1-24-2022

Cross-disciplinary learning index: A quantitative measure of cross-disciplinary learning about energy

Todd Haskell
Western Washington University

Emily Borda
Western Washington University, emily.borda@wwu.edu

Andrew Boudreaux
Western Washington University, andrew.boudreaux@wwu.edu

Recommended Citation

Haskell, T., Borda, E., & Boudreaux, A. (2022). Cross-disciplinary learning index: A quantitative measure of cross-disciplinary learning about energy. Phys. Rev. Phys. Educ. Res., 18(1), 010108. https://doi.org/10.1103/PhysRevPhysEducRes.18.010108

Editors' Suggestion

Cross-disciplinary learning index: A quantitative measure of cross-disciplinary learning about energy

Todd Haskell
Department of Psychology, Western Washington University, Bellingham, WA 98225, USA

Emily Borda
Departments of Chemistry and Science, Mathematics, and Technology Education, Western Washington University, Bellingham, WA 98225, USA

Andrew Boudreaux
Department of Physics, Western Washington University, Bellingham, WA 98225, USA

(Received 18 August 2021; accepted 7 December 2021; published 24 January 2022)

The structure of many science programs at the college level assumes that students are able to draw on and integrate ideas from multiple disciplinary contexts. However, most assessment tools focus on learning in the context of a single discipline. We describe the development and validation of an instrument to measure how well students are able to combine energy ideas from different disciplines into a coherent understanding of a phenomenon. The final version of the instrument consists of a pair of multiple-choice online assessments, along with a metric calculated from the assessment scores: the cross disciplinary learning index (CDLI). The items on both assessments were found to have satisfactory psychometric properties for our sample. Furthermore, CDLI scores correlated with other relevant factors such as amount of science coursework. The CDLI is an easy-to-use metric that could be a useful component of program-level assessment for science, technology, engineering, and mathematics majors. Furthermore, the broader measurement approach, involving a pair of assessments set in different disciplinary contexts, provides a model for assessing cross-disciplinary learning that could be utilized for other cross-cutting scientific concepts.

DOI: 10.1103/PhysRevPhysEducRes.18.010108

I. INTRODUCTION

At our institution a student pursuing a degree in biology or geology must also take introductory courses in physics and chemistry. This requirement reflects the view that physics, chemistry, biology, and geology comprise subdomains of a larger, coherent approach to making sense of the natural world. For such coherence, explanations in physics, chemistry, biology, and geology must be grounded in the same principles and utilize the same concepts, even if the details differ. This view is reflected in the description of cross-cutting concepts presented by the Framework for K-12 Science Education: "These concepts help provide students with an organizational framework for connecting knowledge from the various disciplines into a coherent and scientifically based view of the world" (Ref. [1], p. 83).

In practice, however, students may face many barriers to turning this vision into reality. The siloing of academic disciplines creates the potential for students to emerge from an undergraduate science, technology, engineering, and mathematics (STEM) major with a fragmented rather than coherent vision of the natural sciences. Indeed, the Framework for K-12 Science Education alludes to this, noting that students are often left to their own devices to integrate knowledge across disciplines.

Having students complete foundational courses in multiple disciplines involves a major commitment of time and resources by students and institutions alike. To justify this investment, it would seem important to show that as students progress through such courses, they develop a coherent view of the STEM disciplines. In particular, students taking a later course should be able to productively draw on relevant concepts from an earlier course to help make sense of the new material they are encountering, even in cases in which the two courses are drawn from different disciplines. We refer to this as cross-disciplinary learning, and have provided a general overview of this construct in a previous article [2].

Determining the extent to which cross-disciplinary learning is actually fostered by courses and programs of instruction requires assessment tools that are appropriate for this purpose. However, perhaps in part due to the siloing

mentioned above, the literature on assessing how students integrate ideas across disciplinary boundaries is much less developed than the literature on assessing student understanding within a particular disciplinary context. The work that does exist, described in Secs. II and III, provides a useful starting place, but our review of existing assessment instruments failed to unearth any that were a good fit for measuring the kind of integration of ideas across disciplinary boundaries that we described above. Thus, we embarked on the process of developing our own.

To keep our project to a reasonable scope, we chose to focus on one particular concept that plays an important role across many science disciplines, namely, energy. Energy provides a shared framework for constructing explanations in all science disciplines [3]. Indeed, it functions as both a core idea and a cross-cutting concept in the Next Generation Science Standards [4]. Student learning of energy in one discipline, such as physics, should support student ability to productively apply energy concepts in a related discipline, such as chemistry.

The primary context for our work was a sequence of four science content courses for preservice K-8 teachers, focusing on physics, geology, biology, and chemistry. The physics course is a prerequisite for the other courses, which may be taken in any order. In each course, students develop explanatory models based on foundational energy concepts, including specific energy forms, as well as energy transfer, transformation, and conservation. The courses frame and present these basic energy constructs in a similar fashion, and make use of consistent representations. Because of the curricular coherence around energy ideas, we expected students in these courses would more readily develop a unified concept of energy than students taking more traditional courses that are not part of a coherent sequence. Thus, if our measure was able to detect cross-disciplinary learning at all, it should certainly be able to do so in this context, and therefore this course sequence served as a good test case.

