CBE—Life Sciences Education
Vol. 14, 1–11, Winter 2015

Article

The Laboratory Course Assessment Survey: A Tool to Measure Three Dimensions of Research-Course Design

Lisa A. Corwin,* Christopher Runyon,† Aspen Robinson,‡ and Erin L. Dolan*

*Texas Institute for Discovery Education in Science, College of Natural Sciences, and †Department of Educational Psychology, University of Texas at Austin, Austin, TX 78712; ‡Department of Psychology, University of North Carolina at Charlotte, Charlotte, NC 28223

Submitted March 25, 2015; Revised June 24, 2015; Accepted July 22, 2015
Monitoring Editor: Graham F. Hatfull

Course-based undergraduate research experiences (CUREs) are increasingly being offered as scalable ways to involve undergraduates in research. Yet few if any design features that make CUREs effective have been identified. We developed a 17-item survey instrument, the Laboratory Course Assessment Survey (LCAS), that measures students' perceptions of three design features of biology lab courses: 1) collaboration, 2) discovery and relevance, and 3) iteration. We assessed the psychometric properties of the LCAS using established methods for instrument design and validation. We also assessed the ability of the LCAS to differentiate between CUREs and traditional laboratory courses, and found that the discovery and relevance and iteration scales differentiated between these groups. Our results indicate that the LCAS is suited for characterizing and comparing undergraduate biology lab courses and should be useful for determining the relative importance of the three design features for achieving student outcomes.

INTRODUCTION

National efforts aimed at improving undergraduate science education call for the involvement of undergraduates in research (National Research Council, 2003; American Association for the Advancement of Science, 2011; President's Council of Advisors on Science and Technology, 2012). These calls result from a growing body of evidence that undergraduates benefit from engaging in research (reviewed in Laursen *et al*., 2010; Lopatto, 2010; Corwin *et al*., 2015). Students who participate in research internships report cognitive gains such as the development of knowledge and skills (Kardash, 2000); affective gains, such as satisfaction with their research experience (Thiry and Laursen, 2011); psychosocial gains, such as feeling like a scientist (Thiry and Laursen, 2011; Adedokun *et al*., 2012); and conative gains, such as increased perseverance in the face of obstacles (Lopatto, 2007). Additionally, a number of studies document that undergraduates who participate in research internships persist in science, though it is unclear whether these students persist because of self-selection into these experiences or because of the experience itself (Schultz *et al*., 2011; Eagan *et al*., 2013; Linn *et al*., 2015).

Course-based undergraduate research experiences (CUREs), which involve groups of students in addressing research problems or questions in the context of a class, have been proposed as scalable ways to involve undergraduates in research. CUREs offer other advantages beyond scalability. First, they are accessible to students early in their undergraduate careers when they have greater potential to influence a student's academic and career trajectory (Dolan *et al*., 2008; Lopatto *et al*., 2008; Wei and Woodin, 2011). Second, many CUREs are open to all students who enroll in a course and thus have the potential to involve students who might not otherwise have access to science research opportunities (Wei and Woodin, 2011; Bangera and Brownell, 2014). Finally, a handful of studies indicate that CURE students report many of the same outcomes as students who participate in research internships, including increased knowledge, improved research skills, increases in research self-efficacy, and greater clarity regarding their career choices (reviewed in Corwin *et al*., 2015).

CBE Life Sci Educ December 1, 2015 14:ar37
DOI:10.1187/cbe.15-03-0073
Address correspondence to: Lisa A. Corwin (lisa.c.a@utexas.edu).

Although questions have been raised about whether enough is known about the efficacy and impact of CUREs and research experiences in general (Linn *et al*., 2015), the potential benefits of CUREs are driving widespread development and implementation of this pedagogy. As with any educational intervention, the design features that make it effective for students need to be identified and maintained to ensure its continued efficacy (Borrego *et al*., 2013). Corwin Auchincloss and colleagues (2014) proposed the following set of design features, or "dimensions," which may contribute to the efficacy and impact of CUREs:

- Collaboration: The degree to which students are encouraged to work together, help each other, build off one another's work, and provide and respond to feedback.
- Discovery: The degree to which students have opportunities to generate new scientific knowledge.
- Broad relevance: The degree to which students' work is of interest to a community beyond the classroom, which can manifest as authorship on a scientific paper or presentations or reports to stakeholders.
- Iteration: The degree to which students have opportunities to revise or repeat aspects of their work to fix problems, improve validity of their own and others' results, understand variation in data, or further test hypotheses.
- Use of science practices: The degree to which students engage in asking questions, building and evaluating models, proposing hypotheses, designing studies, selecting methods, gathering and analyzing data, and developing and critiquing interpretations and arguments. Students are likely to engage in several but not all scientific practices during a single CURE.

We describe here the development and validation of a survey instrument, which we call the Laboratory Course Assessment Survey (LCAS), intended to measure students' perceptions of these design features in biology lab courses. We chose to focus on measuring how students perceived course design rather than how instructors intended to design their courses, because students' perceptions are likely vary within a single course offering in ways that relate to the outcomes they realize from the course (Prosser and Trigwell, 1999). Through the use of established methods for instrument design and validation, we produced a 17-item survey that we anticipate will be useful to instructors and researchers for characterizing CUREs and for linking particular CURE design features to specific student outcomes (Corwin *et al*., 2015). We also present data showing that the LCAS is useful for distinguishing between biology CUREs and traditional biology lab courses.

METHODS

Participants

The participants in this study were 213 undergraduate students enrolled in 16 different biology laboratory courses at 11 colleges and universities in the United States. Students were recruited through emails forwarded to them by the instructors of their laboratory courses. These courses included both upper-division and introductory courses, and students in the courses included freshmen through seniors. Instructors were recruited from the authors' personal networks and from the CUREnet website (https://curenet.cns.utexas.edu). All participants received an email inviting them to participate in the study by completing an online survey. A $10 gift card was offered as an incentive. All responses were collected within a year of students completing their laboratory course. The study was conducted with approval from the institutional review boards at the two institutions from which the authors conducted the study, the University of Georgia (STUDY00000793) and the University of Texas at Austin (2014-07-0028). This approval was deemed sufficient by individuals at participating institutions. Out of the 213 participants, 187 completed at least 80% of the survey. All exploratory factor analyses (EFAs) described below were conducted only on surveys that were at minimum 80% complete to minimize imputation error. The demographics of our student sample show low representation of black and Hispanic/Latino(a) students and higher representation of Asian students, with the highest representation being white students (Table 1). We had no representation of Alaska Native

Table 1. Student demographic information[a]

