undergraduate biology students, because without explicit information about students' problems, we would not be able to intervene with appropriate guidance.

METHODS

Study Design

Four cohorts of ~300 undergraduate biology majors participated in the study at a research university in the Midwest region of the United States, across four semesters in three consecutive years (2009–2012). These students were enrolled in a first year–level lecture course, Development, Structure, and Function of Organisms. As described by Clase *et al.* (2010), according to the expected outcomes for this course, students would learn about development, structure, and function of organisms based on information from biological research such as experiments.

Many published assessment instruments for experimental design were tested, of which three were selected, based on the claims of the authors (SRI International, 2003; College Board, 2006, 2009) that the assessment instruments probe the difficulties consistent with previous literature (see Figure 1). These three were used as pre- and posttests with our undergraduate biology student sample (Figure 1B) at the beginning and end of the semester during three consecutive years (Figure 1C). All assessments had been professionally validated (SRI International, 2003; College Board, 2006, 2009) for use with high school students as measures for experimental design knowledge in areas I–V (Table 1). As a result of using each assessment with two different cohorts, we developed the RED to summarize areas in which students consistently demonstrate difficulties with experimental design. Thus, this study examined whether these assessments also provide useful diagnostic information about college students.

Addressing RQ1: What Types of Difficulties Do Undergraduate Biology Students Have with Experimental Design?

This question was addressed under the above literature review section. Studies of experimental design difficulties with children were included, because the same types of difficulties were also reported in studies with undergraduate students (Table 1).

Addressing RQ2: To What Extent Do Published Assessments Reveal Evidence of First-Year Undergraduate Biology Students' Knowledge and Difficulties with Experimental Design?

Motivation for Selection of Assessments. For this study, three published assessments were used as diagnostic questions. With a list of important experimental design difficulties as the target (Table 1), the first criterion for selecting such assessments was whether publishers claim that a test probes for the difficulties documented in the literature. The published assessments that probe for experimental knowledge relevant to each category of difficulty (Table 1, I–V) used in this study will be referred to as the *shrimp*, the *drug*, and the *bird* assessments, published by the College Board (2006), SRI International (2003), and the College Board (2009), respectively (Figure 1).

For the shrimp assessment, students had to propose an experiment to combine nutrients and salt levels to find their effect on the growth of tiger shrimp. The drug assessment asked students to design an experiment with appropriate patients to test a new drug for reducing high blood pressure. The bird assessment was framed around the design of an experiment to treat pesticide granules with two different colors and patterns to learn which of the two treatments the various bird species (blackbirds, zebra finches, and geese) will avoid eating and whether there is a difference for males and females. The actual probes and scoring guidelines are included with permission and a URL for the original source of each assessment as Supplemental Material. In the *Results*, we compare features of experimental design probed by each assessment to the difficulties identified from a review of the literature (Table 1).

The Shrimp Assessment. According to the published source, an assessment from the 2006 College Board AP Statistics test (henceforth shrimp assessment) is useful for evaluating abilities to: "(1) identify the treatments in a biological experiment; (2) present a completely randomized design with replications to address the research question of interest; (3) describe the benefit of limiting sources of variability; and (4) describe the limitations to the scope of inference for the biologist" (College Board, 2006, Scoring Guidelines, p. 16). As per Table 1, this assessment measures knowledge about the experimental subject (difficulty I), treatment or independent variables (difficulty II, C, D, and F), measurement of results (difficulty III), how to deal with variability with randomization and replication of treatments (difficulty IV, C and D), and by selecting one shrimp species as the experimental subject (difficulty IVE), and interpretation of experimental findings (difficulty V). Thus, this assessment clearly was appropriate for the present study, as it is claimed to cover a wide range of difficulties. In the present study, we aimed to confirm this claim and to establish whether other difficulties were revealed by this assessment.

The Drug Assessment. The drug assessment, from an online database, Performance Assessment Links in Science (SRI International, 2003), asks students to design a controlled study to develop a new experimental drug for high blood pressure patients. This assessment was developed by the New York State Education Department to test for experimental design abilities in a medical context. According to the authors, this assessment is designed to measure experimental reasoning abilities such as "(1) stating hypothesis, (2) organizing experimental groups, (3) selecting participants in an experiment, (3) measurement of experimental results, and (4) drawing cause and effect claims from experimental findings." Based on these claims, this assessment probes for various difficulties listed in Table 1. The assessment asks students to propose a hypothesis by associating appropriate treatment and outcome variables (difficulty II, C and D), organize appropriate treatment and control groups (difficulty I and difficulty II, C and D), propose measurable outcomes (difficulty III), and account for variability sourced from unrelated variables through randomization and replication of treatments (difficulty IV, A–E). In addition, the assessment probes for cause-and-effect claims (difficulty V) by which the authors make reference to interpretation of findings (difficulty V) and the need to closely match the groups carrying treatment and control variables (difficulty II, C and E).

The Bird Assessment. A modification of the 2009 AP Statistics assessment was framed around the design of an experiment to study feeding habits of various bird species (henceforth bird assessment). This assessment was centered on statistical abilities for experimental design. According to the authors, the primary goals of this assessment were to assess students' ability to "(1) describe assignment of experimental units to treatments in a block design and (2) provide ways to increase the power of an experiment." These goals align with some of the Table 1 difficulties, because groups of experimental subjects to be tested should be considered based on a variable property appropriate for the goal of an investigation (difficulty I), and a treatment was to be applied to groups of birds as experimental subjects (difficulty II, C and F). The power of an experiment can be increased by replication of treatment conditions (difficulty IVD) and also by reducing influence of the unrelated variables (difficulty IVE). Finally, a good experiment would focus on appropriate measurements (difficulty III) for the proposed interpretation of the experimental findings (difficulty V).

Based on Table 1, one would expect to find the same established or partially established difficulties identified in previous research in the responses from undergraduate students to the assessments. In addition, one would expect data that will permit the above partially established difficulties to be reclassified as established. To test these predictions, we administered the three assessments to diagnose difficulties with experimental design among our own undergraduate student population.

For identification of difficulties undergraduate biology students have with experimental design, more than 1100 responses to three assessments completed by undergraduate biology student were examined and coded for their correct ideas or difficulties with experimental design. A range of responses gathered both before and after a first-year biology course included more than 500 responses to the shrimp assessment, more than 400 responses to the bird assessment, and 236 responses to the drug assessment, as illustrated in Figure 1B. Both inductive analysis of student responses to the assessments and the scoring materials from the publisher were used to characterize both the correct ideas and the difficulties expected from the literature review in Table 1.

Development of the RED. Using both the published difficulties in Table 1 and all responses to each published assessment from volunteers collected over a period of 3 yr, two coders started examining and coding for the students' difficulties. The coders had both completed graduate course work in education research and both were experienced lab scientists who are familiar with experimental design. Each coder coded responses independently, and then the coders came together to discuss codes to resolve any coding discrepancies. Coding was done blindly as to whether a particular response was from pre- or postinstruction. First, qualitative analysis was performed on responses to the shrimp assessment, using inductive coding to detect recurrent mistakes. The analyses involved discriminating accurate and flawed responses and assigning unique codes for each type of error. During inductive analysis, difficulties and accurate responses were read a number of times in order to discover similarities and emerging themes. Themes with similar meaning were coded together and grouped into a particular category (Table 2). Any discrepancy with categorizing responses under existing codes or creating new ones was discussed until agreement was reached. This method resulted in development of the RED as a rubric that represents all the difficulty themes under a particular category.

Addressing RQ3: Can a RED Be Usefully Deployed to Detect Changes in Undergraduate Students' Experimental Design Knowledge during a First-Year Biology Course?

Administering the Assessments. All assessments were administered, both pre- and postinstruction, via online Qualtrics survey software, and open-ended responses were collected as part of a regular homework assignment at the beginning and end of the semester each year. Students were given up to 10 points for providing their own ideas and thoughtfully written responses to the questions without consulting other sources. The survey took up to 30 min of their time. Most students enjoyed knowing that their ideas would be used to help improve instruction for students like them, and they appreciated the opportunity to get points for explaining their own ideas. Different assessments were used for pre- and posttests during a given semester to control for the same students absorbing knowledge by remembering and discussing what was asked when they attempted the test at the beginning of the course (Figure 1C).

Analysis of Responses. Student performance across four cohorts was examined to test our null hypothesis that the shrimp, drug, or bird assessment is *not* appropriate for showing differences in the proportion of students with correct ideas or difficulties in an area of experimental design knowledge at the beginning compared with the end of a semester. Our alternate hypothesis is that the shrimp, drug, or bird assessment *is* appropriate for showing differences in the proportion of students with correct ideas or difficulties in an area of experimental design knowledge at the beginning compared with the end of a semester. To test our hypothesis, we sampled responses using a random sampling approach and examined student responses for experimental design difficulties. In spite of groups being of different sizes across four cohorts (A–D), during random sampling, each response had an equal probability of selection for all students (Kish, 1965). Pre- and posttest responses were deidentified and blind coded to control for bias during analysis. Using the RED, sampled responses were coded independently by the first author once two independent coders achieved a high degree of interrater reliability, as reported below. As responses were coded, the sample size was gradually increased, until student difficulties appeared in a consistent manner and finally reached saturation. In this study, saturation was found with a sample of 40 responses per assessment. This means that after analyzing 40 responses, we recurrently found all difficulties listed in Table 2 and did not detect any new difficulties.

