Chemistry Education Research and Practice

Measuring integrated understanding of undergraduate chemistry research experiences: Assessing oral and written research artifacts

Journal:	*Chemistry Education Research and Practice*
Manuscript ID	RP-ART-04-2021-000104.R2
Article Type:	Paper
Date Submitted by the Author:	31-Oct-2021
Complete List of Authors:	Helix, Max; University of California Berkeley, Graduate Group in Science and Mathematics Education Cote, Laleh; University of California Berkeley, Graduate Group in Science and Mathematics Education; Lawrence Berkeley National Laboratory, Lawrence Berkeley National Laboratory Stachl, Christiane; University of California Berkeley, Department of Chemistry Linn, Marcia C.; University of California Berkeley, Graduate Group in Science and Mathematics Education Stone, Elisa; University of California Berkeley, CalTeach Program Baranger, Anne; University of California Berkeley, Department of Chemistry; University of California Berkeley, Graduate Group in Science and Mathematics Education

Measuring integrated understanding of undergraduate chemistry research experiences: Assessing oral and written research artifacts

Authors: Max R. Helix, Laleh E. Coté, Christiane N. Stachl, Marcia C. Linn, Elisa M. Stone, and Anne M. Baranger

Abstract

Understanding the impact of undergraduate research experiences (UREs) and course-based undergraduate research experiences (CUREs) is crucial as universities debate the value of allocating scarce resources to these activities. We report on the Berkeley Undergraduate Research Evaluation Tools (BURET), designed to assess the learning outcomes of UREs and CUREs in chemistry and other sciences. To validate the tools, we administered BURET to 70 undergraduate students in the College of Chemistry and 19 students from other STEM fields, comparing the performance of students who had less than one year of undergraduate research to those with more than one year of research experience. Students wrote reflections and responded to interviews during poster presentations of their research project. BURET asks students to communicate the significance of their project, analyze their experimental design, interpret their data, and propose future research. Scoring rubrics reward students for integrating disciplinary evidence into their narratives. We found that the instruments yielded reliable scores, and the results clarified the impacts of undergraduate research, specifically characterizing the strengths and weaknesses of undergraduate researchers in chemistry at our institution. Students with at least a year of research experience were able to use disciplinary evidence more effectively than those with less than one year of experience. First-year students excelled at explaining the societal relevance of their work, but they incorporated only minimal discussion of prior research into

their reflections and presentations. Students at all levels struggled to critique their own experimental design. These results have important implications for undergraduate learning, suggesting areas for faculty members, graduate student research mentors, and CURE or URE programs to improve undergraduate research experiences.

Keywords: undergraduate research, assessment, knowledge integration, URE, CURE, postsecondary chemistry education, scientific poster presentations, instrument

Introduction

Opportunities to conduct research are a critical component of undergraduate education for many students majoring in science, technology, engineering, and mathematics (STEM) disciplines, allowing them to engage with the larger scientific enterprise while still completing relevant coursework. Although research experiences vary widely in nature, they generally share common goals across settings, such as developing research skills, improving understanding and application of scientific content knowledge, expanding scientific reasoning skills, increasing confidence for doing science, and integrating students into scientific culture (Linn *et al.*, 2015; Robnett *et al.*, 2015; Rodenbusch *et al.*, 2016; National Academies of Sciences and Medicine, 2017). Numerous studies have focused on the relationship between participation in a research experience and the development of self-efficacy, confidence, and attitudes in/toward science, which are linked to academic retention and career choice (e.g., Shuster et al., 2019; Ashcroft et al., 2020; Avargil et al., 2020; Esparza et al., 2020). Research experiences can serve as a positive influence for career aspirations involving science, despite other challenges students may have faced since entering their college or university (Seymour and Hunter, 2019). As a result, many institutions and funding organizations across the U.S. have dedicated considerable resources to

support these programs each year (Laursen *et al.*, 2010; Auchincloss *et al.*, 2014; Krim *et al.*, 2019).

Despite a national call to replace traditional introductory laboratory courses with research-based courses, this practice is still emerging at U.S. colleges and universities (Olson and Riordan, 2012; Laursen, 2019). Those course-based undergraduate research experiences (CUREs) that have been developed in chemistry allow students to develop self-confidence and project ownership, as well as contributing to novel research in chemistry (Kerr and Yan, 2016; Ghanem *et al.*, 2018; Cruz *et al.*, 2020). Additionally, students who complete these courses believe that they have learned more chemistry content than they would have in traditional lecture and laboratory courses (Chase *et al.*, 2017). These research-based courses support student interest in chemistry, as students find them to be more enjoyable than "cookbook" laboratories with predetermined project outcomes (Clark *et al.*, 2016; Mutambuki *et al.*, 2019; Muna, 2021). Several studies have explored the benefits of group-based approaches to supporting undergraduates as they learn about and conduct chemistry research (Danowitz *et al.*, 2016; Hauwiller *et al.*, 2019).

Due to their prevalence and potential impact, it is important to assess the effects of science research experiences on student learning, in order to determine how students progress over time and to identify how research experiences can be improved to better serve participants (Auchincloss *et al.*, 2014). Such assessments are relevant to both CUREs and undergraduate research experiences that take place in research laboratories (UREs). Most previous studies that assess learning outcomes of science research experiences are limited to a description of the research experience or self-report data; fewer studies validate self-reports with analysis of research products, direct measures of mastery of scientific content or practice, or observations of

student activities (Linn *et al.*, 2015; National Academies of Sciences and Medicine, 2017; Krim *et al.*, 2019; Lin *et al.*, 2019). Scholars such as Pagano *et al.*, 2018 and Stone *et al.*, 2020 have commented on the limited number of studies dedicated to examining the impacts of CUREs in chemistry, when compared to the life sciences. Thus, there is a need for additional assessment tools that can be applied to undergraduate research experiences in chemistry, both inside and outside of the classroom.

Literature on undergraduate research and educational policy documents have identified the following scientific practices as foundational to research experiences for undergraduates (Laursen *et al.*, 2010; Sadler *et al.*, 2010): formulating research questions or hypotheses, designing experiments, analyzing and interpreting data, making conclusions, iteratively planning next steps, and explaining the significance of the research project. Collectively, these scientific reasoning skills are widely regarded as a critical component of science education; educators have moved away from the idea that such skills involve a single cognitive activity, and they are most often viewed as a "set of different but coordinated skills" (Opitz *et al.*, 2017). Thus, the goal of this study was to develop assessment tools to be used with STEM majors, with a particular focus on chemistry, that measure the extent to which they understand research as a set of connected practices. A specific aim for this work was to focus on assessing students' understanding of scientific practices in the context of their own research project, rather than investigating their ability to answer questions about a hypothetical scenario. In this study, we address the following research questions:

1. Do the tools we developed distinguish between undergraduate students with different levels of prior research experience?

2. What do our tools tell us about what students understand about research and what they are still learning at different stages of their undergraduate careers?

Theoretical Framework

The theoretical basis for our work comes from Knowledge Integration (KI), a framework that has been used extensively in the design of learning environments and instruments to assess K-12 student knowledge of scientific content and practices (Linn, 1995; Linn and Eylon, 2011; Ryoo and Linn, 2012; Stone, 2014; Linn *et al.*, 2018). This learning science framework emphasizes that coherent understanding occurs when students make deep connections between their prior and new ideas. KI specifies four key components that support student learning (Linn and Eylon, 2011). The first is to elicit student ideas and prior understandings about a given topic. Students already have a repertoire of knowledge to draw on, and new knowledge will ultimately be built on these existing structures. Second, as students engage more with a particular concept, they discover new, scientifically normative ideas, some of which may challenge or contradict existing ideas. Third, as students explore the ideas they discover, they begin to distinguish between competing ideas and the contexts in which they are applicable. This process leads to a more nuanced understanding of the topic. Finally, students reflect upon their new knowledge in order to consolidate it into a coherent narrative.

The process of conducting research generates knowledge in a way that parallels the KI framework (Linn *et al.*, 2015). Activities such as predicting and hypothesizing allow for *eliciting* undergraduate students' initial ideas. Undergraduate researchers then begin *discovering* new ideas over time as they gather data and participate in other research practices (Linn and Eylon, 2011; White and Gunstone, 2014). They gradually learn to *distinguish* between possible interpretations for their data, and *reflecting* on their research enables learners to consolidate

knowledge and generate new ideas for future work (Brown *et al.*, 1989; Linn and Eylon, 2011). KI guides our expectation that as undergraduates progress in research, they will become more proficient in understanding and discussing their research project, linking their insights to relevant discipline-specific content knowledge to form coherent arguments.

Literature Review

Impacts of Student Participation in Science Research Experiences. A number of studies suggest that gains related to retention in STEM (e.g., graduation rates, entry into the STEM workforce, graduate school attendance) are supported through participation in research experiences, especially for students from groups historically underrepresented in STEM fields (Schultz *et al.*, 2011; O'Donnell *et al.*, 2015; Estrada *et al.*, 2016; Carpi *et al.*, 2017). There is increasing evidence to link participation in authentic scientific research with the development of science identity through immersive learning of discipline-specific practices, referred to as "legitimate peripheral participation" in situated learning theory (Lave and Wenger, 1991; Robnett *et al.*, 2015). Factors such as a positive science identity, self-efficacy development, access to mentoring, and engagement in research at the undergraduate level are important for persistence in STEM and are critical for supporting students from groups historically underrepresented in STEM fields (Carlone and Johnson, 2007; Chang *et al.*, 2011; Mondisa and McComb, 2018; Ortiz *et al.*, 2020). In chemistry, participation in research experiences contribute to retention, enthusiasm for chemistry-related careers, and appreciation for the process of engaging in research in this discipline (e.g., Kerr and Yan, 2016; Williams and Reddish, 2018; Muna, 2021).

Measures of Student Learning in Science Research Experiences. Various performance assessments have been developed to directly measure multiple dimensions of student knowledge

and skills gained from participation in science research experiences (Butz and Branchaw, 2020). For example, the Danczak–Overton–Thompson Chemistry Critical Thinking Test (DOT) was designed to measure critical thinking skills in chemistry students, regardless of prior student knowledge in chemistry (Danczak *et al.*, 2020). The Biological Experimental Design Concept Inventory (BEDCI) measures knowledge and diagnoses non-expert-like thinking in experimental design by analyzing open-ended responses to different scenarios (Deane *et al.*, 2014). The Assessment of Critical Thinking Ability (ACTA) is an open-ended survey that assesses critical thinking skills in biology and chemistry students (White *et al.*, 2011). The Rubric for Experimental Design (RED) identifies areas of experimental design in which undergraduates struggle (Dasgupta *et al.*, 2014, 2016). The Performance assessment of Undergraduate Research Experiences (PURE) instrument measures experimental problem solving and quantitative literacy skills in chemistry students participating in UREs through a series of multipart questions about real-world scientific problems (Harsh, 2016; Harsh *et al.*, 2017). The Test of Scientific Literacy Skills (TOSLS) consists of multiple-choice questions about real-world problems and measures student skills related to scientific literacy (Gormally *et al.*, 2012). Crawford and Kloepper, 2019 developed an exit interview involving a series of written and oral exercises that assess the ways in which chemistry students connect course content to laboratory activities.

Some instruments are "authentic assessments," which are meaningful opportunities for students to integrate and apply their knowledge to novel, complex, and/or realistic situations that simulate typical activities of scientists (Wiggins, 1998; Doğan and Kaya, 2009; Laungani *et al.*, 2018). For example, the Experimental Design Ability Test (EDAT) gives students a real-world scenario and research question and tasks them with designing an appropriate experiment, and has been used in chemistry and the life sciences (Sirum and Humburg, 2011; Goodey and Talgar,

2016). Several studies suggest that writing activities can support student understanding of chemistry concepts (e.g., Lewis dot structure model) and methods (e.g., spectroscopy), as well as confidence in communicating about the material (Shultz and Gere, 2015; Moon *et al.*, 2018; Watts *et al.*, 2020). The Rubric for Science Writing and the Tool to assess Interrelated Experimental Design (TIED) are two assessment tools designed for use in undergraduate science courses, which involve students in activities that scientists engage in (Timmerman *et al.*, 2011; Killpack and Fulmer, 2018).