	Total sample	EFA sample	Whole instrument comparison	Collaboration scale comparison	Iteration scale comparison	Discovery/ relevance scale comparison	National % graduates with biology majors
Sample size	212	187	115	141	134	133	100
Men	60 (28.3%)	58 (31.1%)	37 (32.2%)	43 (30.5%)	42 (31.3%)	41 (30.8%)	40.4
Women	128 (60.4)	125 (66.8)	76 (66.1)	96 (68.1)	90 (67.2)	90 (67.7)	59.6
Not reported	24 (11.3)	4 (2.1)	2 (1.7)	2 (1.4)	2 (1.5)	2 (1.5)	n/a
White	85 (40.1)	84 (44.9)	54 (47.0)	68 (48.2)	65 (48.5)	65 (48.9)	58.4
Hispanic/Latino(a)	13 (6.1)	12 (6.4)	5 (4.3)	8 (5.7)	6 (4.5)	6 (4.5)	8.9
Black	18 (8.5)	18 (9.6)	11 (9.6)	13 (9.2)	14 (10.4)	14 (10.5)	7.1
Asian	60 (28.3)	60 (32.1)	42 (36.5)	46 (32.6)	46 (34.3)	45 (33.8)	15.7
Other	7 (3.3)	7 (3.7)	1 (0.9)	1 (0.7)	1 (0.7)	1 (0.8)	7.0
Not reported	29 (13.7)	6 (3.2)	2 (1.7)	5 (3.5)	2 (0.7)	2 (1.5)	2.9

[a]Participants who responded to at least 80% of the LCAS scale items were included in the EFA. Only students who could be clearly identified as being part of a traditional lab or CURE lab and who had no missing responses were included in the comparisons. Nonparenthetical values indicate absolute numbers of respondents. Parenthetical values indicate percent of each sample.

or Native American students. Our sample contains roughly twice as many female as male participants.

Overview of the Development and Validation Process

Dimensionality, reliability, and validity are established concepts in the development and use of survey instruments. Dimensionality refers to the number of underpinning constructs represented by a set of survey items. We assessed dimensionality of the LCAS using EFA, which examines similarity and dissimilarity among survey items based on the covariance among item responses. Two or more items that consistently elicit similar response patterns, or persistently covary, are likely to describe the same construct or dimension. Groups of items that persistently covary and measure a single construct are called scales.

Reliability refers to the degree of interrelatedness of an item to all other items in the same scale, or the internal consistency. It also refers to the consistency of this interrelatedness across many administrations of an instrument in similar conditions and with a similar population (Bachman, 2004; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, and NCME], 2014). We used Cronbach's alpha (Cronbach, 1951) to assess the baseline internal consistency within each scale of the LCAS and thus for each construct that we propose is being measured by the LCAS (Netemeyer *et al.*, 2003).

Validity refers to the evidence that a scale is measuring what it was intended to measure (AERA, APA, and NCME, 2014). We used several approaches to demonstrate validity for the LCAS. We tested content and face validity by conducting cognitive interviews to ensure the items were clear, transparent, and meaningful to students and instructors from undergraduate biology courses and that they perceived the fundamental features of their courses were represented in the items. We tested construct validity by examining the relationship between the research and theory that informed the scale's development and the data that resulted from testing it. We also tested construct validity of the LCAS to determine whether it differentiates between groups whose responses can be expected to differ, in particular, between students in CURE versus traditional lab courses (Netemeyer *et al.*, 2003). We describe the specific methods for data collection and analysis below.

Data Collection

Development and validation of the LCAS followed the three-stage process (substantive, structural, and external stages) detailed by Benson (1998; Figure 1). During the substantive stage, we determined and refined the substance or topic of the survey instrument. Specifically, we designed the LCAS to assess five dimensions presented above: collaboration, discovery, broad relevance, iteration, and use of science practices (Corwin Auchincloss *et al.*, 2014). This is a form of content validity, with each dimension comprising a separate construct and thus testable with a separate scale. We wrote original Likert-like survey items hypothesized to indicate the presence of each dimension. We performed cognitive interviews with three instructors of self-designated CUREs (two white males, one white female), two instructors of traditional (non-CURE) laboratory courses (one white male, one white female), three students of CUREs (two white females, one white male), and three students of traditional lab courses (two black females, one Asian male).

We confirmed whether courses were CUREs by asking instructors and students to describe during the cognitive interviews what happens in their course. For the purposes of this study, a CURE was defined as "a course that involves a research experience with the potential to produce results that are of interest to a scientific community." A traditional lab course was defined as "a course in which students perform experiments and investigations that are 'good exemplars of phenomena' and in which the 'correct' results of each investigation are known prior to execution" (adapted from Brownell and Kloser, 2015). During the interviews, we asked participants to respond to each item to allow us to evaluate whether our items were comprehensible and to determine whether items prompted responses relevant to

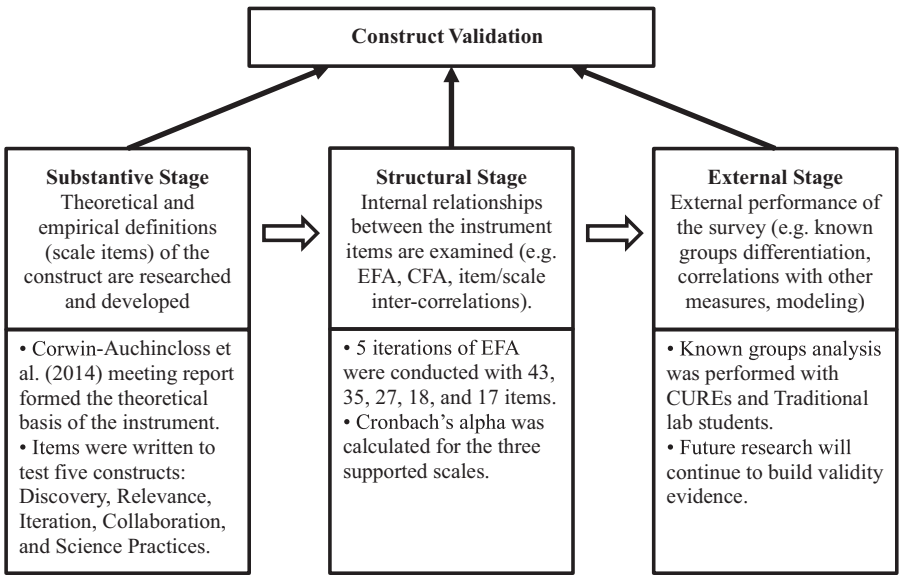

Figure 1. Description of Benson's (1998) construct validation framework and the corresponding steps used to develop and validate the LCAS.

our survey aims (Bachman, 2004). We revised the wording and order of the questions based on the interviews and eliminated unclear and ambiguous questions. Only questions that were clear and interpreted as the authors intended by both instructors and students were retained. This resulted in a total of 55 survey items on the pilot survey to be given to students. We then entered the survey into Qualtrix online survey software to allow for data collection from the participants described above ($N = 213$).

