JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc Article

Development and Validation of Scientific Practices Assessment Tasks for the General Chemistry Laboratory

Norda S. Stephenson, Erin M. Duffy, Elizabeth L. Day, Kira Padilla, Deborah G. Herrington, Melanie M. Cooper, and Justin H. Carmel*

Cite This: J. Chem. Educ. 2020, 97, 884−893 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The development of proficiency in the practices used by scientists and engineers is considered an important student outcome of laboratory instruction. We developed tasks to assess students' use and development of selected scientific and engineering practices in the general chemistry laboratory using an adapted evidence-centered design approach. In this paper, we provide a detailed description of the process of development and validation of these assessment tasks, using one of our tasks to illustrate the process. The tasks show strong evidence of validity and reliability for revealing students' understanding of scientific and engineering practices within the research context.

KEYWORDS: *Chemical Education Research, First-Year Undergraduate/General, Laboratory Instruction, Testing/Assessment*

FEATURE: Chemical Education Research

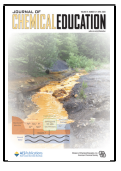
INTRODUCTION

The development of proficiency in the practices used by scientists and engineers is considered an important student outcome of laboratory instruction. The college science laboratory has been identified as an arena which has the potential to help students develop understandings of the ways that scientists behave.[1] With its longer duration, lower instructor−student ratio, and less formal environment, the laboratory is viewed as a place conducive to learning and providing opportunities for students to practice scientist-like behaviors such as asking questions, analyzing and interpreting data, constructing explanations, and argumentation.[2−10] However, there has been much imprecision in the vocabulary used to describe the actions and behaviors of scientists, resulting in various descriptions. For example, in addition to the behaviors mentioned above, there has been significant research examining students' proficiency in more amorphous ideas such as critical thinking,[11−13] problem-solving,[14] and inquiry.[15]

The *Framework for K-12 Science Education*[16] and the *Next Generation Science Standards* (NGSS)[17] describe eight scientific and engineering practices (SEPs) which detail the performances with which students are expected to gain proficiency in their K−12 experience. These practices—asking questions and defining problems, developing and using models, using mathematics and computational thinking, analyzing and interpreting data, engaging in argument from evidence, planning and carrying out investigations, constructing explanations and designing solutions, and obtaining, evaluating and communicating information—might be considered the disaggregated components of scientific inquiry, behaviors that scientists and engineers regularly negotiate. These behaviors embody what we want students to be able to do with their science content knowledge. While the *Framework* and the NGSS were primarily intended for a K−12 audience, curricula aligned with the principles set forth in these documents have been used with general chemistry students,[18,19] and the *Framework* has potential for application at all levels of chemistry education from K−20.[20−24] Table 1 provides a listing of the SEPs and their descriptions for grades 9−12.

The vision of the *Framework* and the NGSS is to use the SEPs as a means for students to demonstrate that they are able to apply knowledge for a more coherent and meaningful learning experience. Making determinations about students' understanding and proficiency in using the SEPs requires assessment that is well-suited to measuring these constructs. However, assessments have traditionally focused on what students know (the science content), rather than what they can do with that knowledge (the SEPs),[25−27] and so there is a great need for assessments that incorporate the SEPs. One major challenge with developing assessments targeting the SEPs is that the process is difficult and time-consuming.[28,29] The development of high-quality assessments of the SEPs requires an in-depth understanding of the practices themselves, knowledge of the inter-relationship between the practices and the chemistry content, and a "realization that practices are not merely pedagogical strategies".[28] While cognizant of the challenges of developing assessments targeting the SEPs, responding to the challenge also represents an opportunity to present a more complete and holistic picture of science that is more reflective of the scientific enterprise. There is also opportunity to influence

Received: September 24, 2019
Revised: February 17, 2020
Published: March 14, 2020



https://dx.doi.org/10.1021/acs.jchemed.9b00897
J. Chem. Educ. 2020, 97, 884−893

Table 1. Scientific and Engineering Practices and Their Descriptions[a]

Scientific and Engineering Practice	Brief Description of Practice
Asking questions and defining problems	Composing, refining, and evaluating empirically testable questions
Developing and using models	Constructing and applying models to predict and demonstrate relationships between systems and their components in the natural world
Planning and carrying out investigations	Designing and executing investigations that provide evidence for, and test questions and models (conceptual, mathematical, empirical, physical)
Analyzing and interpreting data	Comparing data sets for consistency, and using models to generate and analyze data
Using mathematics and computational thinking	Using mathematical and computational tools to analyze, represent, and model data
Constructing explanations and designing solutions	Constructing and revising explanations that are supported by multiple sources of evidence consistent with scientific ideas, principles, and theories
Engaging in argument from evidence	Using appropriate and sufficient scientific reasoning to defend and critique claims and explanations about the natural world
Obtaining, evaluating, and communicating information	Acquiring, assessing and communicating information, evidence, and ideas through multiple channels

[a]See ref 17.

curriculum development, instruction, assessment, and ultimately student learning.

The National Research Council's introduction of the eight SEPs in which scientists regularly engage has served to dispel much of the imprecision which has surrounded the things that scientists do. Elaboration of the SEPs brings a degree of consensus and clarity around a construct that was previously awash with ambiguity and uncertainty, and provides "an enhanced professional language for communicating meaning".[26] Having this clearer and more consensual understanding of the practices of scientists and engineers also allows for more systematic measurement and assessment.[24] Our work focuses on the development and validation of assessment tasks to measure students' use and development of the SEPs in the general chemistry laboratory. We have chosen to "foreground" the SEPs because they encapsulate the behaviors that we want students to demonstrate in a general chemistry laboratory and can therefore serve as standards for assessing what we want to happen in the laboratory. In this paper, we describe our process of development and validation for these items.