There is compelling evidence to suggest that participation in a CURE leads to significant gains in research skills and academic outcomes and can support the subsequent advancement to (and success in) a URE (Rodenbusch *et al.*, 2016; Krim *et al.*, 2019). Studies that measure student learning gains typically consider these gains only over the course of a semester-long research experience, though there is evidence to suggest that undergraduates need to participate in high-impact research experiences spanning more than one semester to develop their understanding of the research process (Deane *et al.*, 2014; Corwin, Runyon, *et al.*, 2015; Griffeth *et al.*, 2015; Harsh, 2016; Remich *et al.*, 2016; Hernandez *et al.*, 2018). A longitudinal study by Szteinberg and Weaver (2013) suggests that CURE students retain chemistry content knowledge longer, as compared to students in traditional laboratory courses.

Prior Gaps Identified in Learning for Undergraduate Researchers. Previous studies point to a lack of mastery among undergraduate researchers in fully understanding their research projects in several areas (Airey and Linder, 2009; Coil *et al.*, 2010; Gormally *et al.*, 2012). Prior findings suggest it is possible for a student to participate in research without understanding the scientific or societal significance of their work, though this skill supports more expert-level reasoning in the discipline of the research project (Bransford *et al.*, 2000; Coil *et al.*, 2010).

Many students, and in particular those from groups underrepresented in STEM fields, choose a STEM education/career in order to make a positive contribution to their communities and/or society, and this interest is likely to influence their commitment to a career in STEM (Bonous-Hammarth, 2000; Harackiewicz and Hulleman, 2010; Chang *et al.*, 2014). However, undergraduates do not always develop the ability to articulate answers to questions about the context of their research project, such as: "Why is this question important to others in this discipline?" or, "What is the 'big picture?'" (Timmerman *et al.*, 2011).

In order to become independent researchers, undergraduates are also expected to develop an understanding of experimental design (Sirum and Humburg, 2011; Killpack and Fulmer, 2018). Undergraduates are typically presented with narratives about previously completed experiments as part of their STEM coursework, but training in designing experiments is less common (Gormally *et al.*, 2012). When reading scientific papers, undergraduates commonly struggle with evaluating and critiquing the design elements used in the studies being discussed (Varela *et al.*, 2005; Coil *et al.*, 2010). Guided-inquiry laboratories, CUREs, and UREs, in which students design their own experiments, can be used to support experimental design skills in chemistry (Goodey and Talgar, 2016). Multiple studies make the case that instruments are needed to measure experimental design and other skills critical for the development of students as scientists as they prepare to advance in their professional career (e.g., Sirum and Humburg, 2011; Dasgupta *et al.*, 2014, 2016; Danczak *et al.*, 2020).

Science research experiences often require that students contribute to data interpretation, but many undergraduates enter introductory-level STEM courses with insufficient skill in understanding how to work with data (e.g., reading graphs, analyzing and interpreting data, creating data visualizations), and STEM coursework does not necessarily cover this content (Coil

et al., 2010; Maltese *et al.*, 2015). Comparing the data analysis skills of various researchers showed that novices are more heavily reliant on personal beliefs, while those with more expertise focus on empirical consistency to draw conclusions from their observations (Hogan and Maglienti, 2001). Chemical education studies suggest that students need to be taught explicitly how to generate and interpret the kinds of visualizations they will need for a particular project, and instruction should be intentional about connecting data to relevant concepts and addressing misconceptions (Connor *et al.*, 2019; Rodriguez *et al.*, 2019). Relatively few studies in chemistry have focused on assessing student skill level in this area, though there is consensus that data interpretation is critical for developing chemists (Maltese *et al.*, 2015; Peteroy-Kelly *et al.*, 2017).

Undergraduate students should also be able to develop hypotheses and conduct appropriate experiments to test these hypotheses by the time they graduate with a STEM bachelor's degree (White *et al.*, 2013). When students are provided with the space to encounter challenges, revise their research goals, and repeat their work, this iterative process can have a powerful impact on their sense of ownership as they learn to navigate obstacles in their scientific discipline (Corwin *et al.*, 2018; Gin *et al.*, 2018). CUREs focused on chemistry have been shown to improve students' project ownership in lower-division, upper-division, and large-enrollment undergraduate courses (Williams and Reddish, 2018; Cruz *et al.*, 2020; Heller *et al.*, 2020).

The BURET Study

We have drawn from this literature to develop four **Berkeley Undergraduate Research Evaluation Tools (BURET) Indicators** that describe areas where undergraduates are expected to integrate their understanding of foundational scientific practices:

I. Communicate the significance of their specific project to the overarching research questions of the laboratory and the broader scientific field

II. Justify their experimental design as appropriate for their research question

III. Analyze and interpret data in order to construct explanations and models that are relevant to their research question

IV. Generate hypotheses and plan future experiments relevant to their research question in response to their analysis and interpretation of data

These Indicators provide the focus for the new instruments described in this study. We designed an interview protocol and reflective prompts to assess how undergraduates develop an integrated knowledge of these dimensions as they engage in research. The first is the Reflection instrument (BURET-R), which prompts written student reflections about the progress of their research project. The second is the Poster interview (BURET-P), which is administered at capstone poster sessions. Both tools were administered to students in a variety of research settings.

Methods

Participants and Context

Participants were recruited from CUREs and UREs at our institution. Undergraduate researcher volunteers came from five different populations (Table 1), with 78% majoring in chemistry disciplines (chemistry, chemical engineering, or chemical biology). It should be noted that nearly all students in UREs had previously taken a CURE, as is typical for URE students in many science departments at this institution. Our procedures were approved by the University of California, Berkeley Committee for Protection of Human Subjects, Protocol #2016-02-8360.

Table 1. Study populations.

Group	n	Response Rate	Type	Duration	Prior Research Experience	Student Description
1	35	58%	CURE	Semester	Mostly none	Freshman chemistry students
2	6	90%	CURE	Summer	None	New chemistry transfer students
3	28	59%	URE	Ongoing	Variable	Department of chemistry students
4	5	65%	CURE	Semester	None	Pre-service STEM teachers
5	15	88%	URE	Summer+	Variable	Pre-service STEM teachers

Students in Groups 1 and 4 (Table 1) were enrolled in CUREs in which the student was responsible for developing their own research question and choosing the methods used to investigate that question. Students in Group 2 chose from possible research projects that could be investigated using computational chemistry approaches. Students in Groups 3 and 5 had typical apprentice-style research experiences in faculty labs, where the projects varied but fit into the overarching goals of their faculty advisor and were generally related to the projects of their graduate student mentor. The level of independence in designing their own work varied as well, generally according to the amount of time each undergraduate had spent working in their research group.

The students participating in the study ranged from having zero to four or more semesters of research experience prior to study participation. The study population of 89 undergraduates contained a mixture of identities, including gender, race, ethnicity, and first language. Our study participants were 58% female, and 24% stated that English is not their first language. Students who self-identified as American Indian/Alaska Native, Black/African-American, Hispanic/Latinx, or other Pacific Islander, collectively referred to as underrepresented minorities (URM), were intentionally oversampled and comprised 19% of our study population.

Instrument Development

Expert Review of the BURET Indicators

To confirm that the BURET Indicators were aligned with the goals of faculty advisors, all chemistry faculty at this institution working with undergraduate researchers were invited to participate in an interview. A total of 21 faculty agreed to be interviewed, for a response rate of 41%. The faculty in this study ranged from assistant professors to full professors and had a wide variety of research group sizes. During a 1-hour interview, faculty were asked to describe their goals for their undergraduate researchers, discuss mentoring practices, and review the BURET Indicators. They commented on whether these were appropriate goals for their undergraduates. Nearly all of the responses were positive, with some faculty members expressing that the Indicators "exactly" described their overall goals for undergraduate researchers. An additional 12 faculty members from other STEM departments were also interviewed, and they gave largely similar responses.

Assessment Design

We sought data collection and assessment approaches that would both support student learning and allow for direct measures of the Indicators across both CUREs and UREs. Many undergraduate researchers create a poster and present their research project as a capstone requirement, providing an opportunity to assess student integration of scientific content and practices in the context of their own work. A set of interview questions targeting the BURET Indicators were developed to ask at the end of each student's prepared presentation. This interview protocol coupled with a rubric to assess several aspects of these verbal presentations make up the BURET Poster Presentation instrument (BURET-P, Fig. 1).

1. Can you please summarize why your research project and what you've learned is important?
2. Can you explain more about why you (and your lab) chose this general strategy for your research project?
3. Can you choose one experimental technique that is central to this work and say why you used it, rather than other options?
    3b. What are the limitations of this technique?
4. Could you expand on how you interpret these results?
    4b. How confident are you in your data and your conclusions?
5. What would you do if you had another year to work on this project, and why?

Fig. 1. BURET-P interview protocol.

A pair of reflective prompts (BURET-R, Fig. 2) were developed to complement the poster presentation assessment. These prompts can be administered at different points in the research experience to provide information on students' developing progress on the Indicators. In this study, BURET-R was administered a few weeks prior to their poster session. These prompts targeted Indicators 3 and 4, respectively, but many students also incorporate discussions of Indicators 1 and 2 in their responses.

1. *Data Analysis Prompt*: Think about the ways you have analyzed data recently.
   (a) Describe one example of data analysis you have done.
   (b) Reflect on how you used this data analysis to create or change an explanation or a model. Frame your response for an experienced scientist who is unfamiliar with your project.
2. *Next Steps Prompt*:
   (a) If you had another month or two to work, what would be your next steps and why?
   (b) What about if you had another year?

Fig. 2. BURET-R reflective prompts.

Items and Scoring Rubrics

Preliminary rubric development was conducted with a small group of 7 undergraduate and 5 graduate students. All participants responded to BURET-R prompts, and a few also presented posters to the research team. Four undergraduates were also interviewed, during which they were asked to expand on their BURET-R and BURET-P responses. Written responses and audio recordings were reviewed to develop the rubrics. The emergent themes from initial rounds

of coding and a review of the relevant literature were used to develop an overlapping set of specific items aligned with the BURET Indicators, resulting in a set of 6 items for the BURET-R and 11 items for the BURET-P scoring instruments (Table 2). Although not all of these items were explicitly elicited by our prompting questions, they were all commonly discussed in student answers. Our assumption was not necessarily that every response would address every item, but that on average, more expert-like responses would integrate more different types of content into the overall narrative, as suggested by the Knowledge Integration theoretical framework. It should be emphasized that the rubrics are not a direct measure of everything a student knows, but rather of what they choose to present.

Table 2. Scoring rubric items for the BURET instruments align to BURET Indicators.

Item	BURET-R	BURET-P
Placing their work in a broader context	X	
Placing their work in a broader scientific context		X
Placing their work in a broader societal context		X
Providing rationale for an experimental design choice	X	X
Addressing limitations of an experimental design choice		X
Comparing alternatives to an experimental design choice		X
Number of experimental design choices with some rationale		X
Identifying and discussing the key variables	X	
Describing their data analysis procedures OR Interpreting their data	X	
Interpreting their data		X
Analyzing sources of error and uncertainty		X
Proposing next steps for the project	X	X
Incorporating references to previous work		X
Integrating additional content knowledge	X	X

Although many of the research projects assessed during this study involved experiments, this was not universally the case. To account for other types of research, “experimental design choice” was defined as any approaches, strategies, techniques, or other decisions made during the study design process. This definition was sufficiently broad to encompass the work described by all students who participated in this study.

To develop a KI rubric for each item, we were guided by prior KI rubrics for similar items. The KI rubrics score items on a 0-5 scale. Each score represents a progressively more integrated and connected response. Applying the KI framework to BURET scoring, descriptions were written for each possible score on each item (see Appendix A for the complete rubric). These were anchored by the idea that a 2 should be a correct statement about an isolated part of the research process, and a 4 should be a clear, basic link from a relevant part of the research process to evidence from scientific content and research practices. For example, a score of 2 on "addressing the limitations of an experimental design choice" could be obtained by simply noting a drawback for a particular technique, whereas a score of 4 would require that students explain that limitation by integrating underlying scientific principles into their discussion or by making a clear reference to the research question. The remaining levels were defined as follows: 0 indicates that responses relevant to the item are absent, 1 indicates a vague statement, 3 indicates a partial link between an assertion and relevant scientific content or practices, and 5 indicates a complex link of 3 or more isolated concepts. The highest level descriptions were informed in part by the graduate students who responded to the BURET-R and BURET-P assessments as scoring categories were being refined. A partial scoring rubric with example responses can be found in Appendix B.