Data Analysis

After data collection was complete, we assessed the dimensionality, reliability, and validity of the survey instrument, consistent with Benson's (1998) structural and external stages of development. First, we assessed the psychometric properties and dimensionality (structure) of the instrument. We started by determining the potential utility of each of the original 55 survey items by assessing its response rate. We also examined a polychoric correlation matrix, including all items, to ensure they correlated with at least two other survey items and had potential to contribute to at least one survey scale. We removed items with low response rates and items lacking strong correlations with other items (described in detail in *Results*). We then performed iterative EFA to assess the dimensionality of the instrument and describe relationships among items. We characterized three useful scales as a result of this analysis and established the internal consistency (a measure of reliability) of each scale (Cronbach, 1951).

In the external stage, an instrument is applied in research to understand how it functions in certain populations or how resulting data relate to data collected using similar or dissimilar instruments. For this stage, we administered the survey to "known groups"—students whom we expected to have different average ratings for each survey scale. Specifically, we compared survey responses from students in CUREs versus traditional lab courses by using *t* tests. CUREs and traditional labs were identified based on qualitative content analysis of two data sources: descriptions of courses by 1) students and 2) instructors. We describe the structural and external stages in more detail in the *Results* to illustrate how we used results of different analyses to inform the development and validation of the LCAS.

RESULTS

Refinement of Scales

The original instrument was composed of 55 new Likert-type items intended to measure five constructs of interest: collaboration, discovery, relevance, iteration, and use of science practices. (This initial item bank is available upon request from L.A.C.) Upon survey administration, we noted that not all items appeared to contribute to assessment of the intended constructs. Thus, we conducted preliminary analyses to systematically remove nonuseful items. First, we examined item response rates. We considered responses of "I don't know" or "I prefer not to respond" to be nonresponses, and we removed four items with a response rate of 60% or lower. When we examined these items, it was clear that missing responses were not missing at random. All of these items asked the students to indicate the frequency of *other* students engaging in certain activities likely to take place outside the classroom. It is probable that students did not have access to this information and could not accurately respond to these items. Second, we examined a polychoric correlation matrix of all remaining items to identify items with low interitem correlations. We removed five items with consistently low interitem correlations (coefficients <0.3 for at least 80% of correlations) because of their low potential to contribute to measurement of underlying constructs as part of a cohesive multi-item scale. We identified and removed three other ambiguous items. Each of these items asked students about the time they had available during the course to complete a task, ending with "as originally planned" (e.g., "In this course I had time to collect and analyze data as originally planned"). Because we had no way of knowing what instructors originally planned, these questions were uninformative. At the end of this process, 43 items remained.

We conducted several iterations of EFA with the remaining 43 items to arrive at our final 17-item scale (Supplemental Material Appendix A). We conducted all EFAs with Mplus 7.2 (Muthén and Muthén, 1998–2012), using weighted least-squares means and variances adjusted estimation. This method of estimation is appropriate with categorical data and nonnormal data, as is the case with the present data (Finney and DiStefano, 2006). We used a GEOMIN (oblique) rotation, as we hypothesized the factors would correlate with one another. Missing responses (3.17% of the data) followed an arbitrary pattern, and none of the remaining items had a comparatively high number of missing responses. Thus, we assumed data were missing at random for the remaining scale items and used full-information maximum likelihood (FIML) estimation to impute missing values. FIML has been shown to be a valid method of accounting for missing data that does not bias the subsequent analyses when data are assumed to be missing at random (Schafer and Graham, 2002; Enders, 2010; Little *et al.*, 2013). For each factor analysis, we allowed a maximum of 10 underlying factors in order to inspect a wide range of models. We did not expect results with more than five factors to be useful, but we examined these solutions to allow for unanticipated results.

Our initial EFA yielded a four-factor solution and revealed that eight items written to assess participation in science practices were not useful. Two items did not clearly load on any factors, and six items loaded on multiple factors (had loadings >0.35 for more than one factor). This is not surprising, considering that many science practices are associated with the other constructs of interest (e.g., analyzing data may be associated with discovery or iteration). We concluded that "participation in science practices" itself does not constitute a single independent construct that could be completely and accurately measured using our approach. Thus, these eight items, which constituted the dimension "science practices," were removed from the survey. After this, we focused on assessing only the four remaining constructs of interest: collaboration, discovery, relevance, and iteration.

We originally wrote nine items to be reverse scored to measure traditional lab course practices and to contrast with items that represent common practices in CUREs. For example, "I was expected to conduct an investigation to

find an anticipated result" was hypothesized to contrast with the discovery item "I was expected to conduct an investigation to find something previously unknown to myself, other students, and the instructor." However, responses to these items indicated that students did not see these traditional lab practices as opposites of CURE practices, making reverse scoring inappropriate. In addition, traditional lab items loaded together on a single, indistinct factor with many of the science practice items. Thus, we removed these nine items from the analysis, resulting in a measure that included 27 positively worded and scored items.

We conducted two more iterations of EFA with the remaining 27 items to identify and remove items that did not contribute further to the functioning of the instrument. Through these analyses, we identified 17 items that consistently factored onto three factors, loaded strongly onto their designated factors (>0.42), and were not redundant with other items in the same scale (i.e., their wording was substantially different from other items). We excluded items that did not meet these criteria. We conducted a final EFA with the 17 remaining items. We selected a final three-factor solution, because the items all had strong factor loadings and represented three theoretically meaningful factors (see *Comparison of CUREs and Traditional Lab Courses* in the *Discussion*). After choosing the solution, we consulted the results of a parallel analysis (Horn, 1965; Hayton *et al.*, 2004), the acceleration factor and optimal coordinates (Raîche *et al.*, 2013), and the Kaiser (1960) rule (eigenvalue greater than one) for additional support for determining the number of factors. Each of these analytical tests suggested that three factors were most appropriate for the data. We corroborated this solution by examining the eigenvalues and resulting scree plot (Figure 2). Interitem correlations, means, SDs, and scales of the final 17 items can be found in Table 2. Factor loadings for the final instrument were consistently above 0.4 within each scale, higher than the suggested minimum cutoff of 0.32 (Tabachnick and Fiddell, 2001). Cronbach's alpha for each scale was calculated with the result of $\alpha > 0.8$ for all scales (Nunnally 1978; Lance *et al.*, 2006; Table 3).