Assessment Design Framework

Evidence-centered design (ECD) has been described as a "principled framework for designing, producing, and delivering educational assessments".[30] The ECD approach is based on the idea that assessment is an evidentiary argument,[31,32] borrowing from Toulmin's model of argumentation[33] and Messick's[34,35] work on validity in assessment. As an evidentiary argument, assessment consists of claims about what students should know and do, evidence in the form of observations, behaviors, or performances that students have the desired knowledge, and tasks that the student will perform to communicate their knowledge or elicit the expected evidence.[36,37] The three central components of the ECD are the student model, the evidence model, and the task model.[31,38−40] Table 2 shows these critical components and provides a brief description of each.

The *Framework* and the NGSS support the use of an evidence-centered design approach for developing assessment tasks that target the SEPs. ECD can enhance the traditional assessment development process in a number of important ways through the following:

(i) a demand for specification of evidence (which allows for easy identification of areas in which students are strong or deficient)

(ii) a requirement for evidence forms which are directly linked to claims

Table 2. Central Components of the ECD and Their Descriptions

Central Component of ECD[b]	Brief Description of Component
Student Model	Specifies the knowledge/skills that we want students to acquire
Evidence Model	Points to specific performances, observations, and/or behaviors that provide evidence of students' proficiency in the desired knowledge
Task Model	Provides tasks with which students will engage to demonstrate their proficiency in the desired knowledge

[b]See refs 38−40.

(iii) an emphasis on students' "demonstrations" of what they know

(iv) a focus on improving the specificity and transparency of assessment and curricular materials developed[37,38]

Pellegrino[41] and Harris, Krajcik, Pellegrino, and McElhaney[42] have described how the ECD framework can be used to develop assessment tasks to measure science proficiency. While we also employ an ECD approach in our work, we use an adaptation of the ECD framework, focusing on the three main models of the ECD framework (i.e., student, evidence, and task models). Harris et al.[42] designed assessment tasks for use with middle school science students; we are focused on developing tasks to assess students' proficiency in the SEPs in the general chemistry laboratory.

METHODOLOGY

Research Design

This report on the development and validation of SEPs assessment tasks for the general chemistry laboratory forms part of a larger project which uses a mixed methods design in understanding students' use and proficiency in the SEPs in the general chemistry laboratory.

Below we detail the process that we employed in developing and validating these assessment tasks. To illustrate aspects of this process, we share from one of our tasks, the Combustion Task. Two additional tasks, along with a list of all the tasks developed and the SEPs addressed by each are included in the Supporting Information.

Convening the Project Team

A team of experts ($n = 7$) consisting of four professors and three postdoctoral researchers with over 40 combined years of

https://dx.doi.org/10.1021/acs.jchemed.9b00897
J. Chem. Educ. 2020, 97, 884−893

Table 3. Alignment between Scientific and Engineering Practices, Key Elements, and Evidence Statements

Scientific and Engineering Practice (SEP)	Key Element: Claim	Evidence Statements
Engaging in argument from evidence (EE)	Construct a scientific argument showing how data support a claim	Make a claim based on data
		Identify data that support claim
		Explain how data support claim
Analyzing and interpreting data (AI)	Analyze data systematically, either to look for salient patterns or to test whether data are consistent with an initial hypothesis	Use a graph, table, or equation based on data to make a claim about relationship between variables
		Given data, generate a plot/graph (with appropriate axes labels and units) to illustrate patterns

experience in chemistry education research guided the task development and validation process. The areas of expertise within the team include curriculum development, curriculum materials development, assessment development, laboratory learning, and quantitative and qualitative data analyses. All members of the team have intimate knowledge of the general chemistry curriculum. The establishment of this team (of experts) is consistent with guidelines and recommendations for establishing validity for the kinds of assessment tasks developed through this project.[40,43] The team provided content and construct validity evidence for the tasks as team members made judgments about the alignment between the key elements, evidence statements, prompts, and targeted responses.[44,45] Recursive feedback from the project team throughout the task development process continued until a general level of consensus was achieved. More details about the team are provided in the Supporting Information.

Defining the Research Parameters

With the establishment of the project team, we progressed to identifying the SEPs of interest from the *Framework* and NGSS, narrowing our focus to a few practices: planning and carrying out investigations, analyzing and interpreting data, constructing explanations, and engaging in argument from evidence. This focus on a few, rather than all the practices, was deliberate and consistent with our goal to develop *quality* items that provide valid and reliable data—a goal that requires substantial time for achievement. These four practices were chosen by our team of experts because they involve laboratory work and are practices that all instructors (traditional and nontraditional) are likely to expect their students to be able to demonstrate in a general chemistry laboratory environment.

Having identified our SEPs, we then outlined several criteria for the tasks we wanted to create. They needed to (i) align closely with the specific capabilities included in each practice outlined in the NGSS for grades 9−12, (ii) target the development of the SEPs, (iii) be embedded in chemistry contexts, and (iv) avoid heavy content-dependence.

We recognize that there are differences in the general chemistry laboratory curricula across institutions and so not all general chemistry students are exposed to the same experiments. To make our assessment tasks more equitable and accessible for a wider general chemistry laboratory population, we tried to provide support for any necessary chemistry content or techniques so that students who are not familiar with the contexts in which the tasks are embedded are not disadvantaged.

Deciding on Appropriate Evidence

We identified the statements about what students should know and be able to do (that is, the **key elements**) for each practice of interest for grade 12 from the NGSS and the *Framework*. We chose to use key elements for grade 12 as students near the end of their high school careers are not very different from students early in their university career. The identified key elements served as the **claims** in the student model of our ECD framework. However, given the relative complexity and bulkiness of these learning goals, they were unpacked to reveal smaller, more manageable **evidence statements** that pointed to specific performances that would provide the evidence that students possessed the desired understandings. In unpacking the key elements for each practice, we articulated specific performances, observations, and behaviors to be demonstrated by students that would serve as required evidence of proficiency with the practice.[7,42] We also considered what evidence is important and appropriate for proficiency at the level of general chemistry. Table 3 shows the relationship among the evidence statements, key elements, and the associated SEP.