All responses to a particular assessment were collected as a pretest at the beginning of the semester, and then all responses to the assessment were collected from a different class as a posttest at the end of the semester (Figure 1C). Each pre- and posttest response was assigned an individual random number using the random number generator function within MS Excel. Then, for each assessment, the 40 lowest random numbers were selected from the pretest and 40 more were added from the posttest responses. This sampling process yielded an adequate uniform sample size to focus on the research questions and yet was manageable for classifying experimental abilities, given the qualitative nature of our coding approach. A random sample of the responses was used to reduce bias during coding and to allow for representation of the overall population (Rubin, 1973). When the same assessment was used at the beginning of the semester with one class and at the end of the semester with another class, we would expect to see a difference in results for students who have not taken this course (at the beginning) compared with those who have completed the course (at the end of course), provided these assessments are useful to characterize learning about experiments in this course.

To determine whether each published assessment could detect changes in student knowledge as a result of course participation, we applied Fisher's exact test to detect differences in correct knowledge and difficulties with experimental design knowledge at the beginning and at the end of a semester. The Fisher's exact test is appropriate when dealing with independent samples (Ramsey and Schafer, 2012). For this study, responses from one group of students before the course were compared with responses from a different population at the end of another semester using the same assessment. In other words, data collected from these two independent random samples produced results that fell into one of two mutually exclusive classes; to determine whether they differed, we compared the proportion with answers that were correct or showed a difficulty. Further, in order to characterize how well each assessment probes for experimental design knowledge with each of the three assessments, we calculated the percentage of students who expressed correct knowledge and difficulties for each broad area across responses to three assessments at the beginning and at the end of a semester.

Coding of RED Areas of Difficulty. Each response was assessed for evidence of difficulties. If a problem was found based on the RED, it was coded as a difficulty under the corresponding broad area (Table 2). For example, a difficulty with *randomization* in the shrimp assessment was noted under "Randomized design of an experiment" (Table 2, area of difficulty 4, d–f). For each of the five big areas, if the student

Table 2. Rubric for experimental design—RED[a]

Areas of difficulty	Propositional statements/completely correct ideas	Typical evidence of difficulties
1. Variable property of an experimental subject	Experimental subject or units: The individuals to which the specific variable treatment or experimental condition is applied. An experimental subject has a variable property. A variable is a certain property of an experimental subject that can be measured and that has more than one condition.	a. An experimental subject was considered to be a variable. b. Groups of experimental subjects were considered based on a property *that diverges* from the subjects that were the target for the stated investigation or claim to be tested. c. Variable property of experimental subject considered is not consistent throughout a proposed experiment.
2. Manipulation of variables	Testable hypothesis: A hypothesis is a testable statement that carries a predicted association between a treatment and outcome variable (Ruxton and Colegrave, 2006).	a. Only the treatment and/or outcome variable is present in the hypothesis statement. b. Hypothesis does not clearly indicate the expected outcome to be measured from a proposed experiment.
	Treatment group: A treatment group of experimental subjects or units is exposed to experimental conditions that vary in a specific way (Holmes *et al.*, 2011).	c. Haphazard assignment of treatments to experimental units in a manner inappropriate for the goal of an experiment. d. Treatment conditions proposed are unsuitable physiologically for the experimental subject or inappropriate according to the goal of an investigation.
	Combinatorial reasoning: In experimental scenarios, when two or more treatment (independent) variables are present simultaneously, all combined manipulations of both together are examined to observe combinatorial effects on an outcome.	e. Independent variables are applied haphazardly in scenarios when the combined effects of two independent variables are to be tested simultaneously. f. Combining treatments in scenarios where the effect of two different treatments are to be determined individually
	Controlling outside variables: The control and treatment groups are required to be matched as closely as possible to equally reduce the effect of lurking variables on both groups (Holmes *et al.*, 2011).	g. Variables unrelated to the research question (often showing a prior knowledge bias) are mismatched across treatment and control groups.
	Control group: A control group of experimental subjects or units, for comparison purposes, measures natural behavior under a normal condition instead of exposing them to experimental treatment conditions. Parameters other than the treatment variables are identical for both the treatment and control conditions (Gill and Walsh, 2010; Holmes *et al.*, 2011).	h. The control group does not provide natural behavior conditions, because absence of the variable being manipulated in the treatment group results in conditions unsuitable for the experimental subject. i. Control group treatment conditions are inappropriate for the stated hypothesis or experimental goal. j. Experimental subjects carrying obvious differences are assigned to treatment vs. control group.
3. Measurement of outcome	Treatment and outcome variables should match up with proposed measurements or outcome can be categorical and/or quantitative variables treatments A categorical variable sorts values into distinct categories. A quantitative or continuous variable answers a “how many?” type question and usually would yield quantitative responses.	a. No coherent relationship between a treatment and outcome variable is mentioned. b. The treatment and outcome variables are reversed. c. An outcome variable that is quantitative is treated as a categorical variable.
	Outcome group: The experimental subject carries a specific outcome (dependent variable) that can be observed/measured in response to the experimental conditions applied as part of the treatment (Holmes *et al.*, 2011).	d. Outcome variables proposed are irrelevant for the proposed experimental context provided or with the hypothesis. e. Stated outcome not measurable. f. No measure was proposed for the outcome variable. g. An outcome variable was not listed for an investigation. h. There is a mismatch between what the investigation claims to test and the outcome variable.

(*Continued*)

Table 2. Continued

Areas of difficulty	Propositional statements/completely correct ideas	Typical evidence of difficulties
4. Accounting for variability	Experimental design needs to account for the variability occurring in the natural biological world. Reducing variability is essential to reduce effect of nonrelevant factors in order to carefully observe effects of relevant ones (Box *et al.*, 2005; Cox and Reid, 2000).	a. Claims that a sample of experimental subjects will eliminate natural variability with those subjects.
	Selection of a random (representative) sample: A representative sample is one where all experimental subjects from a target demographic have an equal chance of being selected in the control or treatment group. An appropriate representative sample size is one that averages out any variations not controlled for in the experimental design (College Board, 2006; Holmes *et al.*, 2011).	b. Criteria for *selecting* experimental subjects for treatment versus control group are biased and not uniform. c. Criteria for selecting experimental subjects for investigation are different in a way that is not representative of the target population.
	Randomized design of an experiment: Randomizing the order in which experimental subjects or units experience treatment conditions as a way to reduce the chance of bias in the experiment (Ramsey and Schafer, 2012).	d. Decisions to *assign* experimental subjects to treatment vs. control group are not random but biased for each group. e. Random assignment of treatments is not considered.
	Randomization can be complete or restricted. One can restrict randomization by using block design, which accounts for known variability in the experiment that cannot be controlled.	f. Random assignment of treatments is incomplete, as they show random assignment of the experimental subjects, but what is needed instead is random assignment of treatments.
	Replication of treatments to experimental units or subjects: Replication is performed to assess natural variability, by repeating the same manipulations to several experimental subjects (or units carrying multiple subjects), as appropriate under the same treatment conditions (Quinn and Keough, 2002).	g. Replication means repeating the entire experiment *at some other time* with another group of experimental subjects. h. No evidence of replication or suggested need to replicate as a method to access variability or to increase validity/power of an investigation.
5. Scope of inference of findings	Scope of inference: Recognizing the limit of inferences that can be made from a small characteristic sample of experimental subjects or units, to a wider target population and knowing to what extent findings at the experimental subject level can be generalized.	a. The inference from a sample is to a different target population. Usually students under- or overestimate their findings beyond the scope of the target population. b. No steps are carried out to randomly select experimental subjects' representative of the target population about which claims are made.
	Cause-and-effect conclusions: A cause-and-effect relationship can be established as separate from a mere association between variables only when the effect of lurking variables is reduced by random assignment of treatments and matching treatment and control group conditions as closely as possible. Appropriate control groups also need to be considered also in comparison to the treatment group ([National Institute of Standards and Technology NIST]/SEMATECH, 2003; Wuensch, 2001).	c. A causal relationship is claimed even though the data show only association between variables. Correlation does not establish causation. (NIST/SEMATECH, 2003)

[a] Refer to the glossary of terms in the Supplemental Material (p. 20).

showed evidence of any difficulty with underlying components, that response was coded under difficulty for that big area. A difficulty with any one component under area *accounting for variability* would count as a difficulty for this overall area.

Second, if we found no difficulty, we looked for evidence that shows clear understanding. Finally, if a response did not show evidence (correct or flawed) about a certain broad area, it was listed as "lack of evidence" (LOE) for that area. For example, a shrimp assessment response stating "measure effect of nutrients/salinity on shrimp" was considered as LOE for the area measurement of outcome, because no indication for what to measure (shrimp growth) was characterized by the phrase "measure effect."

At the same time as difficulties were identified, a corresponding statement was written to describe knowledge that represents correct understanding of each area based on clear definitions of key experimental design concepts (refer to the glossary of terms in the Supplemental Material). For the five areas, this was done by reviewing the literature for statements of correct knowledge. Accurate statements were validated with expert faculty and graduate students over a 3-yr period, using an iterative process until consensus was reached. The experts included a biologist who was head of undergraduate programs, a biochemist, four science education graduate students, and members of a faculty learning community that involved faculty members from the biology and statistics departments. Examples of data to illustrate typical difficulties for each correct idea are presented below and in the Supplemental Material (Supplemental Tables S1–6). The corresponding accurate statements are listed in Table 2 under "Propositional Statements/Completely Correct Ideas."