Instrument Testing

Data Collection

To determine whether the BURET instruments could detect a difference between novice and advanced researchers, students enrolled in the three target CUREs (Groups 1, 2, and 4 in Table 1) were invited to be part of this study. A few weeks before their corresponding poster session, student responses to BURET-R were collected in class from all who agreed to

participate (see Table 1 for response rates). At the final poster session for those courses, a sample of the consenting students was interviewed using the BURET-P protocol (see Appendix C for sampling procedures). Students were interviewed by one of 10 different interviewers by first allowing the student to give their prepared presentation uninterrupted, then using the standardized protocol in Fig. 1 to elicit specific elaborations. Poster interviews were recorded and transcribed before coding.

Additionally, students presenting at one of the two target URE poster sessions (Groups 3 and 5 in Table 1) were invited by email to participate in this study. Responses to BURET-R were collected via Qualtrics a few weeks prior to the poster sessions, and all consenting students who provided responses to BURET-R were interviewed at their poster session, using the same protocol that was used with the CURE students. From our full dataset, 80 BURET-R responses and 55 BURET-P interviews were found to be complete and fully legible or audible, and these were used in our subsequent analysis.

Coding and Rubric Reliability

Data from BURET-R and BURET-P were scored for each study participant according to the corresponding rubrics. 60% of the written responses to BURET-R were coded by two different researchers, and discrepancies were resolved through subsequent discussion. A weighted Cohen's kappa of 0.73 was deemed acceptable, and subsequent coding was completed individually. Each poster transcript was deidentified with respect to student experience level and other characteristics. Transcripts were then coded independently by at least two people, and any discrepancies between coders were discussed and resolved. Two different pairs of coders assessed each transcript. The coders varied in their level of chemistry expertise, and we found

that our scoring rubrics could be used by any coders with a basic understanding of different research topics and expertise in scientific research practices. Rather than assessing "correctness," these rubrics focused on our primary goal of assessing integration of content and practices with respect to various parts of each research project. A weighted Cohen's kappa of 0.65 was achieved between coding pairs, using posters that were coded by all researchers. This is considered to be a substantial level of agreement according to Landis and Koch (1977).

Quantitative Analysis

We established the experimental validity of our instruments based on their ability to successfully distinguish between responses from undergraduates with more or less prior research experience. Participants were divided into novice (0-1 semester) and advanced (2+ semesters) groups based on how many semesters of research they had completed prior to the one in which they were presenting a poster. For each instrument, all items were averaged to produce a single test statistic. We compared novice and advanced participants using a t-test. Additionally, student scores on each item were collapsed into either low (KI score of 0-3) or high scores (KI score of 4-5), and chi-squared tests were performed to determine whether high scores were significantly associated with increased research experience for individual items. Further psychometric analysis was performed to establish the internal structure at the instrument level and to investigate the dimensionality of our construct (see Appendix D for details). As a measure of internal consistency, Cronbach's alpha was calculated for each instrument. All statistical analysis was conducted on Stata.

Item-response theory (IRT) analysis was conducted to gather validity evidence based on internal structure at the instrument level. Because our sample was not sufficient to run the

analysis using all thresholds from our rubric, data were collapsed into scores of low (0-2), moderate (3), or high (4-5), and Wright maps for each instrument were generated from the collapsed data. Additionally, exploratory factor analysis was performed and item-test correlations were calculated to determine whether the construct we are measuring is uni- or multi-dimensional. All statistical analysis was conducted on Stata except for the IRT analysis, which was performed on Conquest.

Results

Do the BURET Instruments Distinguish Between Undergraduate Students with Different Levels of Prior Research Experience?

An analysis of student responses showed that both the BURET-R (n = 80) and BURET-P (n = 55) instruments are able to distinguish between more and less experienced undergraduate researchers. Total scores for each instrument revealed statistically significant differences between students with 2 or more semesters of prior research experience and students with less experience ($p < 0.001$; Tables 3 and 4). Average scores on each item also increased with more research experience, with 9 of the 17 items showing statistically significant gains.

Table 3. Mean scores on BURET-R rubric items.

Semesters of Previous Research Experience	0-1	2+	Sig
Sample size (n)	42	38	
Placing work in a broader context	2.9	3.6	*
Providing rationale for expt. design choice	1.9	2.8	**
Identifying and discussing the key variables	2.0	2.4	
Describing OR interpreting data analysis	2.6	3.5	**
Proposing next steps for the project	2.6	3.2	
Integrating additional content knowledge	0.8	2.4	**
Average Score	2.1	3.0	***

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

Table 4. Mean scores on BURET-P rubric items.

Semesters of Previous Research Experience	0-1	2+	Sig
Sample size (n)	24	31	
Placing work in broader scientific context	2.3	3.5	*
Placing work in broader societal context	3.6	3.7	
Providing rationale for expt. design choice	3.5	3.9	
Addressing limitations of expt. design choice	2.8	3.3	
Comparing alternatives to expt. design choice	2.7	3.4	
Expt. design choices with some rationale (max. 5)	2.5	3.3	*
Interpreting their data	3.1	3.5	*
Analyzing sources of error and uncertainty	2.3	2.5	
Proposing next steps for the project	3.1	3.2	
Incorporating references to previous work	1.9	2.9	*
Integrating additional content knowledge	2.3	3.5	*
Average Score	2.7	3.3	***
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$			

As a measure of reliability, Cronbach's alpha is calculated to be 0.78 for both instruments, which is in the range considered acceptable for science education research instruments (Taber, 2018). Further psychometric analysis suggests an acceptable consistency of the items to measure respondent performance and provides evidence that a unidimensional construct is being measured (see Appendix D for more information).

Two other variables that are highly correlated with increased research experience are year in school and whether the research experience was part of a course. As previously mentioned, most of the novice researchers in our sample were enrolled in a CURE, while most of the advanced researchers were participating in a URE in a faculty lab and had previously completed a CURE. To determine which of these variables was the best predictor of total score on our instruments, factorial ANOVAs were run using year in school, semesters of research experience, and URE/CURE as the independent variables. For the BURET-R, only URE/CURE was a significant predictor ($p < 0.05$), whereas the duration of time spent in college or in undergraduate

research were not. For the BURET-P, only semesters of research experience was a significant predictor ($p < 0.05$). On this instrument, the type of research experience did not have a significant effect on student performance, and the URE students with minimal total research experience performed similarly to other novice researchers. No interaction terms were significant for either instrument. We were unable to examine whether there were differential effects for students who identified as a URM because we recruited too few of these students with at least 2 semesters of research experience. Future work will be needed to investigate this aspect of our instrument.

Comparison of the BURET-R and BURET-P

Overall, largely similar final results were obtained from the BURET-R and the BURET-P instruments, which were generally administered several weeks apart. Student scores on the BURET-R and BURET-P instruments were significantly correlated with one another ($r = 0.4$, $p < 0.01$, see Appendix E for a scatterplot). The items on which students tended to excel or struggle were similar across the two instruments, with some variations based on the exact relationships between the items assessed and the specific prompt or interview questions being answered. Targeted questions asked during the poster presentations generally elicit more specific information than the broader reflective prompts, resulting in more items being coded when assessing poster presentations. Poster presentations were also much longer than the written responses to BURET-R; on average, written responses were 248 words in length, while poster presentations (including answers to questions) were 1,682 words in length. In general, students scored higher on BURET-P (average score = 3.1) than on BURET-R (average score = 2.3). This can also be seen by looking at individual participants; 85% of the participants scored higher on their poster presentations than on their written responses.

What do the BURET instruments tell us about what undergraduate chemistry students understand about research and what they are still learning at different stages of their undergraduate careers?

The BURET instruments provide information about the progress students make on each of the BURET Indicators as they gain in research experience. The following sections describe the characteristics of student progression along each Indicator using the KI framework, including undergraduate student performance on each item and the items that most differentiate novice and advanced study participants. Items are grouped by which Indicator they are most closely associated with to provide a more holistic picture of each primary area of assessment and for clarity of interpretation. (Table 5).

Table 5. BURET-P items grouped by most related BURET Indicator.[a]

Indicator	Items
I	Placing their work in a broader scientific context
	Placing their work in a broader societal context
	Incorporating references to previous work
	Integrating additional content knowledge
II	Providing rationale for an experimental design choice
	Addressing limitations of an experimental design choice
	Comparing alternatives to an experimental design choice
	Number of experimental design choices with some rationale
III	Interpreting their data
	Analyzing sources of error and uncertainty
IV	Proposing next steps for the project

[a] The items are grouped by Indicator for clarity of interpretation. Note that there is no evidence from the internal structure of the instruments for the items to be grouped in this way

Indicator 1: Communicating Significance

The first BURET Indicator assesses how well students can communicate the significance of their specific project to the overarching research questions of the laboratory and the broader scientific field. Three of the four items corresponding to this Indicator for the BURET-P instrument showed statistically significant growth between novice and advanced students.

Advanced students demonstrated a more sophisticated understanding of their project's scientific context ($p < 0.05$), referred to previous work more often ($p < 0.05$), and integrated more content knowledge into their presentations ($p < 0.05$) when compared to less experienced students (Table 4). Analysis of students' responses to the BURET-R instrument also provided evidence that they develop in their ability to place their project into a broader context (Table 3).

Advanced students often demonstrated a more integrated understanding of scientific context by explaining the current state of the field or how their research might affect projects in other labs. One chemistry student who received a score of 4 stated, "I was ... working on investigating ... the mechanical properties of polycarbonate urethane. Our research is particularly relevant to joint implants and joint replacements, ... the current industry standard polymer is called ultra high molecular wave polyethylene. ... Polycarbonate urethane or PCU is being pioneered as a new material. ... But it's pretty new so we're still doing research on the very mechanical properties and how it will react to being in the body and in an ionic environment where there is salts and stuff like that, that can affect its microstructure." In this response, the student clearly connects their work on the mechanical properties of PCU to the broader field of material science, particularly in the area of artificial joints.

A student would receive a 2 on the "Incorporating references to previous work" item by clearly referring to previous research but failing to explicitly link that research to their experimental design or compare it to their own results. For example, a student was scored 2 for the following vague reference to previous work, 'A lot of it was help from literature that we've seen online, especially the solvents. I wouldn't have known where to start without using some of these.' As an example of a higher scoring discussion, one student stated that, "There'd been, not a consensus, but almost every single study that we had read previously looking for these heavy

metals in chocolate, but also in other candy, had focused on the cocoa, then being the source and maybe mentioned other possible sources in passing." The student then compared this body of previous work with their own work, which found a possible alternate source of heavy metals, resulting in a score of 4.

Additionally, advanced students scored higher on providing context by integrating more additional content knowledge into their presentation and answers. Additional content knowledge was defined as "exhibiting scientific content knowledge beyond what is required to describe the project." Students received a 2 by simply providing some additional clarification, or a 4 by providing multiple examples or extensive discussions of relevant information. It should be noted that this does not directly measure the content knowledge of a student, but rather the extent to which students have *integrated* that content knowledge into discussions of their research.

Indicator 2: Justifying the Experimental Design

The second BURET Indicator was assessed with three items that focused on how students discussed their experimental design choices, which were defined as approaches, strategies, techniques, or other decisions made during the study design process. As mentioned previously, these design choices did not necessarily have to be strictly "experimental," since not all research projects encountered in this study were based on experiments. When asked to provide a rationale for an experimental design choice, the difference between novice and advanced student responses on the BURET-R instrument is significant ($p < 0.001$). Although more advanced students generally scored higher than novices on the BURET-P instrument on providing rationale, addressing limitations, and comparing alternatives to their experimental design choices, none of these differences were statistically significant.

Examples varied broadly, from why the research group chose to study a certain topic to the specific instruments used to collect raw data. In the BURET-P interview protocol, students were asked *why* they made a given design choice *instead of* something else, and they were also asked about the *limitations* of that choice. In general, both novice and advanced students scored highly on providing a rationale for an experimental design choice related to their project for the BURET-P instrument; over half of the students scored 4 or higher, which requires a clear description of the design choice *and* an explicit rationale that integrates domain-specific content knowledge. For example, "We chose to use micro plasma atomic emissions spectroscopy because of its wide dynamic range. While there were many other instruments that would have worked similarly well, but not within this large range. And we were very uncertain as to whether we were over diluting or under diluting our samples…. We only had rough EPA guidelines to kind of guide our choices." The marginally higher average score for advanced students compared to novices was not significant. However, advanced students did explain a greater number of their decisions than novices. To reflect this, an item was included that simply counted the number of design choices for which the student provided some rationale. This number was significantly higher ($p < 0.05$) for advanced students, reflecting the greater detail in which they described their experimental design.