Designing Tasks and Scoring Rubrics

Consistent with the ECD framework, we progressed to the designing of **tasks** with which the students would interact. Although we kept the contexts of the tasks largely familiar to those completing a general chemistry laboratory course, the tasks and prompts were designed to bear little resemblance to traditional assessment tasks that students were likely to encounter in their texts or online. Keeping in mind issues of cognitive complexity, vocabulary, and educational level, team members crafted initial drafts of tasks with prompts designed to elicit the specified evidence. Generally, initial iterations of tasks were written as multipart, open-ended prompts with simple sentence structures, and were designed to take no more than 15 min of students' time for completion.

Analytic scoring rubrics to evaluate students' responses on each prompt were developed alongside tasks. The development of the rubrics was informed by the evidence statements which specified the performance that would be accepted as evidence that a student had the requisite skill, the target responses (expected responses or performances written by the team), prior rubrics on practices, and levels of sophistication within student responses.[46,47] Student responses on earlier, more open-ended versions of the tasks served as the basis for the different levels of sophistication, highlighting what ideas students have and how they connect together to create a coherent expression.[48,49] The initial iteration of the Combustion Task is shown in Figure 1 below.

Task Testing and Refining

Task testing and refining is a crucial step in the task development process as it contributes to the validity process by providing content and response process validity evidence. Content validity evidence focuses on the extent to which the task prompts align with the construct (SEP) being measured, and includes procedures for prompt development and scoring such as use of an expert panel and pilot testing and revision.[45,50] Response process validity evidence "assesses the alignment between

https://dx.doi.org/10.1021/acs.jchemed.9b00897
J. Chem. Educ. 2020, 97, 884−893

Phlogiston theory is an early theory to explain combustion (i.e. what happens when substances burn). In brief, the theory states that all combustible substances contain a substance called phlogiston, and when these substances burn, phlogiston is released into the air.

A group of students carried out an experiment to investigate what happens when a piece of cleaned magnesium ribbon is burnt in air in a crucible.

i. Predict what will happen when the magnesium ribbon is burnt in air.
- Its mass will increase.
- Its mass will decrease.
- Its mass will remain unchanged.
- Other.

ii. Explain the reason for your choice.

The table shows some of the data the students collected.

	Mass (g)
Mass of crucible	15.216
Mass of crucible + cleaned magnesium ribbon	16.230
Mass of crucible after heating	16.906

iii. Do the data support or refute the phlogiston theory? Support your response with evidence.

Figure 1. Initial iteration of Combustion Task.

participant responses or performance and test construct",[45] and is often obtained through the use of cognitive interviews to ascertain that students' interpretations of prompts and the intended interpretations are in alignment.

The Combustion Task was designed primarily to elicit evidence of students' use and understanding of the engaging in argument from evidence SEP, particularly targeting evidence statements associated with constructing a scientific argument. With the exception of the first prompt which was deliberately written at a low level of difficulty and as a selected response item to help students feel at ease with the task, the prompts on the initial iteration of the task (Figure 1) were written as constructed response items.

Iteration 1 of the Combustion Task (Figure 1) was printed on Livescribe paper (digital pen and paper technology) and administered to a small group of students (n = 5) similar to the target population. Using a concurrent think-aloud protocol, students were instructed to talk aloud about their reasoning as they responded to the prompts using a Livescribe pen. The use of the Livescribe paper and pen allowed for students' written work to be recorded and connected exactly to their verbal expressions. After completing the think-alouds, students were asked to provide feedback through short semistructured interviews aimed at determining readability and clarity of each prompt. (Questions posed to students included: *Looking back at the questions you just answered, were you uncertain about what you were being asked to do in any of these questions? Were there any terms used in the wording of the questions with which you were unfamiliar or did not understand? How would you word this differently to convey the idea that. . ..?*). Student feedback responses were summarized immediately following the interviews, and were used to improve the readability and clarity of the prompts.

Analysis of student responses on this initial iteration of the task revealed that the prompt designed to aim at the heart of argumentation (prompt (iii), *"Do the data support or refute the phlogiston theory? Support your response with evidence"*, did not elicit responses that provided evidence of students' ability to construct a coherent scientific argument. Some students provided unhelpful responses (such as "yes" and "no") to the first part of the prompt (*"do the data support or refute the phlogiston theory?*), which suggested a need for greater clarity in the structuring of the prompt. In addition, students did not use the data provided as a part of the task to *"support your response with evidence"* as intended: only 1 of 5 students attempted to use the data provided to carry out any calculation to provide evidence in support of their claim. Three (3) of the 5 responses made no reference at all to the masses provided. However, student feedback indicated that the prompts were of adequate readability and clarity.

On the basis of our analysis of student responses, we revisited prompt iii —"*Do the data support or refute the phlogiston theory? Support your response with evidence*". The clarity issue which the team detected in the first portion of this prompt was fairly easily resolved by separating the question into parts so that *"do the data support the phlogiston theory"* was distinct from *"do the data refute the phlogiston theory"*. The team also revisited the second portion of prompt iii—"*support your response with evidence*". Previous work focused on designing prompts to elicit evidence of student reasoning has shown that novice learners are often unsure of what to attend to in answering questions. In investigating students' reasoning about acid−base reactions, Cooper et al.[48] found that it was necessary to separate the explanation prompt into two separate prompts (*describe* what is happening and *why* it is happening) in order to provide students with more structural cues about what was expected in their response, and to elicit

https://dx.doi.org/10.1021/acs.jchemed.9b00897
J. Chem. Educ. 2020, 97, 884−893

Table 4. Modifications Made on the First Two Iterations of Prompt (iii) of the Combustion Task