Interrater Reliability. Two raters (first author and another graduate student) coded each response in terms of five areas in the RED (Table 2). For initially familiarizing the second coder with the RED, response examples with correct and flawed responses to each assessment were used to enable the second coder to understand the rubric and further apply it to characterize student responses (see Supplemental Tables S1–3). Once 100% agreement with the RED was reached for coding the sample, the coders separated to code independently. A sample of 10 responses for three assessments each (30 responses total) was coded using the analysis approach described. To examine reliability of coding across raters, we compared overall area codes. In other words, if rater A coded a response showing difficulty for the area measurement of outcome, we checked whether rater B also coded the response as "difficulty" or "correct" under measurement of outcome. To statistically estimate the degree of agreement as per five areas, we coded a Cohen's kappa value for each area on each assessment individually (Cohen, 1960). Cohen's kappa is considered a better measure of interrater agreement than the simple percent agreement calculation, because it adjusts for the amount of agreement due to chance. A resulting Cohen's kappa value of $\kappa = 0.68$ would indicate substantial agreement (Landis and Koch, 1977), meaning that, with careful definition of the coding protocol and well-trained coders, responses to each assessment could be reliably coded and scored.

FINDINGS

In addressing RQ1, the literature review (Table 1) revealed that most authors had identified several major categories of difficulty, all of which were classified by us as established, except for two difficulties, which had limited available evidence and were classified as partially established. It is important to note, though, that most authors failed to present data that allowed them to unpack or characterize each difficulty category into subcategories that would be more useful to instructors. In addressing RQ2, our qualitative data from the undergraduate biology students' responses to the three selected assessment instruments allowed us to significantly extend the literature knowledge to include multiple subcategories of difficulty allowing us to develop the RED. To ensure that the RED would be useful for characterizing both correct and flawed responses, we pooled data from both pre- and posttests, which made it more likely that the full range of qualities of understanding about experimental design would be covered. In addition, to optimize confidence in our data used to inform the RED, we used only established and partially established difficulties based on the literature review (RQ1) that included only primary research reports.

In this section, for reader convenience, we first present and describe the RED, and thereafter we present the detailed data used to inform the development and validation of this rubric.

The RED

To understand what types of difficulties undergraduate biology students have with experimental design, besides the data from the literature review (RQ1), we examined all answers to three assessments to identify difficulties documented in the literature, as well as other flawed responses, using an iterative process over a period of 3 yr. This process led to the development of the RED (Table 2) with five major categories of student difficulties with experimental design as themes: 1) variable property of an experimental subject; 2) manipulation of variables; 3) measurement of outcome; 4) accounting for variability, and 5) scope of inference based on the findings. These five categories form the basic framework for the RED, with multiple subcategories of difficulty under each major category (Table 2). When the RED was tested for interrater reliability as described above, the average kappa value obtained was 0.9 (see Supplemental Tables S7–9 for detailed calculations), assuring high intercoder reliability (Landis and Koch, 1977). Perhaps not surprisingly, when the RED was used as a guide to characterize and distinguish responses with difficulties from accurate responses, those with difficulties were consistent with low scores according to the scoring guidelines published by authors of the assessments (see scoring guideline links in the Supplemental Material). In the sections below, we present (Table 3) and discuss the detailed data that supported the formulation of the RED.

Difficulties with Experimental Design Detected Using the Published Assessments (RQ2)

To understand to what extent published assessments reveal evidence of first-year undergraduate biology students' knowledge and difficulties with experimental design, we used responses to the shrimp, drug, and bird assessments to identify students' correct ideas and difficulties, which, as shown in Table 3, were then classified within all five categories of difficulty. In the following sections, we discuss the examples of student responses from Table 3, demonstrating correct ideas and typical difficulties with five RED areas to each of three assessments. A detailed explanation of each example is provided. For each assessment, a more complete example from a student with an overall correct idea and a typical response from a student who shows difficulties are presented in Supplemental Tables S1–3. For confidentiality, pseudonyms are used to identify students.

Variable Property of an Experimental Subject. Difficulty with identifying an appropriate experimental subject with a variable property to be investigated was a problem for students across all three assessments. Students had trouble recognizing that an experimental subject possesses properties that vary, the sample of experimental subjects must display an appropriate variable property aligned with the given experimental goal, and the variable property needs to be consistently considered when planning an investigation (Table 2, area of difficulty 1, a–c).

As illustrated in Table 3 (1.Shrimp.C), Anna correctly recognizes tiger shrimp as an experimental subject in the shrimp assessment, but Beth shows a difficulty with the experimental subject (tiger shrimp), as she considers it to be a variable and includes it as a part of the experiment control (1.Shrimp.D). Instead, the correct idea would be to think of a variable

Table 3. Examples of student responses with the RED areas of difficulty across three assessments

1. Variable property of an experimental subject

Shrimp assessment
Correct (C) idea from Anna: *"The advantage to having only tiger shrimp in the experiment is that you are only using one single species of shrimp. This leads to an advantage because there is less variability within the growth of shrimp."*
Difficulty (D) from Beth: *"The tiger shrimps act as the control group."* (area of difficulty 1a)

Drug assessment
Correct (C) idea from Josh: *"Patients need to have [same range of] high blood pressure."*
Difficulty (D) from Ken: *"Participants cannot be pregnant simply because it will affect the fetus differently than the adult. People older than 35 should not test the drug."* (area of difficulty 1b)

Bird assessment
Correct (C) idea from Rita: *"Knowing from previous research that male birds do not avoid solid colors . . . Ensuring that all of the birds being tested are as similar as possible except for the treatment is best. This entails that all birds have the same gender."*
Difficulty (D) from Sara: *"The reason for these differences between the two sexes could have to do with the fact that one sex is the main contributor of food to their young . . . You could set up three separate areas having one species assigned to one of the three."* (area of difficulty 1c)

2. Manipulation of variables

Shrimp assessment
Correct (C) idea from Anna: *"1. A Low salinity; 2. A high salinity; 3. B low salinity; 4. B high salinity; 5. C low salinity; 6. C high salinity."*
Difficulty (D) from Beth: *"Low salinity with no nutrient, high salinity with no nutrients."* (area of difficulty 2, c and f)

Drug assessment
Correct (C) idea from Josh: "[Administration of] new drug . . . *lower the blood pressure of people with high blood pressure to a safe level . . . same range of high blood pressure, diet, exercise, eating habits, sleep habits."*
Difficulty (D) idea from Ken: (i) *"This drug will be administered to people at low dosages at first, then we will record results and from there calculate the correct amount of Alamain that should be given to each person."* (area of difficulty 2b)
(ii) *"Experimental groups will receive a couple of different dosages to see how each dose affects blood pressure."* (area of difficulty 2d)
(iii) *"The younger, healthier participants will be the experimental group while the not so young will be the control."* (area of difficulty 2j)

Bird assessment
Correct (C) idea from Rita: (i) *"Each species of bird would be randomly divided into two groups, with one group receiving treatment 1 and the other group receiving treatment 2 (that is, 50 blackbirds would receive treatment 1, 50 blackbirds would receive treatment 2, and likewise for zebra finches and geese)."*
(ii) *"Ensuring that all of the birds being tested are as similar as possible except for the treatment is best. This entails that all birds have the same gender, are roughly the same age, come from very similar habitats, and are in overall good health (no underlying conditions such as currently suffering from a given disease)."*
Difficulty (D) idea from Sara: (i) *"You could repeat the experiment but this time allowing all three of the species to be in the same area."* (area of difficulty 2, d and f)
(ii) *"This experiment would take into account any competition [among all three bird species] that might take place"* (area of difficulty 2g)

3. Measurement of outcome

Shrimp assessment
Correct (C) idea from Anna: *"The advantage to having only tiger shrimp in the experiment is that there is less variability within the growth of a single species of shrimp."*
Difficulty (D) from Beth: *"A researcher can confidently expect to find a repetitive response to a given exposure in a group of genetically identical tiger shrimps."* (area of difficulty 3e)

Drug assessment
Correct (C) idea from Josh: *"If people who take the drug consistently have decreased blood pressure, then the drug is effective."*
Difficulty (D) from Ken: *"If the drug does indeed reduce blood pressure, the percentage of those whose blood pressure [becomes] normal will be significantly higher than that [of the] control group."* (area of difficulty 3g)

Bird assessment
Correct (C) idea from Rita: *"Differences in the response variable (in this case, the frequency of avoiding or not avoiding food given the particular treatment) can be [attributed to] the difference in treatment."*
Difficulty (D) from Sara: *"They [all three bird species] all will be in the same area together and not separated . . . This would increase the power by determining which seed the birds compete over and which seed the birds ignore . . . After the time is up, you could collect the remaining seeds and see which treatment was eaten the most and which treatment the birds avoided the most."* (area of difficulty 3, c and g)

4. Accounting for variability

Shrimp assessment
Correct (C) idea from Anna: "Using only tiger shrimps reduces variance."
"There are two tanks with each treatment."
"In order for randomization to occur it might be easiest to use dice and assign each number to its corresponding treatment number. Example: Roll dice 1+ 2; Outcome Die 1 = 2 and Die 2 = 4. From this you would put treatment two and four in tanks 1 and 2."
Difficulty (D) from Beth: (i) *"A researcher can confidently expect to find a repetitive response to a given exposure in a group of genetically identical tiger shrimps."* (area of difficulty 4, a and h)
(ii) *"With all the shrimp in one tank, one by one randomly assign a shrimp to a tank . . . by doing this, the biologist is aware of which tanks contain which ingredients but the shrimp are completely randomized."* (area of difficulty 4f)

(*Continued*)

Table 3. Continued

Drug assessment

Correct (C) idea from Josh: *"They [experimental subject/participants] will have to be at the same range of high blood pressure, diet, exercise, eating habits, sleep habits."*

"They [participants] will be chosen at random to be part of the experimental or control group that way they do not have an opinion on how the drug may or may not be helping them."