The final three-factor solution of the LCAS consists of:

1. Collaboration. This factor consists of six items that ask students to evaluate the frequency with which they engaged in activities related to collaboration (e.g., provide help to other students, discuss work with other students, or critique other students' work) and metacognition (e.g., "reflect on what I was learning"). The response options are 1, weekly; 2, monthly; 3, one or two times; 4, never.
2. Discovery and relevance. The second factor consists of five items that ask students to rate their agreement with statements about whether their lab work could lead to discovery of something new, development of new arguments, or generation of information of interest to the scientific community. The response options range from 1, strongly disagree, to 6, strongly agree.
3. Iteration. The third factor consists of six items that ask students to rate their agreement with statements about whether they had time or direction to repeat aspects of their work, such as making revisions, changing methods, and analyzing additional data. The response options ranged from 1, strongly disagree, to 6, strongly agree.

Distinguishing between CUREs and Traditional Lab Courses

We designed the LCAS to measure the design features or dimensions that make CUREs distinctive as learning experiences, based on input from experts in undergraduate research and thorough review of research on these experiences (Corwin Auchincloss *et al.*, 2014). Thus, the LCAS already has some degree of construct validity. To further assess its validity, we compared the ratings of students in traditional lab courses with ratings of CURE students for the entire survey instrument and for each scale (Benson and Hagtvet, 1996; Netemeyer *et al.*, 2003). In comparing these known groups, we hypothesized that CURE students would report higher levels of most, if not all, of the measured constructs compared with students in traditional lab courses.

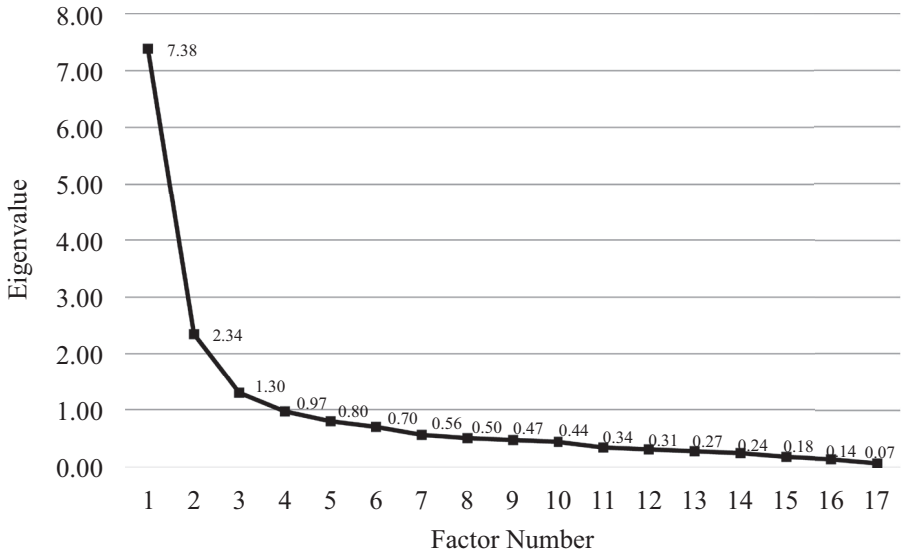

Figure 2. Scree plot of eigenvalues. Plotted points represent the eigenvalues for each added factor.

Table 2. Polychoric correlation matrix for the LCAS[a]

	Item number																
	C1	C2	C3	C4	C5	C6	DR1	DR2	DR3	DR4	DR5	I1	I2	I3	I4	I5	I6
C1	—																
C2	0.62	—															
C3	0.61	0.37	—														
C4	0.41	0.52	0.52	—													
C5	0.50	0.41	0.64	0.64	—												
C6	0.59	0.55	0.81	0.70	0.61	—											
DR1	0.20	0.33	0.33	0.27	0.32	0.26	—										
DR2	0.17	0.21	0.29	0.25	0.29	0.21	0.62	—									
DR3	0.22	0.18	0.31	0.20	0.51	0.24	0.42	0.34	—								
DR4	0.17	0.35	0.29	0.30	0.46	0.30	0.55	0.38	0.67	—							
DR5	0.19	0.29	0.23	0.35	0.39	0.24	0.77	0.53	0.46	0.55	—						
I1	0.08	0.15	0.20	0.34	0.39	0.21	0.29	0.21	0.37	0.43	0.46	—					
I2	0.21	0.20	0.33	0.25	0.38	0.29	0.51	0.39	0.39	0.42	0.46	0.45	—				
I3	0.26	0.23	0.40	0.48	0.39	0.40	0.32	0.25	0.32	0.45	0.38	0.49	0.51	—			
I4	0.24	0.40	0.34	0.35	0.40	0.36	0.39	0.38	0.39	0.50	0.54	0.55	0.63	0.47	—		
I5	0.31	0.41	0.40	0.44	0.43	0.42	0.45	0.37	0.36	0.46	0.50	0.63	0.61	0.46	0.67	—	
I6	0.10	0.17	0.39	0.23	0.42	0.24	0.39	0.38	0.36	0.42	0.40	0.59	0.55	0.43	0.59	0.65	—
Mean	**3.73**	**3.72**	**3.66**	**3.25**	**3.05**	**3.51**	**4.21**	**4.39**	**5.07**	**4.70**	**4.25**	**4.71**	**4.32**	**5.14**	**4.54**	**4.59**	**4.64**
SD	0.67	0.67	0.69	1.04	1.07	0.85	1.49	1.50	1.14	1.24	1.42	1.11	1.31	1.03	1.19	1.29	1.41
Response scale	4	4	4	4	4	4	6	6	6	6	6	6	6	6	6	6	6

[a]$n = 176$ (complete cases only). C refers to a collaboration scale item, DR to a discovery and relevance scale item, and I to an iteration scale item; see Table 3 for specific item codes. Scale indicates the number of response options available for that item.

To conduct this comparison, we first designated courses as CUREs or traditional lab courses. This designation was based on two data sources: student descriptions of their courses and instructor descriptions of the same courses. Students and instructors were asked to respond to three questions about their courses: 1) "In a few sentences, please describe the lab course you are responding about"; 2) "Please list up to 5 words or phrases that describe what makes this course unique or distinctive compared to other lab courses you have taken/taught, other than science topics (for example, not Microbiology)"; and 3) "Please list up to 5 words or phrases you would use, other than science topics, to describe what you/your students did in the course." Students and instructors were blind to one another's responses on all questions.