Prompt	Student Response Examples	Issue	Changes Made
		Iteration 1	
Do the data support or refute the phlogiston theory? Support your response with evidence.	"Yes"	Response was ambiguous and did not provide insight into student thinking.	This part of the prompt was revised to separate the support the theory and refute the theory portions, and force students into a clear choice.
	"No"		Do the data support the phlogiston theory? Refute the phlogiston theory?
	"The emission of a gas is usually followed by the substance becoming more dense."	No active use of the data to provide evidence.	We added clarity and simplified the prompt by breaking it into smaller parts.
	"The mass doesn't really change, it stays pretty similar."	Students seemed uncertain about how to provide evidence.	We were more explicit in our instructions to students to make use of the data; this resulted in the Iteration 2 prompt.
		Iteration 2	
Based on the data, what happened to the mass of the magnesium ribbon when it was burned in air?	—	—	We included this prompt in an attempt to focus students' attention on the data. Because of its purpose in the assessment, we presented it as a selected response prompt.
Describe the calculations you performed on the data (if any) to obtain evidence to support or refute the phlogiston theory. (You may find it helpful to show calculations.)	"Measure the crucible with other substances."	Although some students showed calculations, many wrote responses that did not provide evidence of their thinking.	We changed this prompt from a suggestion to an instruction to carry out a calculation. We also simplified the masses given in the table, using 3 significant figures rather than 5 (as we were not interested in testing students' computational skills, but rather their ability to use data to provide evidence).
	"Subtract to see how much was gained."		
	"Weigh ribbon after it combusted."	Students made errors with the numbers for masses, e.g., changing around the digits and incorrect subtraction/addition.	
	"Add"		
Do the data: Support the phlogiston theory? Refute the phlogiston theory?	—	—	This framing seemed to work quite well. It eliminated all ambiguous "yes/no" responses.
Explain how the data support or refute the phlogiston theory.	"The data should indicate loss of mass."	Students were not tying the data/evidence and the theory together.	We introduced sentence frames to help students look for connections. We also changed the ordering of the prompt pieces in the next iteration because we felt that the new order would be more consistent with students' approach to the question. These changes led to our final version, discussed in the text.
	"Because a few of the magnesium ribbon was released." [sic]		
	No response		

https://dx.doi.org/10.1021/acs.jchemed.9b00897
J. Chem. Educ. 2020, 97, 884−893

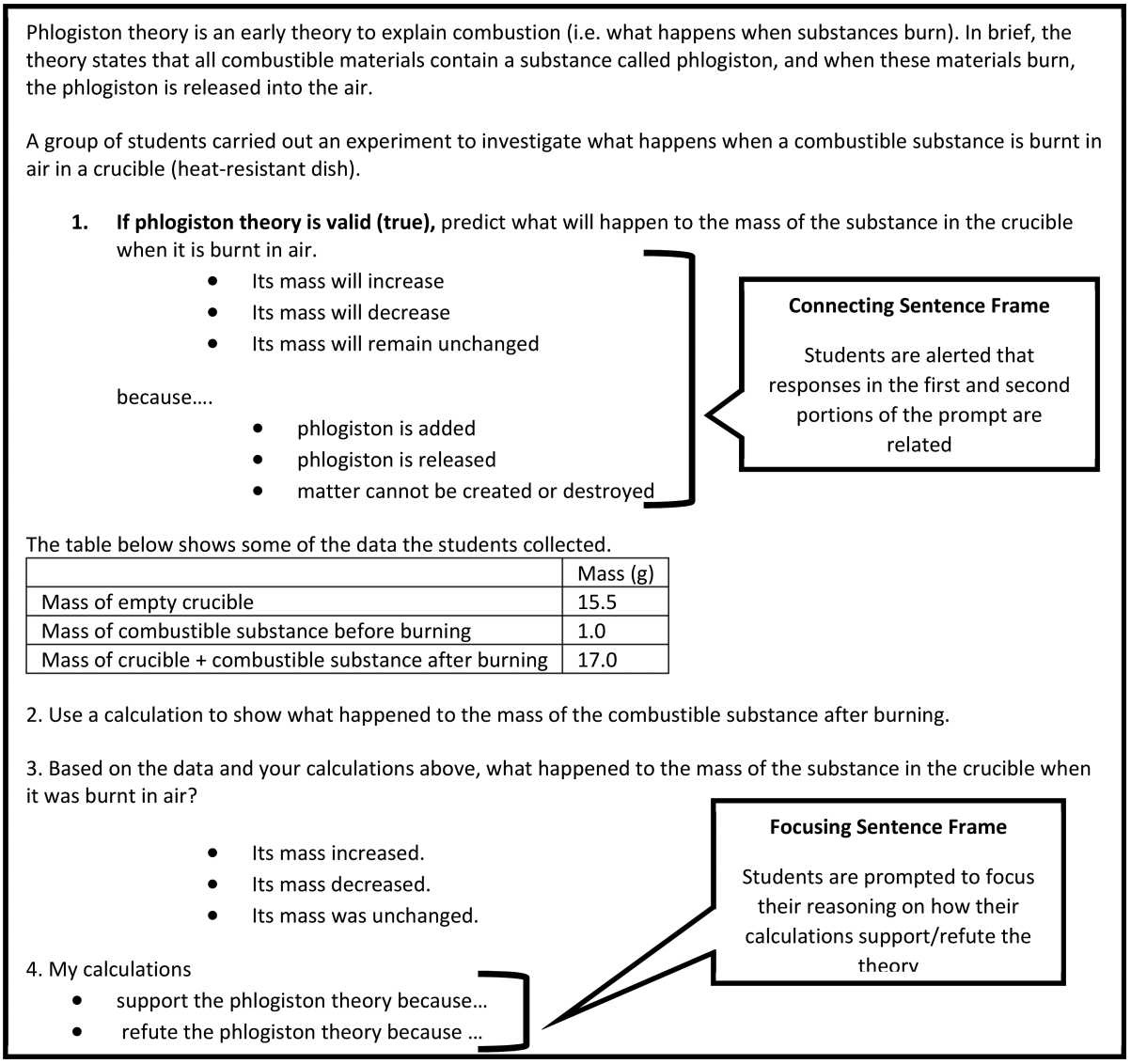


Figure 2. Final version of Combustion task showing use of sentence frames.

their thinking.[48] We reasoned that a similar type of revision was needed to our prompt as students appeared to find the claim to be complex, and so we revised this portion of the prompt to provide students with more structure about what we expected them to include in their responses. The need to provide students with more structure and explicit, careful scaffolding in designing assessments to elicit evidence of student thinking and understanding has also been underscored by a number of other authors.[46,51,52] We therefore divided the *"support your response with evidence"* prompt into