Difficulty (D) idea from Ken: (i) *"People older than 35 should not test the drug. These criteria need to be met and not taken lightly because health problems may arise."* (area of difficulty 4c)

(ii) *"The younger, healthier participants will be the experimental group while the not so young will be the control."* (area of difficulty 4d)

Bird assessment

Correct (C) idea from Rita: *"Each species of bird would be randomly divided into two groups, with one group receiving treatment 1 and the other group receiving treatment 2."*

Difficulty (D) from Sara: *"You could set up three separate areas having one species assigned to one of the three."* (area of difficulty 4e)

5. Scope of inference

Shrimp assessment

Correct (C) idea from Anna: *"One statistical disadvantage to only having only tiger shrimp is that due to the fact we only used one species of shrimp we are not able to make a generalization about all shrimp."*

Difficulty (D) from Beth: *"This fails to demonstrate how a given ingredient may affect another type of shrimp. Ultimately it limits the depth of the study."* (area of difficulty 5, b and c)

Drug assessment

Correct (C) idea from Josh: *"Participants with same range of high blood pressure, diet, exercise, eating habits, and sleep habits . . . blood pressure [will be measured]. . . participants chosen at random."*

Difficulty (D) from Ken: *"Health, hemoglobin, smoking, age under 35, and pregnancy status."* (area of difficulty 5, a and c)

Bird assessment

Correct (C) idea from Rita: *"With all of these potential differences eliminated, the birds would be made different in only one respect: their treatment. In this manner, one would be able to confidently declare that differences in the response variable [in this case, the frequency of avoiding or not avoiding food given the particular treatment] can be laid at the feet of the difference in treatment."*

Difficulty (D) from Sara: *"The reason for these differences between the two sexes could have to do with the fact that one sex is the main contributor of food to their young . . . You could set up three separate areas having one species assigned to one of the three . . . Determining which seed the birds compete over and which seed the birds ignore . . . You could set up three separate areas having one species assigned to one of the three."* (area of difficulty 5, b and c)

property of the experimental subject (Table 2, area of difficulty 1a).

In the drug assessment, Josh suggests maintaining the variable property "blood pressure" constant (Table 3, 1.Drug.C), but Ken proposes experimental subjects divergent from the proposed target population (Table 2, area of difficulty 1b). This is a problem, because Ken considers including patients on the basis of pregnancy status and age (1.Drug.D) instead of sampling an appropriate target population for the drug (people with high blood pressure).

For the bird assessment, one appropriate variable property of birds is the species: blackbirds, zebra finches, and geese. Part of the assessment asks about differences in food preference for zebra finches, but another part focuses on one gender (male) of three different bird species. Rita considers the experimental subject (birds) appropriately with reference to the gender of zebra finches in her initial response, and then she proposes a study with the three species but maintains a consistent reference to the birds' gender (Table 3, 1.Bird.C). This shows that Rita correctly explains the experimental subject in terms of a variable property aligned with the goal of the experiment. In contrast, Sara, in the first part of the response, considers groups of experimental subject based on the gender of zebra finches. But then she shifts to talking about the species with no reference to a specific gender (1.Bird.D). This shows a lack of coherence, because variable property of the experimental subject was not consistently considered (Table 2, area of difficulty 1c).

Manipulation of Variables. Across the three assessments, an appropriate response for manipulating variables would have been to come up with appropriate treatment and control groups and to recognize unrelated variables to a given study. A clear pattern of difficulties was found across the three assessment instruments when students were challenged to hypothesize and manipulate treatment variables during the process of experimental design. Students often did not focus on the right variables. Sometimes they considered irrelevant variables, while at other times, they proposed inappropriate treatments or failed to combine two treatments as required for the experimental goal. Finally, students had trouble matching treatment and control conditions to neutralize effects of lurking/confounding variables for an experiment (Table 2, area of difficulty 2, a–j).

With the shrimp assessment, Anna sets up appropriate treatment groups carrying combinations of two independent treatment variables (nutrient and salinity) applied to the experimental subject (tiger shrimp) (Table 3, 2.Shrimp.C). However, this seems to be difficult for Beth, who haphazardly proposes treatment groups (Table 2, area of difficulty 2c) with missing conditions to keep the shrimp alive (2.Shrimp.D). This also shows a problem with combinatorial reasoning, as Beth fails to combine salt and nutrients appropriately to find their effect on the growth of shrimp (area of difficulty 2f).

Josh's hypothesis for the drug assessment shows a clearly predicted testable association between a treatment and an outcome (Table 3, 2.Drug.C). In contrast, Ken demonstrates difficulty in framing a hypothesis, as he fails to identify a clear expected result from the proposed experiment, as evident from 2.Drug.D.i (Table 2, area of difficulty 2b). Also, Ken proposes treatment conditions such as "different dosages of the blood pressure drug" (2.Drug.D.ii) inappropriate to the original goal of the investigation, which is to test the effect on blood pressure from the presence and absence of drug intake (Table 2, area of difficulty 2d). In an experiment, the control and experimental groups are required to be matched as closely as possible to equally reduce the effect of unrelated variables on both groups. Josh demonstrates this ability well by matching appropriate variables to control lurking variables in a study to develop a high blood pressure drug (2.Drug.C). However, Ken should not have assigned the participants (experimental subjects) carrying obvious differences (young/healthy and not so young) to treatment and control group, respectively (2.Drug.D.iii; Table 2, area of difficulty 2j), because parameters other than the treatment variables need to be identical for both the treatment and control conditions.

For the bird assessment, Rita correctly organizes assignment of experimental units to treatments in alignment with the experimental goal to examine preference in consuming either of two kinds of pesticide granules among three different bird species separated by a block design (Table 3, 2.Bird.C). Sara, on the other hand, tries to combine all three different bird species within a single treatment group (2.Bird.D.i) when, instead, the effect of treatments are to be determined individually for each bird species by "block design." Thus, we conclude Sara shows a difficulty in identification of treatment groups and combinatorial reasoning (Table 2, area of difficulty 2, d and f).

Another measure to identify treatment and control groups by Rita was controlling outside variables by matching up the various treatment groups in terms of lurking variables that could affect bird behavior (Table 3, 2.Bird.C). In contrast, Sara considers "competition among bird species" as a variable unrelated to the intended goal of finding out what pattern or color of pesticide granules each species would avoid eating (2.Bird.D.ii; Table 2, area of difficulty 2g).

Measurement of Outcome. With correct knowledge of measurement of outcome, a student would propose experimental outcomes using appropriate measures. However, in their responses to all three assessments, some students struggled with measures when they either failed to state outcomes that were measurable or they proposed outcomes without specific measures in terms of units or categories. Sometimes those who did propose measurable outcomes suggested variables that were mismatched to a given experimental goal (Table 2, area of difficulty 3, a–g).

The "growth of shrimp" as a measurable outcome is correctly identified in Anna's response to the shrimp assessment (Table 3, 3.Shrimp.C). But for Beth's response (3.Shrimp.D), the phrase "repetitive response" provides no measure for a specific outcome, thereby she demonstrates difficulty for measurement of outcome (Table 2, area of difficulty 3e).

For the drug assessment, Josh suitably suggests "decrease in blood pressure" as an outcome (Table 3, 3.Drug.C). But Ken's proposed outcome (3.Drug.D) illustrates a mismatch between the goal of the investigation and the outcome to be measured (Table 2, area of difficulty 3g). Specifically, this is a mismatch, because having more participants with normal blood pressure is different from saying that participants' blood pressure will be lower if the drug is effective. In other words, an effective drug is one that simply reduces high blood pressure for the treatment group participants but not necessarily down to normal levels.

In the bird assessment, an appropriate measure for an outcome variable is suggested by Rita (Table 3, 3.Bird.C). Sara shows a problem with her proposed measurement of outcome (3.Bird.D) when she indicates that the bird species will "compete" for seeds, which is irrelevant to the stated goal of this investigation (Table 2, area of difficulty 3c). There is a mismatch between what the question asked and the investigation goal, because "which treatment was eaten the most" is not a relevant outcome when the goal is to find out whether or not the birds consume seeds, not "how much" they consume (area of difficulty 3g).

Accounting for Variability. Correct ideas about accounting for variability would require recognizing natural variation among experimental subjects while trying to reduce variation sourced externally from unrelated factors. We found that, across three assessments, students showed flawed ideas concerning variability in multiple ways. Either they completely failed to recognize natural variation or they failed to account for variability with appropriate methods like replicating and randomizing treatment assignments (Table 2, area of difficulty 4, a–h).

For the shrimp assessment, Anna shows a correct understanding of how to deal with natural biological variability (Table 3, 4.Shrimp.C). In contrast, Beth reveals a difficulty with variability (4.Shrimp.D.i) as the phrase "genetically identical tiger shrimps" incorrectly claims that having only tiger shrimp eliminates natural variability. In fact, some variability exists even within a sample of the same species (Table 2, area of difficulty 4a). Another component for this area includes "replication of treatment conditions" as a measure to assess natural variability within an experimental unit carrying multiple experimental subjects. This is included in Anna's response (4.Shrimp.C), but Beth does not consider replication of treatment (4.Shrimp.D.ii; Table 2, area of difficulty 4h).