We have previously defined cross-disciplinary learning and suggested approaches to measuring it [2]. The central goal of this particular project was to create an easy-to-use instrument for measuring this construct, and in this paper we describe the development and validation of an instrument for this purpose. In sharing this work, we also seek to highlight the value of using paired assessments—or even paired questions—set in different disciplinary contexts. We believe this is a key element for distinguishing between unified and fragmented understanding.

The remainder of the introduction discusses our formulation of cross-disciplinary learning, how it relates to other ways of thinking about application of previously learned knowledge, and why existing instruments were not appropriate for measuring it. Cross-disciplinary learning overlaps with many existing ideas in the learning literature, but is explicitly built around the context of sequenced courses from different disciplines. Then, in Sec. IV, we describe the development of the instrument and report preliminary results. We conclude by discussing implications of the data and potential uses of the instrument as well as the broader measurement approach we used.

II. WHAT IS CROSS-DISCIPLINARY LEARNING?

Thus far we have briefly mentioned the idea of a unified understanding of energy and why we believe it is important. But what exactly does this look like in the classroom? For purposes of illustration, we draw on a vignette we presented in an earlier article [2]. This vignette describes the experience of a fictional student Amara, who is beginning her second term at university. In her first term, she took courses in physics and chemistry. She is now in her second chemistry course, where she is learning about exothermic and endothermic chemical reactions. The professor presents the combustion of methane as an example of a highly exothermic reaction. In her previous physics course Amara learned about the law of conservation of energy (LCE), so when she is thinking about the combustion example, she is confused, because energy seems to be created by the burning. She concludes that this is not possible, and that there is an energy source she is not considering. She then hears her professor mention that the methane is storing energy. That makes Amara recall potential energy, something else she encountered in physics. In particular, she remembers an example of a skateboarder riding down a hill. In that example, potential energy changed form, becoming kinetic energy. She ponders whether the burning of methane also involves a transformation of potential energy. But she is not clear what kind of potential energy would be involved.

In this vignette, Amara uses ideas from her physics course as conceptual building blocks to help make sense of new content in her chemistry course. We refer to this process as "cross-disciplinary learning." Amara is confronted with a learning challenge in her new course, situated in a chemistry disciplinary context. This leads to the activation of knowledge elements from previous coursework, in this case ideas about energy conservation and potential energy from Amara's prior physics course. At the same time, information from the current course is also active, such as the observation of a temperature increase during methane combustion. Some of these activated knowledge elements become transformed as Amara thinks through the situation. For example, the idea that potential energy transforms into kinetic energy, originally learned in a physics context using the example of a skateboarder, is reframed in a more general form, in preparation for integrating that idea with the knowledge learned in chemistry, that heat is released when methane combusts. Finally, the individual knowledge elements are integrated to yield a new insight, that the combustion of methane involves the

transformation of potential energy into another form of energy. Tying all these pieces together, we have previously defined cross-disciplinary learning as "a process of sequential learning in which students activate, transform, and integrate knowledge from different disciplines, combining previous learning with new learning to construct new knowledge within a specific discipline" [2].

We find this earlier definition useful for general discussions of what cross-disciplinary learning is. At the same time, for purposes of creating an effective instrument, we found it necessary to be more specific about several elements of the definition, as well as how it is similar to and different from other constructs that have been proposed. To that end, in this section we discuss a number of theoretical ideas that have had an impact on our thinking, how they relate to cross-disciplinary learning, and how they impacted the design of the instrument.

A. What does cross-disciplinary mean?

Our definition of cross-disciplinary learning can be contrasted with how the term interdisciplinary has typically been used. In a typical interdisciplinary learning approach, students draw on concepts and approaches from multiple disciplines in the service of addressing a particular problem. In contrast, we were interested in learning and teaching that is clearly framed in the context of a single discipline, and which foregrounds disciplinary concepts rather than an issue or problem. For example, Amara is taking a chemistry course and thinking about the chemistry topics of exothermic and endothermic reactions. She draws on what she learned in physics to help her do that thinking, but it is clearly chemistry that is foregrounded. Some authors have used the term cross-disciplinary to express this idea. For example, it has been suggested that in a cross-disciplinary approach instructors might create parallel learning activities, each one situated within a specific discipline, but focused on concepts shared across disciplines [5]. The term has been used in a similar way in the context of science writing by middle-level students, where a distinction has been made between discipline-specific academic language, such as technical terms associated with a particular discipline, and cross-disciplinary academic language, which includes elements of writing that are associated with an academic register but not a specific discipline (e.g., "as a result of this process") [6].