- "describe/show the calculations you performed on the data to obtain evidence to support or refute the phlogiston theory"
- "explain how the data support or refute the phlogiston theory"

After these modifications to the initial iteration of the Combustion Task, the task (Iteration 2) was administered to students similar to the intended target population. A random sample of 30 student responses was selected for closer examination by the project team. The team focused primarily on student responses to the modified prompt (*"describe/show the calculations you performed on the data to obtain evidence to support or refute the phlogiston theory"* and *"explain how the data support or refute the phlogiston theory"*). Examination revealed that more than half ($^{16}/_{30}$) of student responses did not make reference to or use the data provided when responding to the prompt *"describe/show the calculations you performed on the data to obtain evidence to support or refute the phlogiston theory"*, and of those who did, only 1 student actually showed a calculation. In fact, 4 students indicated that *"all calculations have already been made that I would make"* and *"no calculations were necessary"*. The intended calculation was a simple subtraction (17.0 g−15.5 g), which we would expect any student in general chemistry to know how to do. However, the fact that more than $^{1}/_{2}$ the students were not doing this indicated to us that our prompt as written was not adequately eliciting students' understandings. Student responses to *"explain how the data support or refute the phlogiston theory"* revealed that only $^{12}/_{30}$ students made a link between the calculation they were asked to perform and phlogiston theory. These observations led to revisions on Iteration 2 of the task. Table 4 summarizes the main modifications made on the first two iterations of this prompt and provides justification for these modifications. The multistage, recursive, and iterative process of

https://dx.doi.org/10.1021/acs.jchemed.9b00897
J. Chem. Educ. 2020, 97, 884−893

revision and testing was repeated until we had an iteration of the task that most students interpreted as intended and that elicited student reasoning. (A similar process was applied in the development of all tasks.)

The change in the ordering of the final prompt/set of prompts from the second iteration (see Table 4) to the final task version (Figure 2) was considered to be a more logical way of organizing the prompt and also made it more accessible to students. Asking students to carry out the calculation and then make a claim based on the result of that calculation provided a more straightforward approach than in the earlier iteration where students made a claim about the data and then performed the calculations to support that claim. The final task version also included the use of sentence frames as additional scaffolding for the task.[46] Focusing sentence frames help students to zero in on the pertinent aspects of the task, and connecting sentence frames prompt students to look for linkages among related components.[46] In this task we used focusing and connecting sentence frames whenever students failed to recognize that two prompts/parts were related. While the frames did not make the connection for students, they helped to focus their attention on key issues and provided them with better opportunities to demonstrate their proficiency in the practices. The use of sentence framing with the Combustion Task is highlighted in Figure 2.

The final task version also shows the simplification of the mass data that students were provided with (referred to in Table 4), as well as the removal of the "Other" option under prompt (i) *"Predict what will happen when the magnesium ribbon is burnt in air"* in the initial iteration. This option was removed as it was not an option that students selected over a number of iterations, indicating that it was not a good distractor. When compared with the initial iteration of the task, the final version incorporated a greater mix of selected and constructed response prompts. The use of constructed response prompts only, or a combination of constructed and selected response prompts for assessment tasks finds support in the literature; the sole use of selected response prompts is not considered adequate to capture students' thinking and understanding nor to provide strong evidence of students' engagement with the practices.[16,40,48,53] Selected response options were used in cases for which we observed that the majority of students gave a small number of relatively short answer options. For example, the writing of the second portion of prompt 1 (Figure 2) with selected response options arose out of the fact that when students (n = 17) were asked to explain the reason for their choice of one of the options in the first portion of the prompt, more than 75% of student responses surrounded phlogiston being released or added, or mass being conserved. Therefore, the selected response options were generated by a reduction of actual student responses (correct and incorrect) on the open-ended prompts after coding for frequently recurring responses. These student responses made the most suitable choices for selected response options as they use actual student language,[54] thereby contributing to the quality and authenticity of the design process.[46] A portion of the rubric for the final version of the Combustion Task is provided in Table 5. The full rubric can be found in the Supporting Information.

In preparation for full implementation of the final task, the task was administered to students similar to the intended population. Randomly selected responses (n = 100) were examined by the project team. This examination showed that on prompt 2 of the final task (Figure 2)—*"use a calculation to show what happened to the mass of the combustible substance after*

Table 5. Portion of Scoring Rubric for Combustion Task

Prompts	Evidence Statement	Level 0	Level 1	Level 2	Target Response
If phlogiston theory is valid (true), predict what will happen when the combustible substance is burnt in air. . .because. . .	Make a claim based on theory/model/data	Incorrect and Inconsistent; Correct but inconsistent	Consistent but incorrect	Consistent and correct	Its mass will decrease because phlogiston is released
Based on your calculation, what happened to the mass of the substance in the crucible when it was burned in air?	Make a claim based on theory/model/data	Inconsistent with calculation	—	Consistent with calculation	Its mass increased

https://dx.doi.org/10.1021/acs.jchemed.9b00897
J. Chem. Educ. 2020, 97, 884−893

burning"—71 students carried out mathematical calculations; while on prompt 4 *"my calculations support/refute the phlogiston theory because. . ."*, 64 student responses provided reasoning focused on changes in mass related to their calculations. Analysis of the responses to the prompt *"use a calculation to show what happened to the mass of the combustible substance after burning"* revealed that just about 25% of the responses were consistent with the targeted responses while another 20% were only partially consistent. The remaining responses were not consistent with the expected responses. This indicates to us that the scaffolding provided, while helping to elicit more student thinking, does not overestimate student thinking.

The final tasks were administered to General Chemistry I and II students across three institutions in the United States (one large research university in the Southeast, one large research university in the Midwest, and one large primarily undergraduate institution in the Midwest) via Qualtrics, an online survey platform. The tasks were administered during the first 2 weeks and the final 2 weeks of the semester.