To account for known variability from lurking variables in an experiment requires randomizing the order in which experimental units experience treatment conditions (Table 2, area of difficulty 4). Randomization is well described in Anna's response, as she illustrates a complete randomization of assignment of both treatment and shrimps to tanks (Table 3, 4.Shrimp.C). Alternatively, an incomplete randomization procedure (Table 2, area of difficulty 4f) is suggested by Beth, who only randomizes assignment of shrimp to tanks but fails to randomize assignment of treatment combinations to each tank (Table 3, 4.Shrimp.D.ii).

For the drug assessment, Josh proposes to deal with variation using a random sample to represent a target population (Table 3, 4.Drug.C). Instead, Ken selects experimental subjects who are not representative of the target demographic population and who are also not randomly chosen (Table 2, area of difficulty 4c; 4.Drug.D.i and ii), because participants with different characteristics are purposefully assigned to treatment and control groups (Table 2, area of difficulty 4d).

In the bird assessment, evaluating how students randomly assign each of three bird species to two treatments provides a measure of how well students address natural variability in an experiment. This is demonstrated well by Rita (Table 3, 4.Bird.C). Alternatively, Sara sets up separate areas for each species but does not specify how treatments are assigned in a randomized manner (4.Bird.D; Table 2, area of difficulty 4e).

Scope of Inference. When a student demonstrates correct ideas about interpretation of experimental findings, he or she estimates an appropriate extent of inference of findings and is also able to draw logical causal claims. But across the three assessments, we found students went wrong with interpretation of experimental findings in several ways. They either over- or underestimated experimental claims, or they made inappropriate inferences about causal relationships, while their experimental procedures only suggested correlation among variables (Table 2, area of difficulty 5, a–c).

For the shrimp assessment, both Anna and Beth recognize the limit of inferences from a small sample of tiger shrimps (Table 3, 5.Shrimp.C). However, Beth still shows difficulty in this area, because she does not mention a measurable outcome or randomization and replication of treatments and fails to recognize natural variability with the experimental subjects. With such flaws, Beth only shows signs of correlation and not of causal association (Table 3, 5.Shrimp.D) between application of variable nutrient and salinity conditions and growth of tiger shrimps (Table 2, area of difficulty 5, b and c).

On the drug assessment, Josh's experimental findings can be generalized to an appropriate sample of the target population of people with high blood pressure. He makes specific considerations during selection of experimental subjects and the identification of experimental groups, and he applies methods to deal with variability (Table 3, 5.Drug.C). Similarly, his proposed measurement of outcome ("blood pressure") and measures for accounting for variability ("participants chosen at random") justify appropriate cause-and-effect conclusions about the effectiveness of the high blood pressure drug. In contrast, Ken's study will apply to a different target population and not the intended subjects with high blood pressure, due to lack of appropriate accounting for variability measures and a skewed participant pool with demographic properties not representative of a larger target population (Table 2, area of difficulty 5a). Similarly, due to selection bias based on irrelevant variables (5.Drug.D), when he selects and assigns participants to treatment groups, causal claims would be inappropriate, because of Ken's flawed comparison groups (Table 2, area of difficulty 5c).

For the bird assessment, careful considerations include appropriate groups of experimental subjects, an organized setup of experimental groups, suitable measurable outcomes, and methods to account for natural variability among bird species for Rita's study, making her design suitable for causal claims. Rita correctly asserts a causal claim in her answer (Table 3, 5.Bird.C). In contrast, Sara's experimental design lacks coherence in several areas. The experimental groups are not considered consistently across different parts of the response, treatment assignments follow a pattern unsuitable to the study goal, proposed outcomes do not match the original investigation goal, and efforts to account for natural variability are inadequate. These flaws make it unfeasible to draw any cause and effect conclusions (5.Bird.D) from Sara's experimental proposal (Table 2, area of difficulty 5, b and c).

Interconnectedness of RED Areas of Difficulty. In examining problems with student interpretation of experimental findings for each of the three assessments, an interesting finding was that student difficulties with two RED categories (Tables 2 and 3) often went together. The categories were not independent but interconnected. For example, it is not surprising that a difficulty with controlling outside variables categorized under manipulation of variables was associated with difficulty accounting for variability, because controlling outside variables provides a way to account for and minimize natural variation in samples. Likewise, proposal of a suitable testable hypothesis with appropriate manipulation of variables was connected to measurement of outcome difficulties because, if the hypothesis carried inappropriate relationships between treatment and outcome variables, the outcome measurements were also flawed. Accounting for variability influenced inferences drawn from experimental findings or scope of inference. Without considering variability, students overestimated or underestimated findings beyond the scope of the participating sample of a "population" in a study (Table 2, area of difficulty 4a). Similarly, correlations were erroneously considered to demonstrate experimental evidence for causal relationships. Causation requires possible lurking variables to be carefully controlled for by random selection of representative experimental subjects.

The various types of typical evidence of difficulties in the RED (Table 2) were confirmed with responses to three different assessments, as illustrated with quotes (Table 3). Supplemental Tables S1–3 provide actual student responses with examples of typical correct ideas and difficulties according to the RED. The difficulties are underlined and coded with a footnote that corresponds to Table 2. But the examples discussed did not illustrate all types of typical evidence of difficulties from Table 2, so actual responses to illustrate other difficulties are provided in Supplemental Tables S4–6. Consistently, a careful analysis of responses revealed difficulties with experimental design in five areas: 1) a property of an experimental subject that is variable, 2) manipulation of variables, 3) measurement of outcome, 4) accounting for and measuring variability, and 5) scope of inference of findings. These five areas were used to develop the RED and thus formed the foundation for subsequent analysis.

Efficacy of the RED to Detect Changes in Students' Experimental Design Abilities (RQ3)

With the various experimental design difficulties now characterized in the RED, we recognized that, for practical purposes, the RED must be validated for its usefulness to detect changes in undergraduate student responses before and after a course (RQ3). We argued that, if the RED is sensitive enough to detect changes in the proportion of undergraduate students with correct responses, a similar measure at the end of course would help us find out whether students are learning about experimental design from our course. To make good decisions about how to focus on student difficulties that needed attention, we needed to know whether some assessments were better than others at probing particular knowledge. The proportion of students who showed correct ideas or difficulties was calculated after coding responses with the RED. For each area, the percentage of students with correct knowledge (dark gray), difficulties (medium gray), or LOE (light gray) is presented in Figure 2. Results show that with the three selected assessments, RED coding is capable of detecting differences in the proportion of students with correct knowledge or difficulties in the five experimental design areas (Table 2).