Students were less proficient at discussing the limitations of experimental design choices. A representative response is, "So if the standards aren't prepared correctly or if they're too high on concentration, it may negatively, it definitely will negatively affect our data. So I think that's a big limitation." which received a 2 for only identifying user error as a possible limitation. However, some students were able to discuss limitations more fluently; for example, the following excerpt scored a 4: "The limitations of that technique are that bringing it under PBS,

which is phosphate buffered saline, only mimics the ionic concentrations. It doesn't mimic the chemical function. So [what] we'd like to do for further research is hydrate it in [inaudible], which … mimics in vivo synovial fluid." Both novice and advanced students showed moderate levels of sophistication on the "comparing alternatives" item but rarely scored as high as 4, for which they needed to make a clear comparison between their choice and the alternative, explaining why their choice was superior. For example, "We decided to use MPAS instead of graphite furnace atomic absorption spectroscopy, even though both measure lead very well. Because MPAS has a larger dynamic range, and we were very uncertain as to the concentration we were gonna get."

Indicator 3: Interpreting the Data

Items for the third BURET Indicator measured the extent to which students were able to analyze and interpret data in order to construct explanations and models relevant to their research question. The data interpretation item for BURET-P focused on constructing explanations that demonstrated domain-specific content knowledge; advanced students were significantly more likely ($p < 0.05$) to score higher on this item. This is consistent with results from the BURET-R instrument, on which advanced students scored higher on describing or interpreting their data analysis ($p < 0.01$).

For example, "And we found that with lower concentrations of silver, we get the same amount of silver conductivity" scored a 2 because there was a clear statement about the experimental results but no additional comments were made about the data or their conclusions. A score of 4 required students to *explain* what they observed: "We stained the plates, which contained the cellulose media, with Congo red, which is a dye that binds to cellulase. So what that allowed us to do is once we washed the excess dye away, we got results that looked like this:

the bacterial colonies that didn't produce any cellulase show no halo, and the whole plate is red, because the cellulose is still there, the dye binds, it's all still there. The ones that you see here have a halo of white, are positive results. They produce cellulase, and we know that because around the bacterial colony, is a halo where the cellulose has been degraded, and the dye doesn't bind." This chemical biology student describes the underlying mechanism of the assay, explaining what is happening on a molecular and cellular level to justify their interpretation.

While scores for data interpretation were generally relatively high, students performed less well on analyzing sources of error and uncertainty. Most students identified a clear potential source of error or expressed skepticism about their results, but less than half of the students elaborated on their answer or connected that source of error to either their experimental design or their conclusions. A more complete response might explain how the experiment was designed to control for possible sources of error. For example, "And also, to avoid error we wanted to use NMR. First, we dissolve our wristbands using deuterated chloroform, and then running that through NMR, and seeing if there are any errors that we can possibly encounter for contamination. We just wanted to make sure the wristbands were mostly silicon. We had a positive control and negative control in just the chemical that we tested." However, most students did not discuss sources of error at this level and there were no significant differences between novice and advanced students.

Indicator 4: Proposing Future Investigations

A single assessment item aligned with the final BURET Indicator measures the extent to which students are able to generate hypotheses and plan future experiments relevant to their research question and in response to their analysis and interpretation of data. This item primarily evaluates the rationale given along with the next steps for the research project proposed by the

student. Interestingly, advanced students did not score significantly higher on this item than novice students for either the BURET-R or -P instrument.

Students who received a 2 on this item typically suggested "more"-based continuations of their work with no rationale: more trials, more substrates, more different temperatures, and so on. In contrast, students who received a 4 would include a rationale that integrates domain-specific content knowledge; for example, one chemistry student said, "In the future we hope to perform confocal microscopy to determine the depth of infiltration, that's another common problem with current scaffolds is that they'll grow in an x-y plane and spread out in a nice flat layer, but they don't go into the bi-layer membrane. So that's what we're hoping to get with these fiber mats later, when you spin onto a mesh collector plate you get these really nice nodes, and we're hoping that cells could easily fit into those pores and infiltrate deeper into the membrane." Most students fell in between these two points; over 50% of participants scored a 3 on this item.

Discussion

We have introduced two novel instruments for assessing how undergraduate researchers grow in their understanding of scientific research. Our instruments assess student discussions of their own research project, complementing previously published instruments that assess the ability of undergraduates to answer questions about completely different research scenarios (e.g., Harsh, 2016) or specific components of the research process like experimental design (e.g., Deane et al., 2014). The BURET instruments use the Knowledge Integration framework to evaluate how undergraduates develop an integrated understanding of the different components of research and the scientific practices and content of their projects. Though developed primarily for chemistry researchers, they can be applied to different types of research situations and across various scientific disciplines, in contrast to most existing tools. Our instruments are able to

distinguish between students at different levels of research experience, and evidence is presented for their validity and reliability. Excerpts from our rich datasets of student responses provided a detailed picture of how students progress in their understanding of research and helped us to identify specific areas where they need more support to fully develop as researchers.

Novice undergraduate students require more guidance to place their research into a larger scientific context

We found that the largest difference between novice and advanced undergraduates was for providing a scientific context for their work. Even though chemistry faculty reported that undergraduates are sometimes given key papers to read when beginning work in a new laboratory, our results show that their understanding of the connection between their experimental work and the broader scientific context is often weak. The faculty interviews we conducted suggest that, at least in some research groups, minimal emphasis is placed on teaching novice undergraduates the scientific context of their research projects. Multiple faculty singled out the first Indicator as important but "hard in some cases for undergrads, they don't necessarily see the big picture at this time." Several faculty also mentioned that having their undergraduates read the literature was a weak point in their mentoring. The lower priority given to these areas by faculty mentors, particularly for novice students, may help explain why there is such an increase in performance once students have been participating in research for at least two semesters. Reading the scientific literature has been shown to be challenging for novice students, but these skills develop over time as they work with their graduate student mentors to read more papers (Nelms and Segura-Totten, 2019). Reisner and Stewart (2020) make the case that incorporating activities to engage students in reading and discussing literature is of critical importance in chemistry research settings, to support students "to think more like disciplinary experts."

However, faculty members may vary in terms of when they feel it is appropriate to better acquaint their undergraduates with the scientific literature. When they are ready, we suggest that mentors use published approaches for teaching students to read the literature (Hoskins *et al.*, 2011; Krontiris-Litowitz, 2013; Sato *et al.*, 2014) in order to help their undergraduates understand the scientific context of their project more rapidly. Additionally, curriculum developers could help facilitate this process by including more interaction with the primary literature in undergraduate coursework.

In contrast, students at all levels performed well on providing an integrated societal context for their work, and more advanced students did not receive higher average scores on this item. The ability to discuss the broader impacts of a research project is a valued skill, with some institutions offering courses explicitly aimed at training students in this area (MacFadden, 2009; Heath *et al.*, 2014). In two of the CUREs included in this study, students developed research questions, often addressing a societal issue of interest to them, and as a result, they could fluently discuss the societal relevance of their project. Because novice students were strong on this item, there was little growth with more research experience.

Support is needed for beginning undergraduate researchers to better justify their experimental design and interpret their data

The extent to which undergraduates are exposed to experimental design in chemistry coursework and research experiences varies widely, and students struggle with this critical skill (Espinosa, 2011; Gormally *et al.*, 2012; Laursen, 2019). Previous attempts to assess gains in experimental design ability during scientific research experiences showed a general trend that participation in a CURE or URE improves student reasoning in this area (Sirum and Humburg, 2011; Dasgupta *et al.*, 2014; Harsh, 2016; Harsh *et al.*, 2017; Shanks *et al.*, 2017). However,

identifying the limitations of an experimental design has been found to be a weak point, even for graduate students (Gilmore *et al.*, 2015). In our work, we found that both novice and advanced students scored relatively high on their ability to rationalize experimental design choices. We also found that more advanced students recognized that rationalizing experimental design, including providing the limitations of and alternatives to their experimental design choices, is an important component of talking about their research. The difference between novice and advanced students' rationalizations of experimental design choices for BURET-R is significant ($p < 0.001$). Similarly, advanced students were more likely to include limitations and alternatives as components of their presentation of their research.

The general trends we observed for experimental design also hold for data interpretation; we showed that students generally performed well on giving straightforward interpretations of their data but were less likely to provide a richer description unless specifically prompted. Scores on the combined data analysis and interpretation items on both instruments were relatively high, with advanced students scoring significantly higher than novice students. This is consistent with other studies showing that data interpretation skills correlate with increased research experience (White *et al.*, 2011; Harsh *et al.*, 2017). In contrast, one of the lowest scoring items for both novice and advanced students was their ability to identify and discuss potential sources of error in their work. Students may deliberately focus on more positive aspects of their project, or the low scores may reveal a genuine deficit among undergraduates, who have been shown to struggle with critically analyzing experimental designs, generating data visualizations, and interpreting chemical data (Varela *et al.*, 2005; White *et al.*, 2011). Our study suggests that students may benefit from targeted interventions in these areas throughout their undergraduate career.

Novice and advanced students were equally proficient at proposing future work for their projects

The advanced and novice students in our sample were equally successful at proposing next steps for their research projects. This was surprising, because when faculty were asked what specifically they look for as signs of progress in their undergraduate researchers, many focused on day-to-day independence, including "thinking about what's next, what would be the next experiment after this one." The faculty interviewed by Laursen *et al.* (2010) also identified taking initiative, making decisions, and acting independently as markers of student progress.

A concept from the literature that is closely related to the item on proposing future work is that of iteration, as students scored higher when the proposed work was linked in some way to their most recent results. Authentic research is an iterative process, where the data from one experiment helps inform the next. Some have suggested that iteration is an essential part of an undergraduate research experience (e.g., Auchincloss *et al.*, 2014), and efforts have been made to explicitly include iteration in CUREs (Light *et al.*, 2019). Although there are instruments that measure whether a student perceives iteration to be a part of their research experience (Corwin, Graham, *et al.*, 2015), to our knowledge, there are no instruments that assess student proficiency in proposing next steps for an ongoing research project.

We anticipated that advanced students would be more experienced at proposing future experiments and would therefore be able to more fluently discuss them in their written responses and poster presentations. Although this was not reflected in the average scores, we observed that only advanced students received the highest possible score for proposing next steps on either the BURET-R or BURET-P instrument. Additionally, most of the advanced graduate students who were interviewed during the development of the instrument (see Methods), scored at the highest level on the BURET-P for this item.

One potential explanation for the discrepancy between expectations and observed results for undergraduate researchers on average is that many of the advanced undergraduate presenters were weeks away from graduation. Those students were likely in the process of concluding their research and were not planning longer-term directions of the project. As a result, their scores on proposing future work might be lower than if we had interviewed them earlier. In contrast, many novice students were enrolled in a one-semester CURE in which they were explicitly instructed to talk about future work as part of their poster presentation. Their relative success in this area suggests that, contrary to faculty expectations, even novice students can be expected to propose the next steps of their research project, and this expectation should be more explicitly integrated into UREs.

The BURET instruments apply to a range of different chemistry subdisciplines

One advantage of the BURET instruments is that they can be applied to very disparate projects spanning the wide range of subfields that fall under the larger domain of chemistry. The BURET instruments attempt to account for subdiscipline-specific knowledge without being restrictive, taking into account the fact that the understanding developed by working on a synthetic organic project is quite different from what one learns doing biophysical chemistry or atmospheric chemistry. To receive a higher score of 4 on the BURET data interpretation item, a student must explain what they observed in a way that demonstrates domain-specific content knowledge relevant to their research project. Because content knowledge for a diverse sampling of undergraduate researchers can be from a variety of disciplines, it has previously been difficult to measure with existing instruments about a single hypothetical scenario. We showed that such knowledge can generally be identified using the BURET instruments. For example, all of the excerpts in Table 6 scored a 4 on data interpretation except the atmospheric chemistry passage,

which scored a 3 because domain-specific content knowledge was vaguely alluded to instead of explicitly stated. We envision the BURET instruments being used by educational researchers to monitor student progress in a unified way across UREs and CUREs.

Table 6. Excerpts from poster presentation transcripts: Interpretation of observed results across various chemistry subdisciplines.