We used the same criteria to analyze student and instructor responses. Specifically, two authors (L.A.C. and E.L.D.) read student responses and instructor responses separately and coded each course as a CURE, a traditional lab course, or ambiguous. We designated as a CURE any course described using terms that reflected a research component with potential to produce results of interest to a community outside the classroom. Relevant terms included "research," "inquiry," "real," "novel," "new," and "publishable," as well as some indicator of the importance of the results beyond the classroom. We designated as a traditional lab any course described as a series of discrete "labs" or "exercises" with a focus primarily on mastery of specific content or techniques. These courses rarely mentioned "inquiry" or "research" and did not include any multiweek investigations. When we were not able to definitively identify a course as a CURE or traditional lab course, we designed the course as "ambiguous" and excluded the associated survey responses from subsequent analyses. We calculated a Cohen's kappa of 0.897 as a measure of intercoder reliability, which is considered very good by conventional standards (Landis and Koch, 1977; Fleiss, 1981). We also calculated a Cohen's kappa of 0.794 as a measure of student and instructor agreement, which is considered good by conventional standards. In the two cases lacking student/instructor agreement, we examined laboratory manuals as a third form of evidence to determine course type. The laboratory manuals provided evidence that both courses were traditional, which aligned with designations based on student responses but did not align with the designation of "ambiguous" based on instructors' responses. Through this process, we were able to clearly identify 60 students in CURE courses and 55 students in traditional courses ($n = 115$) who had complete records for the full 17-item instrument.

We performed preliminary tests to ensure all statistical assumptions of our *t* tests were met. For all comparisons, Levene's test of equal variances was significant—indicating that the assumption that variance is equal throughout the data set was *not* met. Specifically, CURE students had less variation in their responses than traditional students. To confirm that the three LCAS scales were internally consistent and reliable for each group, we calculated Cronbach's alpha values for each scale within each group. Cronbach's alpha values for the traditional group were 0.83 for collaboration, 0.84 for discovery/relevance, and 0.90 for iteration, which are considered very good. Cronbach's alpha values for the CURE group were slightly lower, 0.76, 0.76, and 0.75 respectively, but still demonstrate good internal consistency (Nunnally 1978; Lance *et al*., 2006). Thus, we interpret the difference in variation among groups to indicate that students are interpreting the items in the same way but there is more variation in design features among traditional courses than among CUREs. To accommodate this difference when performing our *t* tests, we adjusted the independent-samples *t* test degrees of freedom with the Welch correction (Welch, 1947).

Table 3. Rotated factor loadings for the LCAS[a]

In this course, I was encouraged to …		Collaboration	Discovery and relevance	Iteration
C1	discuss elements of my investigation with classmates or instructors.	0.767	—	—
C2	reflect on what I was learning.	0.694	—	—
C3	contribute my ideas and suggestions during class discussions.	0.893	—	—
C4	help other students collect or analyze data.	0.708	—	—
C5	provide constructive criticism to classmates and challenge each other's interpretations.	0.617	—	—
C6	share the problems I encountered during my investigation and seek input on how to address them.	0.954	—	—
In this course, I was expected to …				
DR1	generate novel results that are unknown to the instructor and that could be of interest to the broader scientific community or others outside the class.	—	0.938	—
DR2	conduct an investigation to find something previously unknown to myself, other students, and the instructor.	—	0.592	—
DR3	formulate my own research question or hypothesis to guide an investigation.	—	0.421	—
DR4	develop new arguments based on data.	—	0.462	0.306
DR5	explain how my work has resulted in new scientific knowledge.	—	0.701	—
In this course, I had time to …				
I1	revise or repeat work to account for errors or fix problems.[b]	—	—	0.822
I2	change the methods of the investigation if it was not unfolding as predicted.	—	—	0.589
I3	share and compare data with other students.	—	—	0.451
I4	collect and analyze additional data to address new questions or further test hypotheses that arose during the investigation.	—	—	0.702
I5	revise or repeat analyses based on feedback.	—	—	0.764
I6	revise drafts of papers or presentations about my investigation based on feedback.	—	—	0.779
Cronbach's alpha		0.8	0.82	0.85
Factor correlations				
C		—		
DR		0.409	—	
I		0.453	0.528	—

[a]Factor loadings less than 0.25 were omitted. All collaboration items had four response options: "never," "one or two times," "monthly," and "weekly." All other items had six response options ranging from "strongly disagree" to "strongly agree." All items also included additional response options of "I don't know" and "prefer not to respond."

[b]For item I1, the item stem is "In this course, I was expected to ...," unlike the other items in this set.

This correction makes the tests more conservative and reduces the likelihood of making a type I error when comparing groups with unequal variances.

We hypothesized that students in CUREs would have significantly higher ratings overall, as well as higher ratings for each individual scale, compared with students from traditional labs. Overall, our results supported our hypothesis (Table 4). CURE students ($M = 75.10$, $SD = 8.67$) had significantly higher total ratings than traditional lab students ($M = 68.15$, $SD = 14.76$; $t(85.70) = 3.05$, $p < 0.01$, $d = 0.13$). CURE students also rated their courses higher on the discovery and relevance ($p < 0.001$, $d = 0.71$) and iteration ($p < 0.05$, $d = 0.39$) scales than students in traditional lab courses. Cohen's d ranged from 0.39 to 0.71, specifying medium to medium-large effect sizes for these scales (Cohen, 1992; Kotrlik and Williams, 2003). In contrast, we found no difference on the collaboration scale between groups ($p > 0.05$, $d = 0.07$). Both CURE and traditional lab course

Table 4. Group differences on the LCAS[a]

	CURE students		Traditional students						
	Mean	SD	Mean	SD	Welch *df*	*t*	*p*	*d*	Possible range of scores
Collaboration scale	21.11	3.20	20.87	4.02	128.06	0.40	>0.05	0.07	6–24
Discovery scale	24.35	4.04	20.77	5.82	104.37	3.64	<0.001	0.71	5–30
Iteration scale	28.71	4.15	26.53	7.00	95.91	2.47	<0.05	0.39	6–36
LCAS total score	75.10	8.67	68.15	14.76	85.70	3.05	<0.01	0.13	17–90

[a]Only student responses with complete cases on each scale were used. LCAS total score $n = 115$ (60 CURE students, 55 traditional lab course students). Collaboration $n = 141$ (73 CURE students, 68 traditional lab course students). Discovery and relevance $n = 133$ (72 CURE students, 61 traditional lab course students). Iteration $n = 134$ (72 CURE students, 62 traditional lab course students). Welch's *df* adjustment was made because the assumption of homogeneity of variance was not met.

students reported participating in collaborative practices on a weekly or monthly basis (Table 1). Thus, it appears that both CUREs and traditional labs engage students in collaborative and metacognitive practices.

DISCUSSION

Comparison of CUREs and Traditional Lab Courses

The aim of this study was to present and evaluate a new instrument, the Laboratory Course Assessment Survey, or LCAS, intended to measure design features that distinguish CUREs from traditional lab courses in biology. We found that students and instructors showed very high agreement about whether their courses were CUREs or traditional lab courses. For this reason, we have presented the LCAS as a measure of lab course design rather than student perceptions of course design. Concerns have been raised by us and others about relying on student reporting of outcomes from research experiences (Corwin Auchincloss *et al*., 2014; Corwin *et al*., 2015), primarily because of lack of evidence that student reports of knowledge or skill gains accurately reflect their learning (Falchikov and Boud, 1989). Our results suggest that there is a single "lab course design" construct that is reflected in both student perceptions of their lab course experiences and instructors' intended course designs, at least at the level of whether a course could be considered a CURE or a traditional lab course.