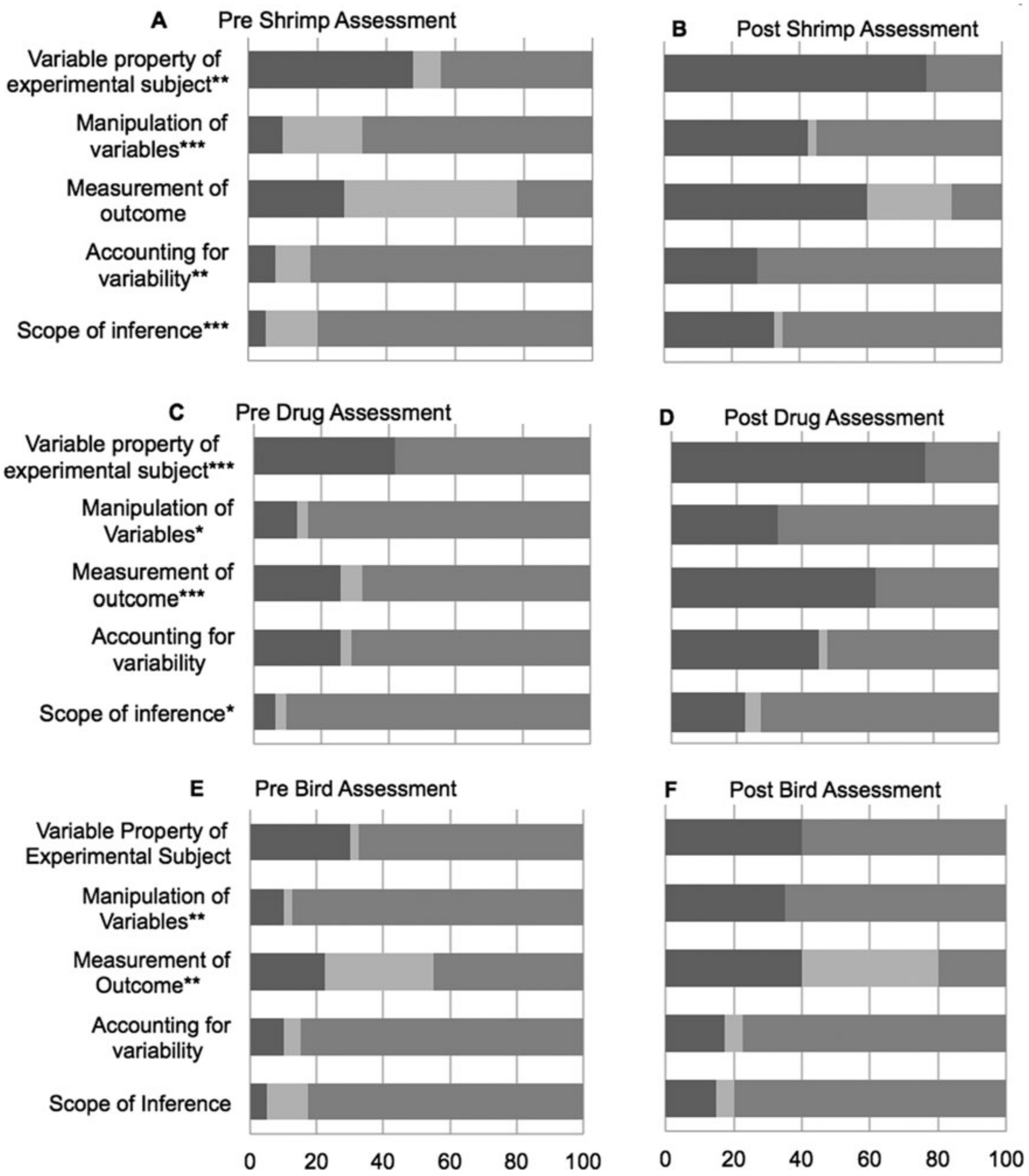

Figure 2. Proportions of students who had correct ideas (dark gray), difficulties (medium gray), and LOE (light gray) for knowledge of experimental abilities as probed by three assessments administered at the beginning and at the end of a semester. The shrimp assessment was given as a posttest during 2009 to cohort A (panel B; $n = 40$) and as pretest during 2010 to cohort B (panel A; $n = 40$). The drug assessment was used as a posttest in 2011 for cohort C (panel D; $n = 40$) and as a pretest in 2012 for cohort D (panel C; $n = 31$). The bird assessment was assigned as a posttest in 2010 to cohort B (panel F; $n = 40$) and as a pretest in 2011 to cohort C (panel E; $n = 40$). The y-axis topics are areas of difficulty from Table 2. Fisher's exact test was applied to compare responses at the beginning and at the end of a semester to detect differences in correct knowledge and difficulties in each area of difficulty for each assessment. *, $p < 0.1$ significance level; **, $p < 0.05$ significance level; ***, $p < 0.01$ significance level.

Our analysis showed that, in the case of certain RED areas, there were significant differences between pre- and posttest with p values ranging from ≤ 0.01 to ≤ 0.1, which implies that each assessment was capable of measuring changes in student knowledge with respect to certain RED areas. We consider a significance level of $p < 0.1$ to be adequate because with written response data, our understanding of changing knowledge is limited to what students write. Thus, we might have a 10% chance of being uncertain about the precision of these assessments in demonstrating experimental design knowledge. However, for research purposes with a cutoff at $p < 0.05$ significance levels, each assessment would still be a useful measure of certain RED areas. For example, the shrimp and drug assessment report pre- versus posttest p values for areas like variable property of experimental subject at <0.05 significance levels.

Looking across the data for the three assessment instruments (Figure 2), a clear pattern of differences at the beginning and end of a course is revealed when the RED was used to code a sample of responses. The manipulation of variables is an area that consistently showed significant difference between the pre- and posttest for all three assessments. This difference was detected even though, for all three assessments, more than half of the students still showed difficulty with manipulation of variables at the end of the course. Figure 2 shows that even though a significant difference was not found on one of the tests for variable property of an experimental subject, measurement of outcome, and scope of inference, the trend was the same as for two of the assessments that did show a significant difference at the beginning and end of a course in these areas. Although one area showed significant difference between the pre- and posttest for only one assessment, accounting for variability trends were also similar for this area across all three tests.

All three assessments showed similar differences in the proportion of students with correct ideas about experimental design and the areas of difficulties that need to be addressed. Next, we present Figure 2 findings, first in terms of the magnitude and direction of change in the proportion of students with correct ideas about experimental design, and then by considering the proportion of students who have difficulties in each area when responses are coded using the RED.

The proportion of students with correct responses at the beginning and the end of the course are aligned for all areas across three assessments in Figure 2, A–F. For the shrimp assessment, by the end of semester, variable property of experimental subject, manipulation of variables, and measurement of outcome showed the largest differences in proportion of students with "correct" ideas (Figure 2, A and B). (Supplemental Table S11 shows actual differences in proportion of students with ideas that were "correct" or showed "difficulty" at the beginning or end of a semester with each assessment.) Similarly, the drug assessment showed more differences in "correct" responses for variable property of experimental subject and measurement of outcome, but it was less sensitive for detecting differences in the proportion of students with correct ideas for manipulation of variables (Figure 2, C and D). The bird assessment was most sensitive in detecting pre- to posttest differences in the proportion of students with "correct" ideas in the areas of manipulation of variables and measurement of outcome, but it was less sensitive for prompting correct ideas about variable property of experimental subject at the end of the course (Figure 2, E and F). A small portion of students had correct ideas about accounting for variability at the end of the course, except in the drug assessment, which similarly prompted nearly a fourth of the students to account for variability at the start of the course. Differences were small, but the trend was the same across all three assessments. According to all three assessments, although some differences are apparent, only a small portion of students had correct ideas about scope of inference even at the end of the course. We acknowledge that, because the assessments were used for diagnostic purposes, we did not give partial credit for distinguishing average students from those with poor understanding corresponding to each RED area. A relatively stringent cutoff was appropriate, because we did not use students' responses for grading purposes. The assessments simply provided opportunities for students to demonstrate their thinking, so we would know what the problems are when students design experiments.

In addition to detecting correct ideas, each assessment also captured information about the proportion of students who demonstrated difficulty with five experimental knowledge areas. From the beginning to the end of the semester, the shrimp assessment measured the largest differences in difficulty for variable property of experimental subject and scope of inference, but for measurement of outcome, the difference found was only 8% (medium-gray bars in Figure 2, A and B). For the drug assessment, the biggest differences in proportion of students with difficulty were detected for variable property of experimental subject and measurement of outcome, and it was less sensitive for detecting difference in difficulties for manipulation of variables (medium-gray bars in Figure 2, C and D). Similarly, for the bird assessment, the largest differences in the proportion of students with difficulties were found for the areas measurement of outcome and manipulation of variables, while difficulties involving accounting for variability and scope of inference remained almost unchanged at the end of semester (medium-gray bars in Figure 2, E and F). Note that all three assessments were good at exposing students' difficulties in the five areas, which is useful for students and the instructor to know, so the problems can be fixed.

An assessment with a large portion of LOE responses is less useful for diagnostic purposes. The drug assessment showed the lowest prevalence of LOE responses (light-gray bars in Figure 2, C and D). The measurement of outcome area was most problematic for LOE on both the shrimp assessment and the bird assessment (light-gray bars in Figure 2, A, B, E, and F).

In general, looking across the three assessments, the areas variable property of an experimental subject and measurement of outcome were easier for most students at the end of the course than manipulation of variables, accounting for variability, or scope of inference. However, variable property of an experimental subject for the bird assessment was harder than for the shrimp and drug assessment. Also, the bird assessment did not probe well for measurement of outcome. Accounting for variability was slightly easier in the drug assessment than in the shrimp and bird assessment, perhaps because the drug assessment specifically probes for ways to deal with variability, like selecting a representative sample and randomized design of an experiment (Table 2, area of difficulty 4). A reason why accounting for variability was more difficult with the other assessments could be that the assessments did not guide students to address variability. Finally, it is interesting to note that scope of inference was problematic for students according to all three assessments, even though a slightly larger proportion of students demonstrated correct ideas in this area at the end of the course for all three assessments (Figure 2, A–F, row 5).

DISCUSSION

In summary, our study yielded the following major findings:

1. All established difficulties documented in our literature review (Table 1) were consistently found in responses from our own undergraduate biology students.
2. Data from our undergraduate biology students permitted the reclassification of one partially established difficulty, the variable property of experimental subject, to established.
3. Data collected from undergraduate biology students, together with difficulties data from a review of the literature, confirmed five major areas of difficulty with experimental design: 1) a property of an experimental subject that is variable; 2) manipulation of variables; 3) measurement of outcome; 4) accounting for and measuring variability; and 5) scope of inference of findings.
4. All the above data were used to inform the development of a rubric for experimental design, or RED, consisting of descriptions of correct ideas and typical difficulties within each of the above-mentioned five major areas.
5. The RED was shown to be an effective tool for detecting changes in undergraduate students' experimental design knowledge during instruction.

In response to RQ1, our comprehensive literature review (Table 1) summarized for the first time the full range of published experimental design difficulties and classified five categories and 13 subcategories of difficulty on a framework that told us whether they required further research or not in order to be fully identified. In fact, nearly all reported difficulties were confirmed to be fully established and therefore ready to be incorporated into our rubric. The one partially established difficulty, concerning variable property of experimental subjects, had previously been identified in only one study by Salangam (2007) with undergraduate biology students who were not science majors. We then reclassified this difficulty as established from data obtained when addressing RQ2, and thus we had a full complement of all the known difficulties for our rubric.

In addressing RQ2, we found that our undergraduate biology students demonstrated the full range of difficulties documented in Table 1, confirming the important need to address such difficulties in instruction. Indeed, we were concerned to find that several of the experimental design difficulties identified as long as 50 yr ago by Karplus (Fuller, 2002) still persist today among our students. In addition, a difficulty with scope of inference, previously reported by Chen and Klahr (1999) in a study involving elementary school–level students, was shown by us to persist as a problem at the undergraduate level. All the above findings convinced us of the important need to develop the RED, so it could serve as an important tool for assessing students in this crucial area of biological expertise while also informing intervention and remediation strategies.