Chemistry Sub-Discipline	Excerpt from Student Poster Presentation
Biochemistry	My interpretation of these results is that the R-pal is utilizing the thiosulfate to grow and produce ammonia, so that's the main takeaway of this experiment and that if we took out thiosulfate and replaced it with another electron donor then they would grow with those electrons donated from that.
Inorganic Chemistry	What I've done here is I've synthesized a magnet that targets the lanthanide that has a strongly axial crystal field, but also a radical bridge, and this works very well because the 2,2'-bipyrimidine, that is substituted with chlorines, is a very weak epineural donor and so the crystal field becomes more axial because you have such a weak epineural donor even though you still have a radical lanthanide bridge.
Materials Chemistry	But the decrease is that prevention of growth that I was talking about, [due to] the charge neutralization of the bromide ions on the ends of the surfactant. So, if the surfactant is more packed, no more gaps are available for precipitation to occur, and so you can't grow any nanorods per se. All you're gonna be left with is a bunch of spherical nanoparticles, no growth curve. So that's the reason for this decrease.
Physical Chemistry	We first ran them on the mass spec to know that we know, that it's working as a control. So we can tell there's one peak for the full rotaxane, and then over time we can see the cleaved product come off, and that peak grows in over time. So then after 16 hours it works. So we go to do it with xenon NMR we can see the same thing after 16 hours you have a pretty full peak come in for CB-6. Here, this is the water peak for CB-6, we always see that in the xenon NMR experiment, and you see CB-6 peak, that's the xenon going in and out there.
Atmospheric Chemistry	What is shown here is the VOC reactivity to show it's relatively constant, and then the NO_x concentrations, and the ozone concentrations. So the NO_x decreases from weekday to weekend because there are less giant trucks driving. Then this is showing that ozone decreases, but it doesn't really decrease that much, it's basically the same.

Limitations

We identify four potential limitations of our study.

Self-selection bias is a limitation of undergraduate research studies, because those who participate are likely to be among the most highly motivated and high performing students. We expect selection bias to be minimal in our case, as approximately 70% of chemistry majors, who

make up the majority of our sample, participate in undergraduate research, giving them an opportunity to participate in the poster session from which we recruited our participants.

Non-uniform experimental conditions. Data was necessarily collected from a variety of CURE and URE contexts, such that some students typed their responses, while others submitted hand-written responses, leading to differences in length of response. We have attempted to counteract these issues by designing the BURET instruments specifically for different contexts and to deeply probe the quality of the responses. However, to ensure that accurate comparisons can be made between students, it is suggested that users of the BURET-R instrument provide an additional statement clarifying the desired length of the responses.

Low numbers of advanced URM participants. Although we attempted to oversample students who identify as a URM, we were only able to recruit four such students who had completed at least two semesters of undergraduate research. With so few advanced researchers, we were unable to determine whether our instrument has any intrinsic bias regarding URM students. This will be an important feature to assess in subsequent studies. The goals identified in faculty interviews were consistent with the categories we used in our coding rubrics, but researchers may assume certain cultural norms about the "correct" way to answer a question that is posed in a scientific setting. Although enculturation into a research program will likely result in more homogeneity over time, novice students in particular may differ in their interpretation of the questions purely as a result of their demographic background.

Preliminary CURE Experience. The most common trajectory for undergraduates in chemistry at our institution is to take a CURE prior to starting a URE in a faculty research group. At some institutions, students may start a URE without any prior CURE experience or enroll in a CURE concurrently with or after participating in a URE. Because we found that BURET-R

scores appear to be sensitive to the type of research experience, more work will be needed for its use in different universities and for comparisons across different sequencing of CURE and URE experiences.

Implications

In recent years, there has been increasing attention on the development and assessment of research experiences for undergraduate STEM majors. However, as compared to the life sciences, chemistry has produced fewer studies about the development and assessment of these opportunities, and fewer instruments have been developed to support student learning of research skills such as experimental design, data analysis, and reading the primary literature in chemistry. We have used the BURET-R and BURET-P instruments to characterize the progression of student expertise and reveal weaknesses in the learning outcomes of undergraduate researchers. While time-intensive to code, we envision the use of the BURET instruments to be highly valuable in the contexts of mentoring and future research on undergraduate research experiences in chemistry. The BURET-R instrument is more generally applicable, as it can be quickly administered. For assessing poster presentations in a variety of educational contexts, the corresponding BURET-P instrument can provide a more detailed picture of student knowledge integration. These instruments offer a method of assessing student learning in relationship with students' own chemistry research projects. Because the focus is on the student's project and not on answering questions about a hypothetical scenario, the instruments are authentic and can be used across the breadth of chemistry subdisciplines. Following initial development, many surveys, interview protocols, and performance-based instruments designed to measure student learning in the sciences have been applied as stand-alone assessments or in combination with other instruments in subsequent studies. We envision that some may find it helpful to use the

BURET instruments (both protocols and rubrics together) in their entirety, as an undergraduate research inventory, while others may administer a selection to their students.

Moreover, the BURET instruments provide an informal, low-stakes method for mentors to check on the progression of their students. Research mentors can regularly observe students setting up and analyzing the results of experiments, but they often have fewer opportunities to probe how their undergraduate students think about the research project more broadly. The BURET-R can be used as a way to quickly gauge how the student discusses their project in response to open-ended questions. This data can serve to guide research mentors to initiate conversations with the student to strengthen their understanding of the project and to consider how to better turn what they know into an integrated narrative about their project. Additionally, the act of responding to the BURET-R prompts is itself a useful opportunity for the student to reflect on their project, which may not be a regular feature of their research experience. Similarly, answering the BURET-P protocol questions is an inherently useful activity, as it can help students to strengthen their poster talks and provide practice taking questions from the audience.

At a departmental or institutional level, the BURET instruments can be used at regular intervals to assess how well a particular research experience is supporting student learning as they progress from novice to advanced researchers. The BURET instruments complement self-report survey data by enabling educational researchers to directly measure student learning with respect to knowledge and skills that are critical for their development as scientists. In the event that certain BURET Indicators are of greater importance with a particular student group, specific probes, like the interview questions in the BURET-P instrument, can be used to further explore student thinking for different components of the research process. Both of the BURET

instruments can be used to provide students with feedback about their strengths and knowledge gaps with respect to the research project they are working on in a CURE or URE. These instruments can also be used to compare different research experiences, providing individual CUREs or UREs with information about the areas in which students need additional instruction or training from their research mentors.

Conflicts of Interest

There are no conflicts to declare.

Appendices

Appendix A – Abbreviated Coding Rubric

CONTEXT – SCIENTIFIC OR SOCIETAL

Indicator 1: Place the research questions or goals of their laboratory and/or project in the context of the larger field.

Score	BURET-R	BURET-P – Scientific	BURET-P – Societal
0	Does not explain goals of experiment or project	Does not explain goals of project	Does not explain goals of project
1	Partial or unclear description of experiment and/or project goals	Partial or unclear description of project goals	Only discusses "personal" goals, but does not mention a societally relevant topic
2	States goal of experiment OR States goal of project	Clearly states goal of project OR States a very limited scientific application of their work	Collecting data with no further connection to societal importance OR Reader can infer societal importance or application of the data collected (i.e. mentions a societally-relevant topic like semiconductor or cancer)
3	Clearly states goal of experiment AND States goal of project (vagueness allowed)	States a general area of science that their work contributes to OR Vague or implied version of below	Implies societal importance OR Vague statement about the possible benefits or use of results
4	Partial Link (3) AND (Explains how expt advances larger project OR Explicit link of project to broader significance (scientific or societal))	Discusses how future projects (by other labs) might be affected by current project OR Suggests new research paths or projects that could be based on this work OR Provides sufficient	Explicitly connects project to specific societal need OR Explicit statement about the possible benefits or use of results (Accurate content knowledge and coherent argument should be present.

		background for reader to understand current state of field	However, exact mechanism of connection does not need to be stated.)
5	Partial Link (3) AND Explains how expt advances larger project or the portion of the project they are working on AND Explicit link of project to broader significance (scientific or societal)	Basic Link (4) with two out of three of the criteria present OR Two different scientific contexts explained for the project - both at Basic Links (4)	Explicit comparison between current project goals and existing solutions to those problems. Exact mechanism of connection does need to be stated. OR Explicit and specific statement about the possible benefits or use of results, including statement of existing societal issue or need

DESIGN CHOICES

Indicator 2: Justify their experimental design as appropriate for their research question and scientific content of their project.

Score	BURET-R & P – Rationale	BURET-P – Limitation	BURET-P – Comparison
0	Coder cannot identify any design choice discussed	Does not discuss any limitations of design choice	Does not mention any alternative design choices
1	Partial or unclear description of one design choice	Vague statement of a very generic limitation or logistical issue OR Vague or implied description of a thoughtful limitation - implied in the description of results	Mentions the fact that there are alternatives, but doesn't mention what these are.
2	Clear description of one design choice, but rationale is poor or absent	Clear statement of a very generic limitation or logistical issue OR Vague description of a thoughtful limitation	Mentions specific alternative, but no comparison OR Compares to alternative because alternative is, in their opinion "not possible"
3	Clear description of one design choice AND Gives reasonable (sounding) rationale but vague, implied, or invokes little to no content knowledge	One or more thoughtful limitations mentioned, but content knowledge only implied	Compares design choice to an alternative, but is somewhat vague or implied OR Compares to alternative because alternative is, in their opinion "not possible", plus *why* it wouldn't be possible
4	Clear description of one design choice AND Gives explicit rationale for choice of instrument or experiment that integrates domain-specific content knowledge	Gives at least one explicit limitation that integrates domain-specific content knowledge	Comparison to an alternative design choice on a single facet with a clear statement of difference or advantage or reason to use one or the other
5	Basic Link (4) but multiple distinct reasons for design choice are discussed AND Strong evidence of	Basic Link (4) AND (Discusses how limitations affect conclusions OR	Clear comparison to an alternative design choice on

	extensive content knowledge that supports their choices	Discusses how limitation was addressed, minimized, avoided, etc.)	more than one facet OR 3 or more Basic Links (4)

VARIABLES - WHAT IS BEING MEASURED, MANIPULATED, OR COMPARED?

Indicator 2: Justify their experimental design as appropriate for their research question and scientific content of their project.

Score	BURET-R Definition
0	Does not indicate what type of data is being collected or discuss any other relevant variables
1	Isolated Concept but vague or implied (unclear what they are actually measuring, manipulating, comparing) OR Basic instrument verification on a standard
2	Clearly identifies what is being measured (raw OR analyzed) OR Clearly identifies one or more variables being manipulated, compared, or held constant
3	Isolated Concept (2) AND (Provides basic rationale for choice of variables and/or range being investigated OR Gives details on how or to what extent the variables are manipulated) OR Basic link (4), but rationale or predictions are vague or questionable
4	Clearly identifies what is being measured (raw OR analyzed) AND Clearly states one or more variables being manipulated, controlled or compared AND (Provides rationale (clear, but slightly generic okay) for why manipulated variables would affect measurements/output OR Provides reasonable prediction of how manipulated variables will affect output)
5	Basic Link (4) AND Rationale and/or predictions are strong and integrate content knowledge

DATA MANIPULATION AND INTERPRETATION

Indicator 3: Analyze and interpret data in order to construct explanations and models that are relevant to their research question.

Score	BURET-R – Manipulation	BURET-R and P – Interpretation
0	Does not describe any analysis of raw data	Does not describe results OR Has not collected data yet
1	States that no data analysis was performed OR States that results are inconclusive with no elaboration	Unclear how conclusion is supported by results OR Implies data interpretation but does not sufficiently describe
2	States a procedure for analyzing or manipulating data with no elaboration	Summarizes results without interpretation OR Pre-packaged conclusion OR States an interpretation with no connection to data
3	Links raw data to analyzed results, but discussion of data or analysis method/procedure is vague	Summarizes results and links to content knowledge or compares to expectations, but vague or minimal insights
4	Clearly links raw data to analyzed results, including (clear) description of the analysis process	Gives plausible explanation for results (or compares results to expectations in a way) that integrates clear content knowledge

5	Basic Link (4), plus discusses at least one assumption or consequential decision made during analysis	Basic Link (4), but integrates extensive content knowledge OR Discusses alternate interpretations

CONFIDENCE/ERROR ANALYSIS

Indicator 3: Analyze and interpret data in order to construct explanations and models that are relevant to their research question.

Score	BURET-P Definition
0	Does not identify any potential sources of error
1	States that the experiment (or a large part of it) "didn't work" without any elaboration as to why OR Describes confidence in the ability of methods to answer the RQ
2	Identifies a clear "error" in what was done OR Vague reference to limitation of method/technique when discussing confidence in results OR Vague "doubts" about data
3	Identifies potential sources of error that are less "obvious" OR Clear reference to limitation of method/technique when discussing confidence in results
4	Clearly identifies potential reasonable source(s) of error AND Mentions how these connect to at least *one* of the following: 1. Research questions; 2. Experimental design (current or future); 3. Their conclusions
5	Clearly identifies multiple distinct potential reasonable source(s) of error at the level of a Basic Link (4)

NEXT STEPS

Indicator 4: Generate hypotheses and plan future experiments in response to their analysis and interpretation of data and research question.