Establishing Validity and Reliability

Anchoring our design process in a sound theoretical framework was a key step in maximizing content and construct validities. The ECD framework, with its attendant demands for claims, evidence, and tasks, is well-suited for assessing complex abilities and knowledge.[32,40] Content and construct validity were further established through close adherence to the descriptions of the SEPs set forth in the NGSS (Appendix F) and *Framework* documents, and the use of our team of experts in developing each task. Development of the prompts for each task required several rounds of negotiation among the members of the development team, consistent with the iterative nature of ECD. Response process validity was established through the use of think-alouds. The team and a cohort of students representative of the target population established face validity.

To establish the reliability of our rubric, five coders independently scored the same 10 sets of student responses chosen at random from our data set. Interrater reliability was computed using Fleiss' kappa, an extension of Scott's pi index for multiple raters, which measures the overall agreement probability among raters, adjusting for the probability of chance agreement.[55−57] Our initial round of scoring returned kappa values that indicated moderate to substantial agreement among the raters. Following discussion and refinement of our rubric, we then took a second set of 10 randomly chosen responses and repeated the scoring process. The resulting kappa values after this second round were between 0.75 and 1.00 for all prompts, indicating substantial agreement.[58]

SUMMARY

Students' development of proficiency in the practices used by scientists and engineers is an important outcome of laboratory instruction. We have developed assessment tasks that foreground the SEPs to measure students' use and development of the SEPs in the general chemistry laboratory. We have found the process of designing and developing assessment tasks that allow students to demonstrate their proficiency in the practices to be recursive, iterative, and complex. However, without high quality assessment tasks that demand evidence of what students are able to do, there will always be questions about whether meaningful learning has occurred or not.

In this paper we described how we used an adapted ECD approach in the development and validation of assessment tasks targeting the SEPs of constructing explanations and designing solutions, engaging in argument from evidence, analyzing and interpreting data, and planning and carrying out investigations for students in the general chemistry laboratory. The final versions of the tasks show strong evidence of validity and reliability, and are therefore more likely to reveal student thinking and understanding in the SEPs, leading to valid conclusions. In our next paper we will explore the levels of sophistication within students' responses on some of these assessment tasks and discuss students' proficiency in our practices of interest.

IMPLICATIONS

We believe that the tasks that we have developed are useful as research tools for assessing the effectiveness of interventions designed to influence students' proficiency in the SEPs, as well as for pinpointing particular areas of weakness within a practice for more targeted intervention. We also believe that they can provide important evidence to instructors regarding students' abilities to use particular SEPs that can be used to inform laboratory curriculum development and support for students as we move toward three-dimensional instruction.[16,17] Although this paper addresses the development and validation of these tasks for use in the general chemistry laboratory, the tasks may be used across instructional levels.

We do not present these tasks as substitutes for three-dimensional assessment tasks; they do not invalidate or replace the need for three-dimensional assessment tasks. Instructors may use these tasks as scaffolds to prepare students for three-dimensional assessments, or they can work alongside three-dimensional assessments in which instructors want to specifically target students' development and proficiency in the SEPs. Whether instructors use these assessments as preparatory or supplementary, we suggest that they be integrated into the curricula as part of a coherent framework for developing students' proficiency in using the SEPs. This is likely to require some reflection and perhaps reorganization as instructors consider how, where, and why particular tasks might be integrated into their existing laboratory curricula. Without a clear and explicit plan about how these assessments fit into their instruction, these tasks are unlikely to provide very useful feedback to instructors, and may send incorrect signals to students about what is valued in the course.

While the process of developing and validating the assessment tasks was iterative and required several revisions to identify the appropriate level of scaffolding, recognizing the importance of focusing and connecting sentence frames has streamlined this process for subsequent prompts. Additionally, although a common concern with the use of such scaffolding is that you are "giving the answer away" and though it is possible to overscaffold, student responses indicate that without the appropriate level of scaffolding we are not fully eliciting evidence of what they are able to do with respect to each practice. Moreover, even with scaffolding, students not proficient in the practices appear unable to provide consistent and/or complete responses.

LIMITATIONS

These assessment tasks were developed to be used with students at the level of general chemistry laboratory (or above) and have not been tested with other populations. The tasks are, by design, embedded in chemistry contexts and therefore may not readily

https://dx.doi.org/10.1021/acs.jchemed.9b00897
J. Chem. Educ. 2020, 97, 884−893

transfer across science disciplines. Additionally, validity and reliability can only be established for the data used in this study. As these tasks are not reliable or valid for all circumstances and populations, there is a need to re-establish reliability for each new data set with which the tasks are used.[59]

Having limited our scope to four of the eight SEPs, we do not currently have any tasks targeting the practices of asking questions, developing and using models, mathematics and computational thinking, and obtaining, evaluating, and communicating information. The development of tasks in these practices is a possible future direction for the project.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.9b00897.

Combustion task rubic (PDF, DOCX)
Conductivity task (PDF, DOCX)
Effusion task (PDF, DOCX)
Team description (PDF, DOCX)
List of tasks developed (PDF, DOCX)

■ AUTHOR INFORMATION

Corresponding Author

Justin H. Carmel − *Department of Chemistry & Biochemistry and STEM Transformation Institute, Florida International University, Miami, Florida 33199, United States;* orcid.org/0000-0001-9281-3751; Email: jcarmel@fiu.edu

Authors

Norda S. Stephenson − *Department of Chemistry & Biochemistry and STEM Transformation Institute, Florida International University, Miami, Florida 33199, United States;* orcid.org/0000-0002-5214-3578

Erin M. Duffy − *Department of Chemistry and SMATE, Western Washington University, Bellingham, Washington 98225, United States;* orcid.org/0000-0003-0073-9529

Elizabeth L. Day − *Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0002-8770-841X

Kira Padilla − *Department of Chemistry, Universidad Nacional Autònoma de Mèxico, Ciudad de Méxic̀o, Distrito Federal 04510, Mexico*

Deborah G. Herrington − *Department of Chemistry, Grand Valley State University, Allendale, Michigan 49401, United States;* orcid.org/0000-0001-6682-8466

Melanie M. Cooper − *Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0002-7050-8649

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.9b00897

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The project team thanks the Carmel and Underwood research groups at Florida International University for their feedback and input, the student participants for their responses and insights on early versions of the task, and the National Science Foundation for funding (No. 1708506, No. 1708666). Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect those of the National Science Foundation.