To answer RQ3, we then used the RED in a pre/posttest comparison of experimental design knowledge and difficulties to find out whether it can be usefully deployed with published assessments to discriminate changes in knowledge during course participation. The RED was found to be useful with all three assessments. It further helped us organize the changes in student knowledge according to five areas of difficulty. The scoring process we used to discriminate changes before and after the course can be applied for practical purposes. Although we gathered hundreds of responses at the beginning and end of each semester from four cohorts, our random sample of 40 responses was sufficient to successfully demonstrate changes in students' knowledge. During scoring, for research purposes, we scored students for evidence of difficulties in an all-or-none manner. However, these assessments were low stakes and provided students a forum to express their ideas freely. Alternatively, an instructor might decide to assign partial credit to let students know where they stand on a continuum.

Once developed, the RED made it possible to evaluate the strengths and weaknesses of the three assessment instruments (Figure 2). For example, we now know that the bird assessment was more difficult for students in this study, perhaps because the context, ecological behavior, was not covered in this particular course (Clase *et al.*, 2010). In this study, prior knowledge, such as "competition among species," can lead students astray. Lack of knowledge about the context may also lead to LOE responses. An assessment with a high frequency of LOE responses could potentially be improved by providing background information, so all students designing an experiment start with the same contextual knowledge. We do not know whether students who show LOE with manipulation of variables in fact had difficulties and thus chose to not write much. Other areas with LOE problems on the pretest showed a decline in LOE for the posttest, indicating the problem may reflect how much students chose to write in their response rather than indicating a flawed probing design for the assessment. By more specifically probing for the LOE, as directed by the RED, students would be better prompted to reveal their knowledge. In contrast, the other two assessment instruments performed better than the bird instrument for the sample of biology students tested in this study. Now that we can use the RED to consistently grade student knowledge and to help students recognize and address their difficulties, it will be useful to gather a collection of assessments that specifically address each aspect of the RED.

An alternative explanation for why students struggle with identifying components of experimental design in an unfamiliar context could be that novice students, unlike experts, frequently have trouble identifying two problems as having the same theoretical features if the context is changed (Chi *et al.*, 1988). It is especially important to determine whether students are having trouble because they lack knowledge about experimental design concepts as defined in our glossary (see the Supplemental Material), or if they know about experiments but have trouble applying what they know in an unfamiliar context. In other words, certain features might allow students to call on particular knowledge about experiments in one domain, but they may have trouble transferring what they know to a completely different domain (Chen and Klahr, 1999; Barnett and Ceci, 2002; Sternberg, 2003). To resolve this uncertainty, more research is needed with additional experimental design assessments.

We envision the RED being potentially useful with a variety of existing assessment instruments, including the three used in the present study, for measuring progress from experiential learning in laboratory courses, research internships, or course-based undergraduate research experiences and not just in lecture courses like in the one in the current study. According to Laursen *et al.* (2010), undergraduate research experiences are often evaluated by faculty, and some "ask students to 'demonstrate their understanding of the processes of science' by framing a research question, developing a hypothesis, designing an experiment to test it, analyzing real data, writing a research report, and presenting their own work. These examples were sparse, and institutional evaluation efforts were often described as poorly developed or even perfunctory" (Laursen *et al.*, 2010, p. 176). The RED might be a useful guide for assessing experimental design-based assignments developed by faculty mentors who also consider the various components of experimental design appropriate for their local situation. Thus, to get a complete picture of student understanding of experimentation, multiple assessments should be applied to meet the RED criteria.

In considering the advantages that the RED brings to the issue of experimental design in the classroom, this rubric makes it possible to consistently diagnose and score student experimental design knowledge with different assessments. It can guide identification of student deficiencies and difficulties in certain aspects of experimental design, and these can reveal a need for new learning objectives, along with activities and remediation strategies to fix such deficiencies and difficulties. The RED can also be applied toward designing instructional strategies to alert both students and instructors as to pitfalls to avoid and areas in need of instruction to promote proficiency with experimental design. With information about student difficulties, the propositional statements of the RED can be of further use in helping target the problems with specific instruction based on practicing experimental design tasks. The RED helped us find useful information about our own students as we strive to teach students not just knowledge of the subject matter but how biology is performed as a research endeavor. Thus, the RED is useful for guiding all stages of learning, including objectives and instruction, in addition to assessment of experimental design.

Instructors who may want to use the RED could track their students' development of experimental design knowledge and abilities in a few different ways. Considering the RED difficulties ("Typical evidence of difficulties" column Table 2), an instructor could place examples for each difficulty from Table 3, plus examples found in the Supplemental Material (Supplemental Tables S4–6) or examples from his or her own students, in a scoring rubric. As examples for scoring a particular assessment, a table with difficulties from the shrimp assessment and drug assessment are posted online (http://tinyurl.com/REDShrimp and http://tinyurl.com/REDDrug). Instructors might create their own assessments, informed by the RED, and use them to examine the quality of their instruments. The RED outlines five major areas of difficulty, and, if an assessment fails to probe for a target area, the instructor could modify the directions to convert his or her own assessment into a more effective probe.

For the educational researcher, the RED can be used to guide and focus the design of educational research concerning experimental design and causal explanations, because the rubric details the components of experiments to consider. Thus, it can guide the coding of expert and novice explanations of experimental design, as well as the content analysis of textbook portrayals of experiments, and how those impact learning. For example, biology textbooks tend to show experiments with visualizations such as graphs. The three assessments used in the current study had no visualizations, which was a limitation. One way for an educational researcher to understand whether experts differ from students in their knowledge about experimental design could be to have them visualize the concepts of their experimental design with graphs. A graph might help students organize their approach to using experimental design concepts. Visuals such as graphs might represent the five areas of experimental design difficulties from the RED in a visual form. For instance, instructors can alert their students that the experimental subject is typically stated in the graph legend (Table 2, area of difficulty 1), the *x*-axis represents the treatment variables (area of difficulty 2) and the *y*-axis generally shows the measurable outcomes (area of difficulty 3). Students can also be alerted to graphically make attempts to represent the variation (area of difficulty 4), say in the form of error bars, and to the need, when interpreting a graph, to consider the sample, the controls, treatment and outcome variables, and to explain the extent to which claims can be inferred for a given experiment (area of difficulty 5).

With the RED to diagnose experimental design difficulties, future research can target specific difficulties with interventions to teach beginner researchers what to do and what not to do by using graphs or other drawings to focus their attention on each of the five component areas in Table 2. Clearly, much work remains to be done to help biology students understand research to meet academic standards and to gain a competitive employment edge upon graduation. We suggest that biologists might use the RED as a framework based on empirical evidence to guide beginner researchers to develop competence in experimental design.

ACKNOWLEDGMENTS

The shrimp and bird assessments were developed by the College Board; for the drug assessment, SRI International is the publisher and the copyright owner. We are grateful to the College Board and SRI International for granting permission to use their published assessments. We thank David Eichinger, Jeff Karpicke, and Dennis Minchella for insightful comments and suggestions and Samiksha Neroorkar, Rethabile Tekane, Caleb Trujillo, and other members of our research groups PIBERG (Purdue International Biology Education Research Group) and VIBE (Visualizations in Biochemistry Education) for their contributions to the progress of our study. We also thank our reviewers for their valuable feedback to help improve an early draft of the article. Some material presented here is based on research supported by the National Science Foundation under grant #0837229. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

American Association for the Advancement of Science (2010). Vision and Change: A Call to Action, Washington, DC.

Association of American Colleges and Universities (2013). Scientific Thinking and Integrative Reasoning Skills, Washington, DC.

Association of American Medical Colleges and the Howard Hughes Medical Institute (2009). Report of Scientific Foundations for Future Physicians Committee, Washington, DC.

Barnett SM, Ceci SJ (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychol Bull *128*, 612.

Beck CW, Blumer LW (2012). Inquiry-based ecology laboratory courses improve student confidence and scientific reasoning skills. Ecosphere *3*, 112.

Box GE, Hunter JS, Hunter WG (2005). Statistics for Experimenters: Design, Innovation, and Discovery, 2nd ed., Hoboken, NJ: Wiley.

Bullock M, Ziegler A (1999). Scientific reasoning: developmental and individual differences. In: Individual Development from 3 to 12, ed. FE Weinert and W Schneider, Cambridge, UK: Cambridge University Press, 38–54.

Burns JC, Okey JR, Wise KC (1985). Development of an integrated process skill test: TIPS II. J Res Sci Teach *22*, 169–177.

Chen Z, Klahr D (1999). All other things being equal: acquisition and transfer of the control of variables strategy. Child Dev *70*, 1098–1120.

Chi MT, Glaser RE, Farr MJ (1988). The Nature of Expertise, Hillsdale, NJ: Erlbaum.

Clase KL, Gundlach E, Pelaez NJ (2010). Calibrated peer review for computer-assisted learning of biological research competencies. Biochem Mol Biol Educ *38*, 290–295.

Cohen JA (1960). Coefficient of agreement for nominal scales. Educ Psychol Meas *20*, 37–46.

College Board (2006). AP Statistics Free-Response Questions. http://apcentral.collegeboard.com/apc/public/repository/_ap06_frq_statistics_51653.pdf (accessed 20 December 2013).

College Board (2009). AP Statistics Free-Response Question Form B. http://apcentral.collegeboard.com/apc/public/repository/ap09_frq_statistics_formb.pdf (accessed 20 December 2013).