Using EFA, we show that the LCAS consists of three scales that measure: 1) collaboration, 2) discovery and relevance, and 3) iteration. We established content validity of the LCAS by developing scale items based on hypothesized course design features of CUREs (Corwin Auchincloss *et al*., 2014). We found high interitem reliability for each of the three scales. We used a known-groups comparison to test our hypothesis that the LCAS would differentiate between CUREs and traditional laboratory courses. We found that the LCAS differentiates between CUREs and traditional labs as perceived by both students and instructors. Two of the three scales differentiate between CUREs and traditional labs with medium to medium-large effect sizes. Overall, the psychometric properties of this instrument indicate it is suited for data collection in a variety of undergraduate biology lab courses.

The discovery and relevance scale combines two of the five proposed design features of CUREs: the potential for students to make discoveries and the relevance of their work beyond the classroom. Discovery, as we use it here, emphasizes novelty to both students and other stakeholders (e.g., instructors and members of a scientific community). The discovery and relevance scale places strong emphasis on this by specifying that discovery must be new to students, instructors, and communities outside the course (see items DR1, DR2, and DR5 in Table 3). In addition, these questions are indicative of the relationship between relevance and discovery (e.g., item DR1). The extent to which a result is a discovery depends on its relevance to a broader body of knowledge. Brownell and Kloser (2015) also propose grouping discovery and relevance for pedagogical reasons. They argue that learning experiences that involve students in making discoveries (i.e., discovery) and in relating science to their daily lives (i.e., relevance) lead to similar student outcomes, such as increased excitement and engagement (Brownell and Kloser, 2015).

The iteration scale reflects what it was originally designed to measure—the extent to which students have opportunities to change, revise, or repeat aspects of their work based on feedback. Items in this scale emphasize iteration as performed by a single student in order to make progress toward achieving a research goal or to address a scientific question. This scale does not measure iteration as performed by a class of students (e.g., two or more students working separately on the same sample to replicate results), because students may be unable to accurately assess whole-class forms of iteration. This scale also does not measure iteration as repeated practice of the same procedure to develop technical expertise or to conduct different investigations with goals that are distinct from one another, which may be common in traditional or inquiry courses. This distinction is made in an effort to target the specific form of iteration that is observable in research.

These two scales differentiate between CUREs and traditional lab courses. These results align with our expectations, because traditional lab courses are typically designed to demonstrate well-established phenomena (i.e., not discovery), and iteration is not an essential element for students to make progress in traditional lab courses, because the outcomes are known (Brownell and Kloser, 2015). In contrast, CUREs involve students in work with unknown outcomes (i.e., discovery) and repeating aspects of the work is often necessary to generate results and confirm findings (i.e., iteration). The combination of discovery and relevance has a large effect size and thus explains the majority of the difference between CURE and traditional lab courses. This makes sense, as the potential for discovery is a defining feature of research and thus of CUREs. The medium effect size observed for iteration might be the result of iteration being inherent to the research process and thus to CUREs, while being an optional element of traditional lab courses based on whether instructors offer opportunities for students to repeat aspects of their work.

Both CUREs and traditional labs had similarly high ratings on the collaboration scale. This is not surprising, considering that social learning theory has long been used to explain how people learn (Bandura, 1971; Lave and Wenger, 1991) and collaborative group work is increasingly a part of instructional design in higher education (Springer *et al*., 1999; Felder and Brent, 2007). We anticipate that student responses on the collaboration scale reflect group work in both course types. Brownell and Kloser (2015) suggest that students in CUREs are more likely than students in traditional lab courses to engage in cognitively demanding, or "interactive" forms, of collaboration. Chi and Wylie (2014) define interactive learning experiences as those in which students build off, elaborate on, or add to one another's ideas. This is in contrast to passive, active, and constructive forms of learning, in which students orient and receive information without overtly doing anything with it (passive), students do something overtly active with information provided to them (active), and students produce something beyond what was presented to them (constructive). Although the current scale is useful for assessing collaboration, further development and testing would be necessary to elucidate nuances in collaboration that reflect interactivity.

The collaboration scale also included items related to metacognition, or students' reflection on their own knowledge or

thinking. The relationship between collaboration and metacognition is not surprising, as the process of collaboration demands that students verbalize their thoughts, including their rationales and interpretations. This in turn requires students to reflect on their own thinking, or to be metacognitive (Chi *et al*., 1994; Tanner, 2009). Although metacognition was described as a component of collaboration in the grounding work for this paper (Corwin Auchincloss *et al*., 2014), it was not a central focus in development of the LCAS.

Limitations of the LCAS

As with any research aimed at developing new instruments, this study has several limitations. The items in the collaboration scale are measured on a four-point rating scale in contrast to the six-point rating scales used to assess discovery and relevance and iteration. There is a chance that the four-point rating scale may mask fine-grained differences in collaboration between CURE and traditional courses. The four-point scale may also have influenced how the collaboration items factored. It is unlikely that this is a major influence on the factor analysis results, because the final six items in this scale were part of a larger set of nine items with the same response options. The three items that were removed had low interitem correlations, showing that the clustering of items is not merely as artifact of the rating scale. In future uses of the LCAS, we recommend expanding the collaboration rating scale to five points (i.e., 1, weekly; 2, every other week; 3, monthly; 4, one or two times; 5, never) to determine whether small differences in collaboration can be identified between CUREs and traditional lab courses.

Another limitation is that our comparison of CUREs and traditional lab courses relies on the assumption that the variation in students' responses is due mainly to course design and not other factors, such as instructor experience or students' prior lab course or research experience. For example, as students progress through a degree program and gain experience, they may hone their ability to assess whether their work is truly relevant to the scientific community. These factors should be measured in future uses of the LCAS to determine their influence on students' responses and to better characterize the validity of the LCAS.

Finally, as with many educational studies, our sample was not random. Participation in the study was voluntary, and self-selection into the study may have biased the group of student participants. However, we made efforts to collect data from different institutions across the United States and to include a variety of lab course types. This likely helped us to capture a spectrum of student experience that may have mitigated self-selection bias. In addition, participants were mainly white and Asian students (40–50% and 28–37% respectively, depending on the analysis). An insufficient number of Hispanic/Latino(a), black, and other underrepresented minority (URM) students (<10% in each group) participated in the study for us to test the properties of the survey with students from a full range of backgrounds. Although the percentage of URM students in our survey population was not substantially different from national percentages of URM students who graduated with a bachelor's degree in biology (Table 1; National Science Foundation, 2014), future research should evaluate the instrument's utility with a broader range of students.