Score	BURET-R and P Definition
0	Does not discuss any potential future work
1	Completely different goals for future work with no/minimal relationship to current work OR Implies that they will "continue with the plan" but does not sufficiently describe
2	Simple quantitative extension, modification, or new experiment with no or poor rationale OR "Continue with the plan" OR Repeat experiment with simple issue fixed
3	Simple quantitative extension with good rationale OR Modification or new experiment with credible but vague rationale OR Repeat experiment after difficult-to-predict issue fixed (troubleshooting), link to content knowledge is vague or absent
4	Modification, troubleshooting, or new experiment with clear rationale that integrates content knowledge
5	Multiple Basic Links (4), at least one of which is not a borderline Partial Link (3) OR (Basic Link AND Explicitly links new choices to the *results* of current work)

PREVIOUS WORK

Indicator 1: Place the research questions or goals of their laboratory and/or project in the context of the larger field.

Score	BURET-P Definition
0	Does not mention any prior work
1	Vague references to "other studies" without any specific designs/results or clear specification of how this informs part of project
2	Clear reference to previous work, but no stated connection to current work OR Vague reference to previous work with connection to current project

3	Clear description of previous design or results AND (Vague connection to/influence on current work OR Vague comparison b/w old and new design or results)
4	Summarizes previous work (specific design or results) AND (Explicitly states how it connects to/influenced current work OR Compares to current results)
5	Basic Link (4) AND (Explanation of how current work is different or novel OR Attempts to interpret sim/diff between current and previous results)

INTEGRATION OF (ADDITIONAL) CONTENT KNOWLEDGE

Indicator 1: Place the research questions or goals of their laboratory and/or project in the context of the larger field.

Score	BURET-R	BURET-P
0	Response does not integrate any scientific content knowledge beyond what is necessary to describe the project	Response does not integrate any scientific content knowledge beyond what is necessary to describe the project
1	(not used)	(not used)
2	Weak example of a Partial Link (3)	Weak example of a Partial Link (3)
3	Exhibits scientific content knowledge beyond what is required to describe project	Exhibits scientific content knowledge beyond what is required to describe project
4	(not used)	Exhibits *extensive* scientific content knowledge beyond what is required to describe project
5	Exhibits *extensive* scientific content knowledge beyond what is required to describe project	Multiple Basic Links (4)

Appendix B – Partial Coding Rubric with Examples

Score	Description	Examples
0	- Does not discuss any limitations of design choice	
1	- Vague reference to limitations	- "Again, part of the main problem is that graphite furnace is really temperamental."
2	- Clear statement of a generic limitation, OR - Vague description of thoughtful limitation	- "In terms of that technique, I think it depends on the accuracy in which the solutions are prepared. So if the standards aren't prepared correctly or if they're too high on concentration, it may negatively, it definitely will negatively affect our data. So I think that's a big limitation. And also you have to produce a lot of different samples, which can be time consuming."
3	- One or more thoughtful limitations mentioned, but content knowledge only implied	- "The limitations of Congo Red is that it is visual. So it is qualitative even though we can't measure the radius. The radius isn't really going to tell us anything numerical about how much cellulose the bacteria digests."
4	- Gives at least one explicit limitation that integrates domain-specific content knowledge	- "One experimental technique that we use is hydrating the sample and then putting them under nanoindentation. … So the limitations of that technique are that you're running it under PBS, which is phosphate buffered saline, and that only mimics the ionic concentrations, it doesn't mimic the chemical functionality you'd encounter in in vivo synovial fluid."
5	- Basic Link (4) AND	- "The main limitation is that the scaled particle theory ignores the entropic consideration in the energy of interaction here, so it's hard to say what would happen at different temperatures. In order to predict

(- Discusses how limitations affected conclusions OR - Discusses how limitation was addressed, minimized, avoided, etc.)	the temperature dependence, you need an approximate value of the entropy of dissolution, which isn't known for a lot of these molecules. However, we found that that's actually very easy to predict. For each group of molecules it's approximately constant for a certain chlorination number so you know that if you have a PCB and it has three chlorines that you will know the entropy very well."

Appendix C – Complete Sampling Procedures

All of the students enrolled in the three target CUREs were invited to be part of this study in person by one of the researchers. A few weeks before their corresponding poster session, students were asked to respond to the two reflective prompts, and answers were collected from all students who agreed to participate. This included 135 members of the chemistry CURE, for a response rate of 58%, 11 students in the pre-service teacher CURE, for a response rate of 65%, and 9 participants from the transfer student CURE, for a response rate of 90%. All of the consenting students in the pre-service teacher and transfer student CUREs were then interviewed at the final poster session for those courses. For the chemistry CURE, a subset of 35 consenting students were interviewed at the final poster session. To choose a representative sample, students were stratified by major and prior research experience. After intentionally oversampling 7 URM students, a random sample of 28 was chosen from among the remaining 128 students. From this pool of CURE participants, 6 URM students and a random sample of 9 other students were chosen for further analysis. An additional 20 students for whom we had prompt responses but not poster session interviews were also randomly selected for analysis.

Additionally, all students presenting at one of the two target URE poster sessions were invited by email to participate in this study. For the chemistry poster session, 112 students were invited to participate and 66 consented, for a response rate of 59%. For the pre-service teacher poster session, 23 of the 26 students (88%) responded affirmatively to the invitation. Responses to the reflective prompts were collected via Qualtrics a few weeks prior to the poster sessions.

All consenting students who provided answers to our reflective prompts (30 from the chemistry poster session and 23 from the pre-service teacher session) were interviewed at these poster sessions, using the same protocol. From this pool of URE participants, 6 URM students and a random sample of 24 other students were chosen for further analysis. An additional 15 students for whom we had prompt responses but not poster session interviews were also randomly selected for analysis. In total, the dataset we analyzed included 80 responses to reflective prompts and 55 poster session interviews.

Appendix D – Psychometric Analysis

Factor analysis provided evidence that a unidimensional construct is being measured. For each instrument, only one factor had an eigenvalue greater than 1, and the ratio of the first two eigenvalues was well above 4. Additionally, all items except one had a correlation of at least $r = 0.55$ with the overall score on the corresponding instrument. The sample size is expected to be sufficient for the one factor solution that used 6 variables for BURET-R and the one factor solution that used 11 variables for BURET-P (Mundfrom *et al.*, 2005).

Item-response theory (IRT) analysis was then conducted to establish the internal structure at the instrument level (Wilson, 2005). Because the sample size was not sufficient to run the analysis using all thresholds from the rubrics, data were collapsed into scores of low (0-2), moderate (3), or high (4-5), and Wright maps for each instrument were generated from the collapsed data, and there was at least one response for each possible answer choice in order to fit the data to an item response model. The resulting Wright maps (see Figs. D1 and D2 below) show that the range of instrument item logit values span nearly the entire distribution of respondent logit values, with only a few students falling below all item thresholds on the

BURET-R instrument, and a few Thurstonian thresholds located below the lowest respondent logit value for the BURET-P instrument. The reliability of partial credit model analysis carried out on the data is 0.77 for BURET-R and 0.76 for BURET-P. These values indicate an acceptable consistency of the items to measure respondent performance (Bond & Fox, 2007; Wright & Masters, 1982).

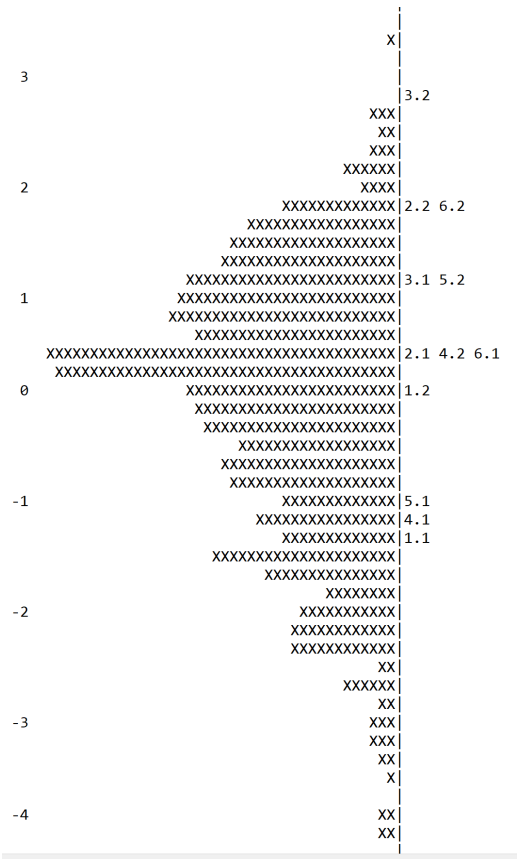


Fig. D1. Wright map for BURET-R

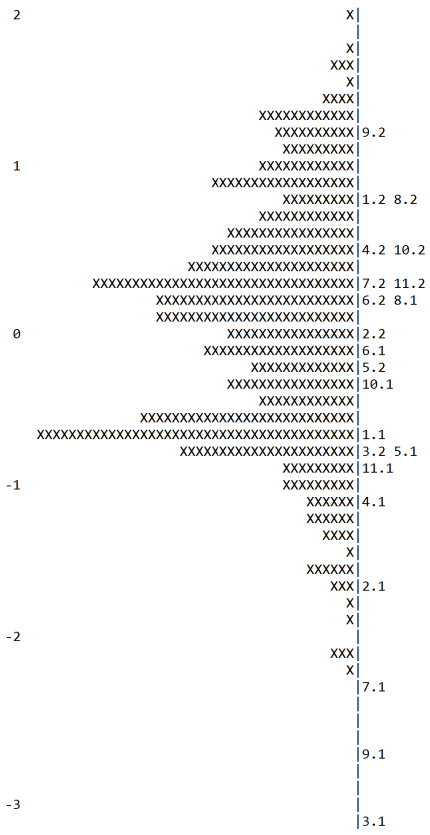
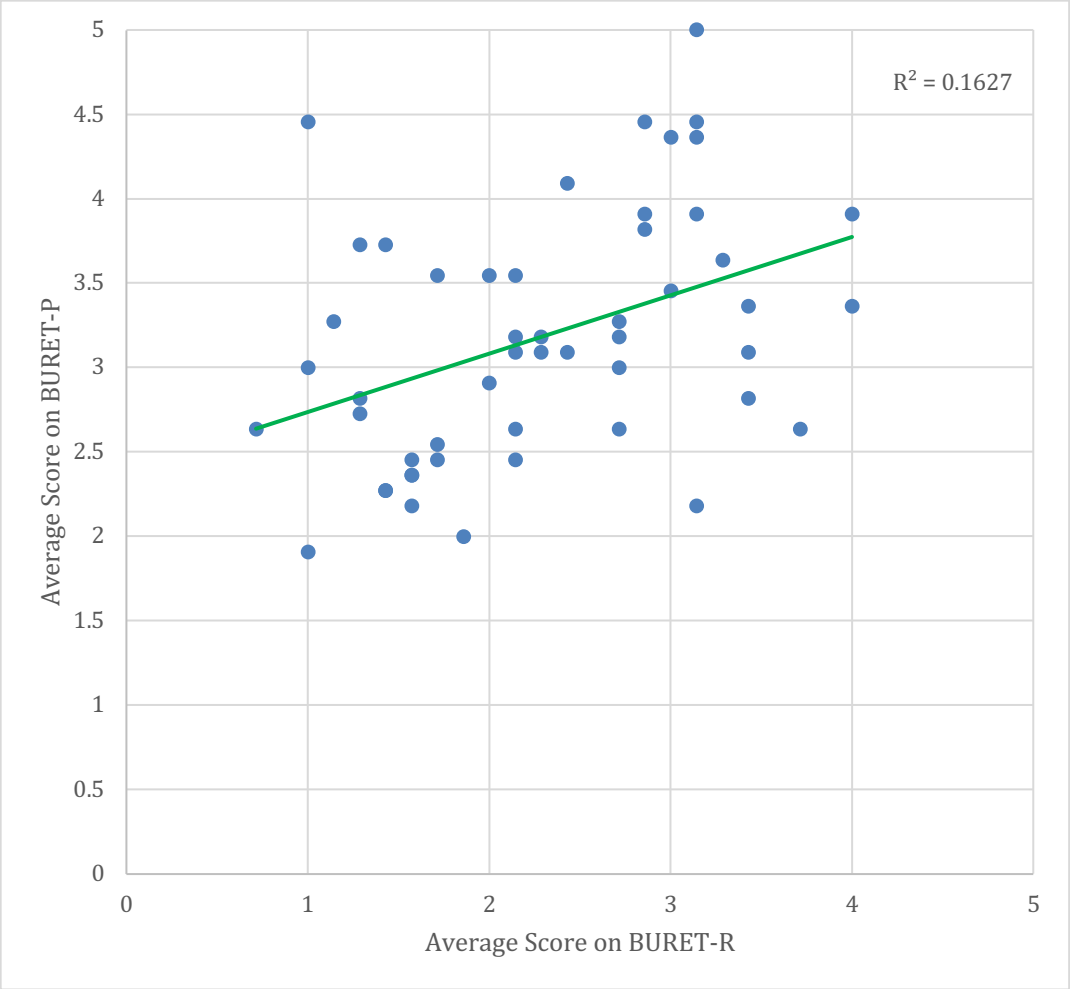