Colon-Berlingeri M, Burrowes PA (2011). Teaching biology through statistics: application of statistical methods in genetics and zoology courses. CBE Life Sci Educ *10*, 259–267.

Cox DR, Reid R (2000). The Theory of the Design of Experiments, Boca Raton, FL: Chapman and Hall/CRC.

D'Costa AR, Schlueter MA (2013). Scaffolded instruction improves student understanding of the scientific method and experimental design. Am Biol Teach *75*, 18–28.

Dolan E, Grady J (2010). Recognizing students' scientific reasoning: a tool for categorizing complexity of reasoning during teaching by inquiry. J Sci Teach Educ *21*, 31–55.

Duschl RA, Schweingruber HA, Shouse AW (2007). Taking Science to School: Learning and Teaching Science in Grades K-8, Washington, DC: National Academies Press.

Feldon DF, Timmerman BC, Stowe KA, Showman R (2010). Translating expertise into effective instruction: the impacts of cognitive task analysis (CTA) on lab report quality and student retention in the biological sciences. J Res Sci Teach *47*, 1165–1185.

Fuller RG (ed.) (2002). A Love of Discovery: Science Education—The Second Career of Robert Karplus, New York: Kluwer.

Gill J, Walsh J (2010). Control group. In: Encyclopedia of Research Design, ed. N Salkind, Thousand Oaks, CA: Sage.

Gormally C, Brickman P, Lutz M (2012). Developing a Test of Scientific Literacy Skills (TOSLS): measuring undergraduates' evaluation of scientific information and arguments. CBE Life Sci Educ *11*, 364–377.

Grayson DJ, Anderson TR, Crossley LG (2001). A four-level framework for identifying and classifying student conceptual and reasoning difficulties. Int J Sci Educ *23*, 611–622.

Griffith AB (2007). Semester-long engagement in science inquiry improves students' understanding of experimental design. Teach Issues Exp Ecol *5*, Research #2. www.ecoed.net/tiee/vol/v5/research/griffith/pdf/Griffith_2007.pdf (assessed 3 May 2014).

Grunwald S, Hartman A (2010). A case-based approach improves science students' experimental variable identification skills. J Coll Sci Teach *39*, 28–33.

Gutwill-Wise J (2001). The impact of active and context-based learning in introductory chemistry courses: an early evaluation of the modular approach. J Chem Ed *78*, 684–690.

Harker AR (2009). Full application of the scientific method in an undergraduate teaching laboratory: a reality-based approach to experiential student-directed instruction. J Coll Sci Teach *29*, 97–100.

Hiebert SM (2007). Teaching simple experimental design to undergraduates: do your students understand the basics. Adv Physiol Educ *31*, 82–92.

Holmes D, Moody P, Dine D (2011). Research Methods for the Biosciences, Oxford, UK: Oxford University Press.

Kanari Z, Millar R (2004). Reasoning from data: how students collect and interpret data in science investigations. J Res Sci Teach *41*, 748–769.

Kardash CM (2000). Evaluation of an undergraduate research experience: perceptions of undergraduate interns and their faculty mentors. J Educ Psychol *92*, 191–201.

Kish L (1965). Survey Sampling, New York: Wiley.

Klahr D, Fay AL, Dunbar K (1993). Heuristics for scientific experimentation: a developmental study. Cogn Psychol *25*, 111–146.

Kloser M, Brownell S, Shavelson R, Fukami T (2013). Effects of a research-based ecology lab course: a study of nonvolunteer achievement, self-confidence, and perception of lab course purpose. J Coll Sci Teach *42*, 72–81.

Koehler DJ (1994). Hypothesis generation and confidence in judgment. J Exp Psychol Learn *20*, 461–469.

Kuhn D, Dean D (2005). Is developing scientific thinking all about learning to control variables. Psychol Sci *16*, 866–870.

Kuhn D, Pearsall S (2000). Developmental origins of scientific thinking. J Cogn Dev *1*, 113–129.

Kuhn D, Schauble L, Garcia-Mila M (1992). Cross-domain development of scientific reasoning. Cogn Instr *9*, 285–327.

Landis JR, Koch GG (1977). The measurement of observer agreement for categorical data. Biometrics *33*, 159–174.

Laursen S, Hunter AB, Seymour E, Thiry H, Melton G (2010). Undergraduate Research in the Sciences: Engaging Students in Real Science, San Francisco: Jossey-Bass.

Lawson AE, Drake N, Johnson J, Kwon YJ, Scarpone C (2000). How good are students at testing alternative explanations of unseen entities? Am Biol Teach *62*, 249–255.

Lawson AE, Snitgen DA (1982). Teaching formal reasoning in a college biology course for preservice teachers. J Res Sci Teach *19*, 233–248.

Libarkin J, Ording G (2012). The utility of writing assignments in undergraduate bioscience. CBE Life Sci Educ *11*, 39–46.

Lopatto D (2003). The essential features of undergraduate research. Counc Undergrad Res Q *24*, 139–142.

Lopatto D (2004). Survey of Undergraduate Research Experiences (SURE): first findings. Cell Biol Educ *3*, 270–277.

Lopatto D (2007). Undergraduate research experiences support science career decisions and active learning. CBE Life Sci Educ *6*, 297–306.

Lopatto D (2008). Exploring the benefits of undergraduate research: the SURE survey. In: Creating Effective Undergraduate Research Programs in Science, ed. R Taraban and RL Blanton, New York: Teacher's College Press, 112–132.

Metz AM (2008). Teaching statistics in biology: using inquiry-based learning to strengthen understanding of statistical analysis in biology laboratory courses. CBE Life Sci Educ *7*, 317–326.

National Institute of Standards and Technology/SEMATECH (2003). Engineering Statistics Handbook, Section 3.1.3.6. www.itl.nist.gov/div898/handbook/ppc/section1/ppc136.htm (accessed 20 December 2013).

National Research Council (2007). Rising above the gathering storm: energizing and employing America for a brighter economic future, Washington, DC: National Academies Press.

Park J, Pak S (1997). Students' responses to experimental evidence based on perceptions of causality and availability of evidence. J Res Sci Teach, ***34***, pp. 57–67.

Picone C, Rhode J, Hyatt L, Parshall T (2007). Assessing gains in undergraduate students' abilities to analyze graphical data. Teach Issues Exp Ecol *5*, Research #1. www.esa.org/tiee/vol/v5/research/picone/article.html (accessed 20 December 2013).

Purdue University (2010). Second Purdue Symposium on Psychological Sciences—Psychology of Science: Implicit and Explicit Processes. http://www.purdue.edu/hhs/psy/research/symposium/2ndPSPS2010.php (accessed 20 December 2013).

Quinn GGP, Keough MJ (2002). Experimental Design and Data Analysis for Biologists, Cambridge, UK: Cambridge University Press.

Ramsey FL, Schafer DW (2012). The Statistical Sleuth: A Course in Methods of Data Analysis, Boston: Brooks/Cole.

Roth WM, McGinn MK, Bowen GM (1998). How prepared are preservice teachers to teach scientific inquiry? Levels of performance in scientific representation practices. J Sci Teach Educ *9*, 25–48.

Rubin DB (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics *29*, 185–203.

Ruxton GD, Colegrave N (2006). Experimental Design for the Life Sciences, Oxford, UK: Oxford University Press.

Salangam J (2007). The impact of a prelaboratory discussion on non-biology majors abilities to plan scientific inquiry. Master's Thesis, Fullerton: California State University.

Schauble L (1990). Belief revision in children: the role of prior knowledge and strategies for generating evidence. J Exp Child Psychol *49*, 31–57.

Schauble L (1996). The development of scientific reasoning in knowledge-rich contexts. Dev Psychol *32*, 102–119.

Schauble L, Glaser R (1990). Scientific thinking in children and adults. Contrib Hum Dev *21*, 9–27.

Shi J, Power JM, Klymkowsky MW (2011). Revealing student thinking about experimental design and the roles of control experiments. Int J Sch Teach Learn *5*, 1–16.

Singer S, Hilton M, Schweingruber H (2006). America's Lab Report: Investigations in High School Science, Washington, DC: National Academies Press.

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

SRI International (2003). Diffusion and Testing a New Drug 1. http://pals.sri.com/tasks/9-12/Testdrug/ (accessed 20 December 2013).

Sternberg RJ (2003). What is an "expert student?" Educ Res *32*, 5–9.

Thiry H, Weston TJ, Laursen SL, Hunter AB (2012). The benefits of multi-year research experiences: differences in novice and experienced students' reported gains from undergraduate research. CBE Life Sci Educ *11*, 260–272.

Timmerman B, Strickland DC, Johnson RL, Payne JR (2011). Development of a "universal" rubric for assessing undergraduates' scientific reasoning skills using scientific writing. Assess Eval High Educ *36*, 509–547.

Tobin KG, Capie W (1981). The development and validation of a pencil and paper test of logical thinking. Educ Psychol Measurement *41*, 413–424.

Tobin KG, Capie W (1982). Relationships between classroom process variables and middle-school science achievement. J Educ Psychol *74*, 441.

Wei CA, Woodin T (2011). Undergraduate research experiences in biology: alternatives to the apprenticeship model. CBE Life Sci Educ *10*, 123–131.

Wuensch KL (2001). When Does Correlation Imply Causation? http://core.ecu.edu/psyc/wuenschk/stathelp/Correlation-Causation.htm (accessed 20 December 2013).