■ REFERENCES

(1) Hofstein, A.; Mamlok-Naaman, R. The laboratory in science education: The state of the art. *Chem. Educ. Res. Pract.* **2007**, *8* (2), 105−107.

(2) Fay, M. E.; Grove, N. P.; Towns, M. H.; Bretz, S. L. A rubric to characterize inquiry in the undergraduate chemistry laboratory. *Chem. Educ. Res. Pract.* **2007**, *8* (2), 212−19.

(3) Martin-Hansen, L. Defining inquiry. *Sci. Teach.* **2002**, *69* (2), 34−37.

(4) National Research Council. *National Science Education Standards*; National Academies Press: Washington, DC, 1996.

(5) National Research Council. *How students learn: History, mathematics, and science in the classroom*; National Academies Press: Washington, DC, 2005.

(6) Driver, R.; Newton, P.; Osborne, J. Establishing the norms of scientific argumentation in classrooms. *Sci. Educ.* **2000**, *84*, 287−312.

(7) Krajcik, J.; McNeill, K. L.; Reiser, B. Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Sci. Educ.* **2008**, *92* (1), 1−32.

(8) Hofstein, A.; Lunetta, V. N. The laboratory in science education: Foundation for the 21st century. *Sci. Educ.* **2004**, *88*, 28−54.

(9) Sampson, V.; Grooms, J.; Walker, J. Argument-Driven Inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study. *Sci. Educ.* **2011**, *95* (2), 217−257.

(10) Walker, J. P.; Sampson, V.; Southerland, S.; Enderle, P. J. Using the laboratory to engage all students in science practices. *Chem. Educ. Res. Pract.* **2016**, *17*, 1098−1113.

(11) Stephenson, N. S.; Miller, I. R.; Sadler-McKnight, N. P. Impact of Peer-Led Team Learning and the Science Writing and Workshop Template on the Critical Thinking Skills of First-Year Chemistry Students. *J. Chem. Educ.* **2019**, *96* (5), 841−849.

(12) Stephenson, N. S.; Sadler-McKnight, N. P. Developing critical thinking skills using the Science Writing Heuristic in the chemistry laboratory. *Chem. Educ. Res. Pract.* **2016**, *17*, 72−79.

(13) Gupta, T.; Burke, K. A.; Mehta, A.; Greenbowe, T. J. Impact of guided-inquiry-based instruction with a writing and reflection emphasis on chemistry students' critical thinking abilities. *J. Chem. Educ.* **2015**, *92* (1), 32−38.

(14) Shadle, S. E.; Brown, E. C.; Towns, M. H.; Warner, D. L. Rubric for Assessing Students' Experimental Problem-Solving Ability. *J. Chem. Educ.* **2012**, *89* (3), 319−325.

(15) White, B. Y.; Frederiksen, J. R. Inquiry, modeling, and metacognition: Making science accessible to all students. *Cogn. Instr.* **1998**, *16*, 3−118.

(16) National Research Council. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*; National Academies Press: Washington, DC, 2012.

(17) National Research Council. *Next Generation of Science Standards: For States, By States*; National Academies Press: Washington, DC, 2013.

(18) Cooper, M. M.; Klymkowsky, M. W. Chemistry, Life, the Universe and Everything: A New Approach to General Chemistry, and a Model for Curriculum Reform. *J. Chem. Educ.* **2013**, *90*, 1116−1122.

(19) Carmel, J. H.; Herrington, D. G.; Posey, L. A.; Ward, J. S.; Pollock, A. M.; Cooper, M. M. Helping students to "do science": Characterizing scientific practices in general chemistry laboratory curricula. *J. Chem. Educ.* **2019**, *96*, 423−434.

(20) Cooper, M. M. Chemistry and the next generation science standards. *J. Chem. Educ.* **2013**, *90* (6), 679−680.

(21) Cooper, M. M.; Caballero, M. D.; Ebert-May, D.; Fata-Hartley, C. L.; Jardeleza, S. E.; Krajcik, J. S.; Laverty, J. T.; Matz, R. L.; Posey, L. A.; Underwood, S. M. Challenge Faculty to transform STEM Learning. *Science* **2015**, *350* (6258), 281−282.