Implications for Research and Teaching

We believe that the LCAS will be useful to both instructors and researchers for a range of assessment purposes. For example, the LCAS could be used to assess the degree to which collaboration, discovery and relevance, and iteration are present in a particular course or vary among different types of lab courses (e.g., CUREs vs. inquiry labs), among different courses of the same type (e.g., different CUREs), and among different offerings of the same course (e.g., different offerings of SEA-PHAGES; Jordan *et al*., 2014). The LCAS could also be used for determining the extent to which each design feature relates to student outcomes. For example, students who collaborate once or twice during a course may develop communication skills, while students who collaborate on a weekly basis may develop communication skills and a sense of community with their peers (Corwin *et al.*, 2015). Results from these kinds of studies will be useful for informing the design and teaching of lab learning experiences.

Research that connects course design features with outcomes will also be useful for future studies of the efficacy and effectiveness of CUREs. Efficacy studies elucidate the key components needed for an innovation to be successful (i.e., What is necessary and sufficient for a CURE to be effective?), and effectiveness studies examine how adaptations of an innovation affect the intended outcomes (i.e., To what extent is there latitude in how CUREs are implemented?; O'Donnell, 2008). Future research should compare the relationship between CURE design features and student outcomes in different institutional and disciplinary contexts. Future studies should also examine the extent to which students experience CUREs differently based on their backgrounds, demographics, and prior experiences. For example, students who are non science majors or early in their undergraduate careers may perceive any lab learning experience with outcomes unknown to them as "discovery," while upper-division majors might view discovery and relevance more narrowly as experiences in which they have the potential to present their work to a broader scientific community or publish their findings. Such differences may have unique and important implications for students' psychosocial development in ways that influence their persistence in science (Estrada *et al*., 2011). For example, high discovery and relevance ratings from introductory students, who are new to higher education and to the practice of science, may influence the extent to which they feel like part of a community. In contrast, high discovery and relevance ratings from upper-division students may influence the extent to which they identify as scientists and thus pursue further research experiences or go on to graduate education in science.

The LCAS could also be tested to determine its usefulness for comparing students' experiences in research internships, since the design features it measures are likely to be present and to vary across internship experiences (Corwin Auchincloss *et al*., 2014). For example, directors of undergraduate research programs may wish to assess the degree to which the internships include these design features and how design features relate to the outcomes realized by students who self-select into these experiences.

Although our results show that the LCAS is useful for measuring several aspects of CURE design, we acknowledge the LCAS does not measure all features that are likely to make CUREs distinctive as learning experiences. For example, we

originally wrote items to measure participation in science practices. However, these items did not form a single unidimensional scale, because each practice represents a distinct activity. Future research should investigate other approaches to measuring student engagement in science practices, such as the creation of an inventory, similar to the teaching practices inventory designed by Wieman and Gilbert (2014), or the design of course-specific tools and rubrics as proposed by Brownell and Kloser (2015). Project ownership has also been identified as an important feature of lab course experiences (Hanauer *et al.*, 2012; Hanauer and Dolan, 2014). Future research should develop and test more complex models of the relationships among design features (ownership, discovery and relevance, iteration, collaboration, etc.), student-level differences, and student outcomes to yield insights into the efficacy and impact of different research experiences for diverse students (Corwin *et al.*, 2015; Linn *et al.*, 2015).

Future research should continue to test the alignment between student perceptions of course design and instructors' intended course designs and to identify, if possible, any variables that allow student perceptions and instructor intentions to be delineated. For example, instructors may design courses to engage students in research. Yet if students do not believe they are doing legitimate work (Lave and Wenger, 1991) that has the potential to contribute to a larger research endeavor, they may report having a different experience than the instructor intended. The experience may be research in the eyes of the instructor but not authentic from a student perspective (Rahm *et al.*, 2003). This may result in a different level of engagement that is limited in its resemblance to the actual practice of science (Sadler and McKinney, 2010) and more closely resembles the process of following steps in a protocol typically observed in traditional lab courses. If student perceptions and instructor intentions can be teased apart, it would be possible to examine whether student outcomes more closely relate to student perceptions, instructor intentions, or the alignment between the two.

ACKNOWLEDGMENTS

Thanks to Melissa Aikens, Lucas Wachsmuth, Sarah Eddy, Patrick Enderle, Stephanie D. Rivale, Josh Beckham, and Stacia Rodenbusch for their careful reading and thoughtful feedback on drafts of the manuscript. Thanks to Sara Merkel for feedback during survey development. Support for this work was provided by a grant from the National Science Foundation (NSF DBI-1450729). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of NSF.

REFERENCES

Adedokun OA, Zhang D, Parker LC, Bessenbacher A, Chilcress A, Burgess WD (2012). Towards an understanding of the processes of the effects of undergraduate research experiences on students' aspiration for research careers and graduate education. J Coll Sci Teach *42*, 82–91.

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC. http://visionandchange.org/files/2011/03/Revised-Vision-and-Change-Final-Report.pdf2011/03/Revised-Vision-and-Change-Final-Report.pdf (accessed 10 January 2014).

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). Standards for Educational and Psychological Testing, Washington, DC.

Bachman LF (2004). Statistical Analyses for Language Assessment, Stuttgart, Germany: Ernst Klett Sprachen.

Bandura A (1971). Social Learning Theory, New York: General Learning Press.

Bangera G, Brownell SE (2014). Course-based undergraduate research experiences can make scientific research more inclusive. CBE Life Sci Educ *13*, 602–606.

Benson J (1998). Developing a strong program of construct validation: a test anxiety example. Educ Meas Issues Pract *17*, 10–17.

Benson J, Hagtvet KA (1996). The interplay among design, data analysis, and theory in the measurement of coping. In: Handbook of Coping: Theory, Research, Applications, ed. M Zeidner and NS Endler, New York: Wiley.

Borrego M, Cutler S, Prince M, Henderson C, Froyd JE (2013). Fidelity of implementation of research-based instructional strategies (RBIS) in engineering science courses. J Eng Educ *102*, 394–425.

Brownell SE, Kloser MJ (2015). Toward a conceptual framework for measuring the effectiveness of course-based undergraduate research experiences in undergraduate biology. Stud High Educ *40*, 1–20.

Chi MT, Wylie R (2014). The ICAP Framework: linking cognitive engagement to active learning outcomes. Educ Psychol *49*, 219–243.

Chi MTH, de Leeuw N, Chiu MH, LaVancher C (1994). Eliciting self-explanations improves understanding. Cogn Sci *18*, 439–477.

Cohen J (1992). Statistical power analysis. Curr Dir Psychol Sci *1*, 98–101.