Fig. D2. Wright map for BURET-P

Appendix E – Scatterplot for BURET-R vs. BURET-P scores



Acknowledgements

This work was funded by the National Science Foundation Improving Undergraduate STEM Education initiative (NSF IUSE 1712001), and the Alfred P. Sloan Foundation (G-2016-7112). Additional support comes from the Berkeley Graduate School of Education Barbara Y. White Fund and a National Science Foundation Research Experiences for Teachers (RET) Site Award (NSF EEC-1542471). L.E.C. was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1106400. Thanks to Gabe Otero, Erica Dettmer-Radtke, Ibrahim Hajar, Seth Van Doren, Jiho Kim, Yuki Watanabe, Michelle Douskey, Zachary Firestein, Eddy Ham, Michelle Wilkerson, Vicky Laina for assistance with

data collection and Laura Armstrong for assistance with data analysis. Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily represent the views of these agencies.

References

Ashcroft, J., Blatti, J., & Jaramillo, V. (2020), Early career undergraduate research as a meaningful academic experience in which students develop professional workforce skills: A community college perspective. In *Integrating Professional Skills into Undergraduate Chemistry Curricula* (pp. 281-299). American Chemical Society. DOI: 10.1021/bk-2020-1365.ch016

Airey J. and Linder C., (2009), A disciplinary discourse perspective on university science learning: Achieving fluency in a critical constellation of modes. *J. Res. Sci. Teach. Off. J. Natl. Assoc. Res. Sci. Teach.*, **46**(1), 27–49.

Auchincloss L. C., Laursen S. L., Branchaw J. L., Eagan K., Graham M., Hanauer D. I., et al., (2014), *Assessment of course-based undergraduate research experiences: A meeting report*. *CBE-Life Sci. Educ.*, **13**(1), 29-40.

Avargil, S., Kohen, Z., & Dori, Y. J. (2020), Trends and perceptions of choosing chemistry as a major and a career. *Chem. Educ. Res. Pract.*, **21**(2), 668-684.

Bond, T. G., & Fox, C. M. (2007). Fundamental measurement in the human sciences. *Chicago, IL: Institute for Objective Measurement*.

Bonous-Hammarth M., (2000), Pathways to success: Affirming opportunities for science, mathematics, and engineering majors. *J. Negro Educ.*, 92–111.

Bransford J. D., Brown A. L., and Cocking R. R., (2000), *How people learn*, Washington, DC: National academy press.

Brown J. S., Collins A., and Duguid P., (1989), Situated cognition and the culture of learning. *Educ. Res.*, **18**(1), 32–42.

Butz A. R. and Branchaw J. L., (2020), Entering Research Learning Assessment (ERLA): Validity evidence for an instrument to measure undergraduate and graduate research trainee development. *CBE—Life Sci. Educ.*, **19**(2), ar18.

Carlone H. B. and Johnson A., (2007), Understanding the science experiences of successful women of color: Science identity as an analytic lens. *J. Res. Sci. Teach. Off. J. Natl. Assoc. Res. Sci. Teach.*, **44**(8), 1187–1218.

Carpi A., Ronan D. M., Falconer H. M., and Lents N. H., (2017), Cultivating minority scientists: Undergraduate research increases self-efficacy and career ambitions for underrepresented students in STEM. *J. Res. Sci. Teach.*, **54**(2), 169–194.

Chang M. J., Eagan M. K., Lin M. H., and Hurtado S., (2011), Considering the impact of racial stigmas and science identity: Persistence among biomedical and behavioral science aspirants. *J. High. Educ.*, **82**(5), 564–596.

Chang M. J., Sharkness J., Hurtado S., and Newman C. B., (2014), What matters in college for retaining aspiring scientists and engineers from underrepresented racial groups. *J. Res. Sci. Teach.*, **51**(5), 555–580.

Chase A. M., Clancy H. A., Lachance R. P., Mathison B. M., Chiu M. M., and Weaver G. C., (2017), Improving critical thinking via authenticity: the CASPiE research experience in a military academy chemistry course. *Chem. Educ. Res. Pract.*, **18**(1), 55–63.

Clark T. M., Ricciardo R., and Weaver T., (2016), Transitioning from expository laboratory experiments to course-based undergraduate research in general chemistry. *J. Chem. Educ.*, **93**(1), 56–63.

Coil D., Wenderoth M. P., Cunningham M., and Dirks C., (2010), Teaching the process of science: faculty perceptions and an effective methodology. *CBE—Life Sci. Educ.*, **9**(4), 524–535.

Connor M. C., Finkenstaedt-Quinn S. A., and Shultz G. V., (2019), Constraints on organic chemistry students' reasoning during IR and 1 H NMR spectral interpretation. *Chem. Educ. Res. Pract.*, **20**(3), 522–541.

Corwin L. A., Graham M. J., and Dolan E. L., (2015), Modeling course-based undergraduate research experiences: An agenda for future research and evaluation. *CBE—Life Sci. Educ.*, **14**(1), es1.

Corwin L. A., Runyon C. R., Ghanem E., Sandy M., Clark G., Palmer G. C., et al., (2018), Effects of discovery, iteration, and collaboration in laboratory courses on undergraduates' research career intentions fully mediated by student ownership. *CBE—Life Sci. Educ.*, **17**(2), ar20.

Corwin L. A., Runyon C., Robinson A., and Dolan E. L., (2015), The laboratory course assessment survey: a tool to measure three dimensions of research-course design. *CBE—Life Sci. Educ.*, **14**(4), ar37.

Crawford G. L. and Kloepper K. D., (2019), Exit interviews: laboratory assessment incorporating written and oral communication. *J. Chem. Educ.*, **96**(5), 880–887.

Cruz C. L., Holmberg-Douglas N., Onuska N. P., McManus J. B., MacKenzie I. A., Hutson B. L., et al., (2020), Development of a Large-Enrollment Course-Based Research Experience in an Undergraduate Organic Chemistry Laboratory: Structure–Function Relationships in Pyrylium Photoredox Catalysts. *J. Chem. Educ.*, **97**(6), 1572–1578.

Danczak S. M., Thompson C. D., and Overton T. L., (2020), Development and validation of an instrument to measure undergraduate chemistry students' critical thinking skills. *Chem. Educ. Res. Pract.*, **21**(1), 62–78.

Danowitz A. M., Brown R. C., Jones C. D., Diegelman-Parente A., and Taylor C. E., (2016), A combination course and lab-based approach to teaching research skills to undergraduates. *J. Chem. Educ.*, **93**(3), 434–438.

Dasgupta A. P., Anderson T. R., and Pelaez N., (2014), Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. *CBE—Life Sci. Educ.*, **13**(2), 265–284.

Dasgupta A. P., Anderson T. R., and Pelaez N. J., (2016), Development of the neuron assessment for measuring biology students' use of experimental design concepts and representations. *CBE—Life Sci. Educ.*, **15**(2), ar10.

Deane T., Nomme K., Jeffery E., Pollock C., and Birol G., (2014), Development of the biological experimental design concept inventory (BEDCI). *CBE—Life Sci. Educ.*, **13**(3), 540–551.

Doğan A. and Kaya O. N., (2009), Poster sessions as an authentic assessment approach in an open-Ended University general chemistry laboratory. *Procedia-Soc. Behav. Sci.*, **1**(1), 829–833.

Esparza, D., Wagler, A. E., & Olimpo, J. T. (2020), Characterization of instructor and student behaviors in CURE and Non-CURE learning environments: Impacts on student motivation, science identity development, and perceptions of the laboratory experience. *CBE—Life Sci. Educ.*, **19**(1), ar10.

Espinosa L., (2011), Pipelines and pathways: Women of color in undergraduate STEM majors and the college experiences that contribute to persistence. *Harv. Educ. Rev.*, **81**(2), 209–241.

Estrada M., Burnett M., Campbell A. G., Campbell P. B., Denetclaw W. F., Gutiérrez C. G., et al., (2016), Improving underrepresented minority student persistence in STEM. *CBE—Life Sci. Educ.*, **15**(3), es5.

Ghanem E., Long S. R., Rodenbusch S. E., Shear R. I., Beckham J. T., Procko K., et al., (2018), Teaching through Research: Alignment of Core Chemistry Competencies and Skills within a Multidisciplinary Research Framework. *J. Chem. Educ.*, **95**(2), 248–258.

Gilmore J., Vieyra M., Timmerman B., Feldon D., and Maher M., (2015), The Relationship between Undergraduate Research Participation and Subsequent Research Performance of Early Career STEM Graduate Students. *J. High. Educ.*, **86**(6), 834–863.

Gin L. E., Rowland A. A., Steinwand B., Bruno J., and Corwin L. A., (2018), Students who fail to achieve predefined research goals may still experience many positive outcomes as a result of

CURE participation. *CBE—Life Sci. Educ.*, **17**(4), ar57.
Goodey N. M. and Talgar C. P., (2016), Guided inquiry in a biochemistry laboratory course improves experimental design ability. *Chem. Educ. Res. Pract.*, **17**(4), 1127–1144.
Gormally C., Brickman P., and Lutz M., (2012), Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE—Life Sci. Educ.*, **11**(4), 364–377.
Griffeth N., Batista N., Grosso T., Arianna G., Bhatia R., Boukerche F., et al., (2015), An Undergraduate Research Experience Studying Ras and Ras Mutants. *IEEE Trans. Educ.*, **59**(2), 91–97.
Harackiewicz J. M. and Hulleman C. S., (2010), The importance of interest: The role of achievement goals and task values in promoting the development of interest. *Soc. Personal. Psychol. Compass*, **4**(1), 42–52.
Harsh J. A., (2016), Designing performance-based measures to assess the scientific thinking skills of chemistry undergraduate researchers. *Chem. Educ. Res. Pract.*, **17**(4), 808–817.
Harsh J., Esteb J. J., and Maltese A. V., (2017), Evaluating the development of chemistry undergraduate researchers' scientific thinking skills using performance-data: first findings from the performance assessment of undergraduate research (PURE) instrument. *Chem. Educ. Res. Pract.*, **18**(3), 472–485.
Hauwiller M. R., Ondry J. C., Calvin J. J., Baranger A. M., and Alivisatos A. P., (2019), Translatable Research Group-Based Undergraduate Research Program for Lower-Division Students. *J. Chem. Educ.*, **96**(9), 1881–1890.
Heath K. D., Bagley E., Berkey A. J. M., Birlenbach D. M., Carr-Markell M. K., Crawford J. W., et al., (2014), Amplify the Signal: Graduate Training in Broader Impacts of Scientific Research. *BioScience*, **64**(6), 517–523.
Heller S. T., Duncan A. P., Moy C. L., and Kirk S. R., (2020), The Value of Failure: A Student-Driven Course-Based Research Experience in an Undergraduate Organic Chemistry Lab Inspired by an Unexpected Result. *J. Chem. Educ.*, **97**(10), 3609–3616.
Hernandez P. R., Woodcock A., Estrada M., and Schultz P. W., (2018), Undergraduate research experiences broaden diversity in the scientific workforce. *BioScience*, **68**(3), 204–211.
Hogan K. and Maglienti M., (2001), Comparing the epistemological underpinnings of students' and scientists' reasoning about conclusions. *J. Res. Sci. Teach. Off. J. Natl. Assoc. Res. Sci. Teach.*, **38**(6), 663–687.
Hoskins S. G., Lopatto D., and Stevens L. M., (2011), The CREATE approach to primary literature shifts undergraduates' self-assessed ability to read and analyze journal articles, attitudes about science, and epistemological beliefs. *CBE—Life Sci. Educ.*, **10**(4), 368–378.
Kerr M. A. and Yan F., (2016), Incorporating course-based undergraduate research experiences into analytical chemistry laboratory curricula. *J. Chem. Educ.*, **93**(4), 658–662.
Killpack T. L. and Fulmer S. M., (2018), Development of a Tool to Assess Interrelated Experimental Design in Introductory Biology. *J. Microbiol. Biol. Educ.*, **19**(3).
Krim J. S., Coté L. E., Schwartz R. S., Stone E. M., Cleeves J. J., Barry K. J., et al., (2019), Models and Impacts of Science Research Experiences: A Review of the Literature of CUREs, UREs, and TREs. *CBE—Life Sci. Educ.*, **18**(4), ar65.
Krontiris-Litowitz J., (2013), Using Primary Literature to Teach Science Literacy to Introductory Biology Students. *J. Microbiol. Biol. Educ.*, **14**(1), 66–77.
Landis J. R. and Koch G. G., (1977), The measurement of observer agreement for categorical data. *biometrics*, 159–174.
Laungani R., Tanner C., Brooks T. D., Clement B., Clouse M., Doyle E., et al., (2018), Finding some good in an invasive species: introduction and assessment of a novel CURE to improve experimental design in undergraduate biology classrooms. *J. Microbiol. Biol. Educ.*, **19**(2).
Laursen S., (2019), Levers for Change: An Assessment of Progress on Changing STEM Instruction: Executive Summary, American Association for the Advancement of Science. https://www.aaas.org/sites/default/files/2019-07/levers-for-change-WEB100_2019.pdf