(22) Matz, R. L.; Fata-Hartley, C. L.; Posey, L. A.; Laverty, J. T.; Underwood, S. M.; Carmel, J. H.; Herrington, D. G.; Stowe, R. L.; Caballero, M. D.; Ebert-May, D.; Cooper, M. M. Evaluating the Extent of a Large-Scale Transformation in Gateway Science Courses. *Sci. Adv.* **2018**, *4* (10), eaau0554.
(23) Carmel, J. H.; Ward, J. S.; Cooper, M. M. A glowing recommendation: A project-based cooperative laboratory activity to promote use of the Scientific and Engineering Practices. *J. Chem. Educ.* **2017**, *94* (5), 626−631.
(24) Rodriguez, J. G.; Towns, M. H. Modifying Laboratory Experiments To Promote Engagement in Critical Thinking by Reframing Prelab and Postlab Questions. *J. Chem. Educ.* **2018**, *95* (12), 2141−2147.
(25) Bybee, R. NGSS and the Next Generation of Science Teachers. *J. Sci. Teach. Educ.* **2014**, *25*, 211−221.
(26) Osborne, J. Teaching Scientific Practices: Meeting the challenge of change. *J. Sci. Teach. Educ.* **2014**, *25*, 177−196.
(27) Stowe, R. L.; Cooper, M. M. Practicing what we preach: assessing "critical thinking" in organic chemistry. *J. Chem. Educ.* **2017**, *94* (12), 1852−1859.
(28) Pruitt, S. L. The Next Generation Science Standards: The Features and Challenges. *J. Sci. Teach. Educ.* **2014**, *25*, 145−156.
(29) Underwood, S. M.; Posey, L. A.; Herrington, D. G.; Carmel, J. H.; Cooper, M. M. Adapting assessment to support three-dimensional learning. *J. Chem. Educ.* **2018**, *95*, 207−217.
(30) Mislevy, R. J.; Steinberg, L. S.; Almond, R. G. *Evidence-centered assessment design*; Educational Testing Service: Princeton, NJ, 1999.
(31) Mislevy, R. J.; Haertel, G. D. Implications of Evidence-Centered Design for Educational Testing. *Educ. Measur. Iss. Pract.* **2006**, *25* (4), 6−20.
(32) Riconscente, M. M.; Mislevy, R. J.; Corrigan, S. Evidence-centered Design. In *Handbook of Test Development*, 2 nd ed.; Routledge: New York, NY, 2016; pp 40−63.
(33) Toulmin, S. *The Uses of Argument*; Cambridge University Press: Cambridge, EN, 1958.
(34) Messick, S. Validity. In *Educational Measurement*, 3 rd ed.; Macmillan: New York, NY, 1989; pp 13−104.
(35) Messick, S. *Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning*; Educational Testing Services: Princeton, NJ, 1994.
(36) Mislevy, R. J.; Steinberg, L. S.; Almond, R. G. On the structure of educational assessments. *Measur. Interdisc. Res. Persp.* **2003**, *1*, 3−67.
(37) Ewing, M.; Packman, S.; Hamen, C.; Thurber, A. Representing targets of measurement within evidence-centered design. *Appl. Measur. Educ.* **2010**, *23*, 325−41.
(38) Zieky, M. J. An introduction to the use of evidence-centered design in test development. *Psicol. Educ.* **2014**, *20*, 79−87.
(39) Pellegrino, J. W.; DiBello, L. V.; Brophy, S. P. The science and design of assessment in engineering education. In *Cambridge Handbook of Engineering Education Research*; Johri, A., Olds, B. M., Eds.; Cambridge University Press: Cambridge, UK, 2014; pp 571−578.
(40) National Research Council. *Developing Assessments for the Next Generation Science Standards*; National Academies Press: Washington, DC, 2014.
(41) Pellegrino, J. W. Measuring what matters: Challenges and opportunities in measuring science proficiency. *Australian Council on Educational Research Conference on Learning Assessments: Designing the Future*; Melbourne, Australia, Aug. 2015. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1263&context=research_conference (accessed 2020/02/17).
(42) Harris, C. J.; Krajcik, J. S.; Pellegrino, J. W.; McElhaney, K. W. *Constructing Assessment Tasks that Blend Disciplinary Core Ideas, Crosscutting Concepts, and Science Practices for Classroom Formative Applications*; SRI International: Menlo Park, 2016.
(43) Harsh, J. Designing performance-based measures to assess the scientific thinking skills of chemistry undergraduate researchers. *Chem. Educ. Res. Pract.* **2016**, *17*, 808−817.
(44) Buffum, P.; Lobene, E.; Frankosky, M.; Boyer, K.; Wiebe, E.; Lester, J. A practical guide to developing and validating computer science knowledge assessments with application to middle school. *SIGCSE* **2015**, 622−627.
(45) Severino, L.; Tecce DeCarlo, M. J.; Sondergeld, T.; Izzetoglu, M.; Ammar, A. A validation study of a middle grades reading comprehension assessment. *RMLE Online.* **2018**, *41* (10), 1−16.
(46) Kang, H.; Thompson, J.; Windschitl, M. Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Sci. Educ.* **2014**, *98*, 674−704.
(47) McNeill, K. L.; Krajcik, J. Scientific Explanations: Characterizing and evaluating the effect of teachers' instructional practices on student learning. *J. Res. Sci. Teach.* **2008**, *45* (1), 53−78.
(48) Cooper, M.; Kouyoumdjian, H.; Underwood, S. M. Investigating students' reasoning about acid−base reactions. *J. Chem. Educ.* **2016**, *93*, 1703−1712.
(49) Becker, N.; Noyes, K.; Cooper, M. M. Characterizing students' mechanistic reasoning about London dispersion forces. *J. Chem. Educ.* **2016**, *93* (10), 1713−1724.
(50) Cook, D.; Hatala, R. Validation of educational assessments: A primer for simulation and beyond. *Adv. Sim.* **2016**, *1*, 31.
(51) Jin, H.; Anderson, C. W. Developing assessments for a learning progression on carbon-transforming processes in socio-ecological systems. In *Learning Progressions in Science*; Alonzo, A. C., Gotwals, A. W., Eds.; Springer: Berlin, 2012; pp 151−181.
(52) Cooper, M.; Stowe, R. Chemistry Education Research—From Personal Empiricism to Evidence, Theory, and Informed Practice. *Chem. Rev.* **2018**, *118* (12), 6053−6087.
(53) Lee, H. S.; Liu, O. L.; Linn, M. C. Validating Measurement of knowledge Integration in Science Using Multiple-Choice and Explanation Items. *Appl. Meas. Educ.* **2011**, *24* (2), 115−136.
(54) Adams, W. K.; Wieman, C. E. Development and Validation of Instruments to Measure Learning of Expert-like Thinking. *Int. J. Sci. Educ.* **2011**, *33*, 1289.
(55) Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 29−48.
(56) Falotico, R.; Quatto, P. Fleiss' kappa statistic without paradoxes. *Qual. Quant.* **2015**, *49*, 463−470.
(57) Hallgren, K. A. Computing inter-rater reliability for observational data. *Tutor Quant. Methods Psychol.* **2012**, *8* (1), 23−34.
(58) Landis, J. R.; Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159−174.
(59) Bretz, S. Designing assessment tools to measure students' conceptual knowledge of chemistry. In *Tools of Chemistry Education Research*; Bunce, D., Cole, R., Eds.; American Chemical Society: Washington, DC, 2014; pp 155−168.

https://dx.doi.org/10.1021/acs.jchemed.9b00897
J. Chem. Educ. 2020, 97, 884−893