Corwin LA, Graham MJ, Dolan EL (2015). Modeling course-based undergraduate research experiences: an agenda for future research and evaluation. CBE Life Sci Educ *14*, es1.

Corwin Auchincloss L, Laursen SL, Branchaw JL, Eagan K, Graham M, Hanauer DI, Lawrie G, McLinn CM, Pelaez N, Rowland S, *et al.* (2014). Assessment of course-based undergraduate research experiences: a meeting report. CBE Life Sci Educ *13*, 29–40.

Cronbach LJ (1951). Coefficient alpha and the internal structure of tests. Psychometrika *16*, 297–334.

Dolan EL, Lally DJ, Brooks E, Tax FE (2008). PREPping students for authentic science. Sci Teach *75*, 38–43.

Eagan MK, Hurtado S, Chang MJ, Garcia GA, Herrera FA, Garibay JC (2013). Making a difference in science education: the impact of undergraduate research programs. Am Educ Res J *50*, 683–713.

Enders CK (2010). Applied Data Analysis, New York: Guilford.

Estrada M, Woodcock A, Hernandez PR, Schultz P (2011). Toward a model of social influence that explains minority student integration into the scientific community. J Educ Psych *103*, 206–222.

Falchikov N, Boud D (1989). Student self-assessment in higher education: a meta-analysis. Rev Educ Res *59*, 395–430.

Felder RM, Brent R (2007). Cooperative learning. In: Active Learning: Models from the Analytical Sciences, ACS Symposium Series 970, ed. PA Mabrouk, Washington, DC: American Chemical Society.

Finney SJ, DiStefano C (2006). Non-normal and categorical data in structural equation modeling. In: Structural Equation Modeling: A Second Course, ed. GR Hancock and RO Mueller, Charlotte, NC: Information Age.

Fleiss JL (1981). Statistical Methods for Rates and Proportions, 2nd ed., New York: John Wiley.

Hanauer DI, Dolan EL (2014). The Project Ownership Survey: measuring differences in scientific inquiry experiences. CBE Life Sci Educ *13*, 149–158.

Hanauer DI, Frederick J, Fotinakes B, Strobel SA (2012). Linguistic analysis of project ownership for undergraduate research experiences. CBE Life Sci Educ *11*, 378–385.

Hayton JC, Allen DG, Scarpello V (2004). Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. Organ Res Methods *7*, 191–205.

Horn JL (1965). A rationale and test for the number of factors in factor analysis. Psychometrika *32*, 179–185.

Jordan TC, Burnett SH, Carson S, Caruso S, Clase K, DeJong RJ, Dennehy JJ, Denver DR, Dunbar D, Elgin SCR, *et al.* (2014). A broadly implementable research course for first-year undergraduate students. MBio *5*, e01051.

Kaiser HF (1960). Directional statistical decisions. Psychol Rev *67*, 160.

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

Kotrlik JW, Williams HA (2003). The incorporation of effect size in information technology, learning, and performance research and performance research. Inf Technol Learn Perform J *21*, 1.

Lance CE, Butts MM, Michels LC (2006). The sources of four commonly reported cutoff criteria: what did they really say? Organ Res Methods *9*, 202–220.

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

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

Lave J, Wenger E (1991). Situated Learning: Legitimate Peripheral Participation, Cambridge, UK: Cambridge University Press.

Linn MC, Palmer E, Baranger A, Gerard E, Stone E (2015). Undergraduate research experiences: impacts and opportunities. Science *347*, 1261757.

Little TD, Jorgensen TD, Lang KM, Moore EWG (2013). On the joys of missing data. J Pediatr Psychol *39*, 151–162.

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

Lopatto D (2010). Science in Solution: The Impact of Undergraduate Research on Student Learning, Washington, DC: Council on Undergraduate Research and Research Corporation for Scientific Advancement.

Lopatto D, Alvarez C, Barnard D, Chandrasekaran C, Chung HM, Du C, Eckdahl T, Goodman AL, Hauser C, Jones CJ, *et al.* (2008). Undergraduate research: Genomics Education Partnership. Science *322*, 684–685.

Muthén LK, Muthén BO (1998–2012). Mplus User's Guide, 7th ed., Los Angeles, CA: Muthén & Muthén.

National Research Council (2003). BIO2010: Transforming Undergraduate Education for Future Research Biologists, Washington, DC: National Academies Press.

National Science Foundation (2014). Higher education in science and engineering. Science and Engineering Indicators, 2014 chap. 2. www.nsf.gov/statistics/seind14/content/chapter-2/chapter-2.pdf (accessed 10 March 2015).

Netemeyer RG, Bearden WO, Sharma S (2003). Scaling Procedures: Issues and Applications, Thousand Oaks, CA: Sage.

Nunnally JC (1978). Psychometric Theory, New York: McGraw-Hill.

O'Donnell CL (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. Rev Educ Res *78*, 33–84.

President's Council of Advisors on Science and Technology (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_feb.pdf (accessed 15 September 2014).

Prosser M, Trigwell K (1999). Understanding Learning and Teaching: The Experience in Higher Education, Buckingham, UK: Open University Press.

Rahm J, Miller HC, Hartley L, Moore JC (2003). The value of an emergent notion of authenticity: examples from two student/teacher–scientist partnership programs. J Res Sci Teach *40*, 737–756.

Raîche G, Walls TA, Magis D, Riopel M, Blais JG (2013). Non-graphical solutions for Cattell's scree test. Eur J Res Methods Behav Soc Sci. *9*, 23–29.

Sadler TD, McKinney L (2010). Scientific research for undergraduate students: a review of the literature. J Coll Sci Teach *39*, 43–49.

Schafer JL, Graham JW (2002). Missing data: our view of the state of the art. Psychol Methods *7*, 147.

Schultz PW, Hernandez PR, Woodcock A, Estrada M, Chance RC, Aguilar M, Serpe RT (2011). Patching the pipeline: reducing educational disparities in the sciences through minority training programs. Educ Eval Policy Anal *33*, 95–114.

Springer L, Stanne ME, Donovan SS (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: a meta-analysis. Rev Educ Res *69*, 21–51.

Tabachnick BG, Fidell LS (2001). Using Multivariate Statistics, Boston, MA: Pearson.

Tanner KD (2009). Talking to learn: why biology students should be talking in classrooms and how to make it happen. CBE Life Sci Educ *8*, 89–94.

Thiry H, Laursen SL (2011). The role of student-advisor interactions in apprenticing undergraduate researchers into a scientific community of practice. J Sci Educ Technol *20*, 771–778.

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

Welch BL (1947). The generalization of "Student's" problem when several different population variances are involved. Biometrika *34*, 28–35.

Wieman C, Gilbert S (2014). The Teaching Practices Inventory: a new tool for characterizing college and university teaching in mathematics and science. CBE Life Sci Educ *13*, 552–569.