Laursen S., Hunter A.-B., Seymour E., Thiry H., and Melton G., (2010), *Undergraduate research in the sciences: Engaging students in real science*, John Wiley & Sons.
Lave J. and Wenger E., (1991), *Situated learning: Legitimate peripheral participation*, Cambridge university press.
Light C. J., Fegley M., and Stamp N., (2019), Emphasizing iterative practices for a sequential course-based undergraduate research experience in microbial biofilms. *FEMS Microbiol. Lett.*, **366**(23), fnaa001.
Lin T.-J., Lin T.-C., Potvin P., and Tsai C.-C., (2019), Research trends in science education from 2013 to 2017: a systematic content analysis of publications in selected journals. *Int. J. Sci. Educ.*, **41**(3), 367–387.
Linn M. C., (1995), Designing computer learning environments for engineering and computer science: The scaffolded knowledge integration framework. *J. Sci. Educ. Technol.*, **4**(2), 103–126.
Linn M. C. and Eylon B.-S., (2011), *Science learning and instruction: Taking advantage of technology to promote knowledge integration*, Routledge.
Linn M. C., Palmer E., Baranger A., Gerard E., and Stone E., (2015), Undergraduate research experiences: Impacts and opportunities. *Science*, **347**(6222), 1261757–1261757.
Linn M., Eylon B.-S., Kidron A., Gerard L., Toutkoushian E., Ryoo K., et al., (2018), Knowledge integration in the digital age: Trajectories, opportunities and future directions, International Society of the Learning Sciences, Inc.[ISLS].
MacFadden B. J., (2009), Training the Next Generation of Scientists about Broader Impacts. *Soc. Epistemol.*, **23**(3–4), 239–248.
Maltese A. V., Harsh J. A., and Svetina D., (2015), Data visualization literacy: Investigating data interpretation along the novice—expert continuum. *J. Coll. Sci. Teach.*, **45**(1), 84–90.
Mondisa J.-L. and McComb S. A., (2018), The role of social community and individual differences in minority mentoring programs. *Mentor. Tutoring Partnersh. Learn.*, **26**(1), 91–113.
Moon A., Zotos E., Finkenstaedt-Quinn S., Gere A. R., and Shultz G., (2018), Investigation of the role of writing-to-learn in promoting student understanding of light–matter interactions. *Chem. Educ. Res. Pract.*, **19**(3), 807–818.
Muna G. W., (2021), Stimulating Students' Learning in Analytical Chemistry through an Environmental-Based CURE Project. *J. Chem. Educ.*, **98**(4), 1221-1226.
Mundfrom D.J. , Shaw D.J., Ke T.L. (2005) Minimum Sample Size Recommendations for Conducting Factor Analyses, *International Journal of Testing*, **5**(2), 159-168.
Mutambuki J. M., Fynewever H., Douglass K., Cobern W. W., and Obare S. O., (2019), Integrating Authentic Research Experiences into the Quantitative Analysis Chemistry Laboratory Course: STEM Majors' Self-Reported Perceptions and Experiences. *J. Chem. Educ.*, **96**(8), 1591–1599.
National Academies of Sciences and Medicine, (2017), *Undergraduate research experiences for STEM students: Successes, challenges, and opportunities*, National Academies Press.
Nelms A. A. and Segura-Totten M., (2019), Expert–Novice Comparison Reveals Pedagogical Implications for Students' Analysis of Primary Literature. *CBE—Life Sci. Educ.*, **18**(4), ar56.
O'Donnell K., Botelho J., Brown J., González G. M., and Head W., (2015), Undergraduate research and its impact on student success for underrepresented students. *New Dir. High. Educ.*, **2015**(169), 27–38.
Olson S. and Riordan D. G., (2012), Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Report to the President. *Exec. Off. Pres.* https://files.eric.ed.gov/fulltext/ED541511.pdf
Opitz A., Heene M., and Fischer F., (2017), Measuring scientific reasoning–a review of test instruments. *Educ. Res. Eval.*, **23**(3–4), 78–101.
Ortiz N. A., Morton T. R., Miles M. L., and Roby R. S., (2020), What About Us? Exploring the Challenges and Sources of Support Influencing Black Students' STEM Identity Development in Postsecondary Education. *J. Negro Educ.*, **88**(3), 311–326.
Pagano J. K., Jaworski L., Lopatto D., and Waterman R., (2018), An inorganic chemistry laboratory

course as research. *J. Chem. Educ.*, **95**(9), 1520–1525.
Peteroy-Kelly M. A., Marcello M. R., Crispo E., Buraei Z., Strahs D., Isaacson M., et al., (2017), Participation in a year-long CURE embedded into major core genetics and cellular and molecular biology laboratory courses results in gains in foundational biological concepts and experimental design skills by novice undergraduate researchers. *J. Microbiol. Biol. Educ.*, **18**(1).
Reisner B. A. and Stewart J. L., (2020), The Literature Discussion: A Signature Pedagogy for Chemistry, in *Advances in Teaching Inorganic Chemistry Volume 1: Classroom Innovations and Faculty Development*,., ACS Publications, pp. 3–20.
Remich R., Naffziger-Hirsch M. E., Gazley J. L., and McGee R., (2016), Scientific growth and identity development during a postbaccalaureate program: Results from a multisite qualitative study. *CBE—Life Sci. Educ.*, **15**(3), ar25.
Robnett R. D., Chemers M. M., and Zurbriggen E. L., (2015), Longitudinal associations among undergraduates' research experience, self-efficacy, and identity. *J. Res. Sci. Teach.*, **52**(6), 847–867.
Rodenbusch S. E., Hernandez P. R., Simmons S. L., and Dolan E. L., (2016), Early engagement in course-based research increases graduation rates and completion of science, engineering, and mathematics degrees. *CBE—Life Sci. Educ.*, **15**(2), ar20.
Rodriguez J.-M. G., Bain K., Towns M. H., Elmgren M., and Ho F. M., (2019), Covariational reasoning and mathematical narratives: investigating students' understanding of graphs in chemical kinetics. *Chem. Educ. Res. Pract.*, **20**(1), 107–119.
Ryoo K. and Linn M. C., (2012), Can dynamic visualizations improve middle school students' understanding of energy in photosynthesis? *J. Res. Sci. Teach.*, **49**(2), 218–243.
Sadler T. D., Burgin S., McKinney L., and Ponjuan L., (2010), Learning science through research apprenticeships: A critical review of the literature. *J. Res. Sci. Teach. Off. J. Natl. Assoc. Res. Sci. Teach.*, **47**(3), 235–256.
Sato B. K., Kadandale P., He W., Murata P. M. N., Latif Y., and Warschauer M., (2014), Practice Makes Pretty Good: Assessment of Primary Literature Reading Abilities across Multiple Large-Enrollment Biology Laboratory Courses. *CBE—Life Sci. Educ.*, **13**(4), 677–686.
Schultz P. W., Hernandez P. R., Woodcock A., Estrada M., Chance R. C., Aguilar M., and Serpe R. T., (2011), Patching the pipeline: Reducing educational disparities in the sciences through minority training programs. *Educ. Eval. Policy Anal.*, **33**(1), 95–114.
Seymour E. and Hunter A.-B., (2019), *Talking about leaving revisited*, Springer.
Shanks R. A., Robertson C. L., Haygood C. S., Herdliksa A. M., Herdliska H. R., and Lloyd S. A., (2017), Measuring and Advancing Experimental Design Ability in an Introductory Course without Altering Existing Lab Curriculum. *J. Microbiol. Biol. Educ.*, **18**(1), 1–8.
Shultz G. V. and Gere A. R., (2015), Writing-to-learn the nature of science in the context of the Lewis dot structure model. *J. Chem. Educ.*, **92**(8), 1325–1329.
Shuster, M. I., Curtiss, J., Wright, T. F., Champion, C., Sharifi, M., & Bosland, J. (2019), Implementing and evaluating a course-based undergraduate research experience (CURE) at a Hispanic-Serving institution. *Interdiscip. J. Probl.*, **13**(2), 1.
Sirum K. and Humburg J., (2011), The Experimental Design Ability Test (EDAT). *Bioscene J. Coll. Biol. Teach.*, **37**(1), 8–16.
Stone E. M., (2014), Guiding Students to Develop an Understanding of Scientific Inquiry: A Science Skills Approach to Instruction and Assessment. *Cell Biol. Educ.*, **13**(1), 90–101.
Stone K. L., Kissel D. S., Shaner S. E., Grice K. A., and van Opstal M. T., (2020), Forming a Community of Practice to Support Faculty in Implementing Course-Based Undergraduate Research Experiences, in *Advances in Teaching Inorganic Chemistry Volume 2: Laboratory Enrichment and Faculty Community*,., ACS Publications, pp. 35–55.
Szteinberg G. A. and Weaver G. C., (2013), Participants' reflections two and three years after an introductory chemistry course-embedded research experience. *Chem. Educ. Res. Pract.*, **14**(1), 23–35.

Taber K. S., (2018), The use of Cronbach's alpha when developing and reporting research instruments in science education. *Res. Sci. Educ.*, **48**(6), 1273–1296.

Timmerman B. E. C., Strickland D. C., Johnson R. L., and Payne J. R., (2011), Development of a 'universal' rubric for assessing undergraduates' scientific reasoning skills using scientific writing. *Assess. Eval. High. Educ.*, **36**(5), 509–547.

Varela M. F., Lutnesky M. M. F., and Osgood M. P., (2005), Assessment of Student Skills for Critiquing Published Primary Scientific Literature Using a Primary Trait Analysis Scale. *Microbiol. Educ.*, **6**, 20–27.

Watts F. M., Spencer J. L., and Shultz G. V., (2020), Writing Assignments to Support the Learning Goals of a CURE. *J. Chem. Educ.*, **98**(2), 510-514.

White B., Stains M., Escriu-Sune M., Medaglia E., Rostamnjad L., Chinn C., and Sevian H., (2011), A Novel Instrument for Assessing Students' Critical Thinking Abilities. *J. Coll. Sci. Teach.*, **40**, 102–107.

White H. B., Benore M. A., Sumter T. F., Caldwell B. D., and Bell E., (2013), What skills should students of undergraduate biochemistry and molecular biology programs have upon graduation? *Biochem. Mol. Biol. Educ.*, **41**(5), 297–301.

White R. and Gunstone R., (2014), *Probing understanding*, Routledge.

Wiggins G., (1998), *Educative Assessment. Designing Assessments To Inform and Improve Student Performance.*, San Francisco, CA: Jossey-Bass Publishers.

Williams L. C. and Reddish M. J., (2018), Integrating primary research into the teaching lab: benefits and impacts of a one-semester CURE for physical chemistry. *J. Chem. Educ.*, **95**(6), 928–938.

Wilson M. Constructing measures: An item response modeling approach. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; 2005.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA press.