Consistent with these uses of the term, in our work we have used "cross-disciplinary" to indicate a situation where (a) there is a concept that is shared by multiple related disciplines, (b) there is a learning activity in which the concept is framed and clearly situated within a single discipline, and (c) students draw on one or more previous learning experiences that occurred in other disciplinary contexts to help them productively apply the concept.

B. Transfer

In order for a student to exhibit cross-disciplinary learning, as defined above, they need to draw on prior learning in another discipline. Issues with cross-disciplinary learning could therefore arise either because a student struggles to apply a concept across disciplinary boundaries or because they never achieved mastery of that concept within the discipline where they first encountered it. It is important to distinguish between these two sources of difficulty, because the appropriate instructional response would differ. We have found ideas from the literature on transfer useful in thinking about how to do this.

Educators typically hope that students are able to take things they learned within a specific class and apply that learning in other contexts. When that kind of application occurs, we say that students exhibited successful transfer. Since cross-disciplinary learning involves applying previously learned knowledge in a new context, it is arguably a form of transfer. As the goal of transfer is so central to the individual and societal investment we make in education, there has been over a century of research on it (see Ref. [7] for a useful review). There is as much inconsistency as consistency in the findings of this research, which has led some authors to shy away from using the term altogether. Nevertheless, one distinction that has proved useful across a broad range of situations is between *near* transfer, meaning there is high similarity between the learning context and the application context, and *far* transfer, meaning the learning and application contexts are quite different.

In the case of cross-disciplinary learning, there are two main application contexts of interest: within discipline and across discipline. Following Dori and Sasson [8,9], we treat within-discipline application as an instance of near transfer and across-discipline application as far transfer. For our purposes, we are particularly interested in how transfer is affected by crossing a disciplinary boundary. As noted above, in order to ascertain this, it is not enough to only examine student performance in a far transfer situation. A student might fail to exhibit far transfer, but if they never mastered the relevant concept in the first place, it is possible that they would have also failed to exhibit near transfer. Our approach to addressing this issue is to give students two transfer contexts: one in the discipline where the concept was originally learned ("near" transfer) and another in a novel disciplinary context ("far" transfer). By comparing their performance in those two situations we can see to what extent disciplinary context is important, and by extension whether they have a fragmented or unified understanding of the concept.

Although we have found it useful to draw on the idea of near versus far transfer, it is important to note that cross-disciplinary learning differs from most traditional views of transfer. In particular, in cross-disciplinary learning new learning can take place at the time of application, in the

form of the transformation and integration of knowledge from different sources. This characteristic is discussed further below.

C. Mapping and abstraction

We have also drawn useful insights from research on one of the most heavily studied forms of transfer: analogical problem solving. Analogical problem solving involves drawing on the solution to a known problem as a basis for solving a novel problem. Although there are a great many theoretical proposals for how analogical problem solving works, one part of the process that has received considerable attention is constructing a mapping between elements of the known problem and elements of the new problem; see, e.g., Ref. [10].

For example, in the vignette presented earlier, Amara learned about potential energy through the example of a skateboarder rolling down a hill. In order to apply that knowledge to the case of methane burning, Amara would need to figure out how to map elements of the skateboarding scenario onto the situation of methane burning. To do that, she might start with the hill and look for some element of methane combusting that would correspond to that. However, it is unclear what that element would be. In order to successfully achieve a mapping, Amara needs to think about the skateboarding situation more abstractly. Rather than trying to map the hill directly to something in the methane scenario, she might think about rolling down the hill as a process that plays out over time. With this reframing, it is now possible to map onto a similar element of the methane scenario, as the combustion is also a process that plays out over time. It has been argued that identifying the appropriate level of abstraction is one of the fundamental challenges of successful mapping [10].

Cross-disciplinary learning does not have to include this kind of direct mapping from one situation to another. However, by definition it involves integrating concepts drawn from different disciplinary contexts. Successfully linking them may require some kind of modification of the concepts first. That modification could involve abstraction, but in order to avoid suggesting that this is the only kind of modification a learner might make, we have adopted the more generic term transformation.

D. Preparation for future learning

Although we have found certain concepts from the traditional transfer literature to be very helpful, as described in the previous two sections, in other ways the traditional way of thinking about transfer has felt like an obstacle to doing the kind of assessment we wished to do. Typically, transfer has been seen as the direct application of previous learning. However, cross-disciplinary learning also involves learning at the time of application. Specifically, the kind of transfer we are interested in involves gradually building knowledge over time through encounters with various learning opportunities. The new learning in one situation becomes the foundation for further learning in the next situation. Transfer has not typically been assessed in a way that reflects this idea of on-going learning. Bransford and Schwartz have noted that efforts to measure transfer have often involved what they call "sequestered problem solving" (SPS) [11]. In this approach, learners are asked to solve novel problems in the absence of any opportunity to seek out resources, propose and test ideas, or refine their thinking based on feedback. However, in the case of using knowledge gained in one course to help master content in a subsequent course, learners often do have these opportunities. Thus, SPS-style assessment may underestimate what learners in this situation are capable of.

Bransford and Schwartz contrasted the direct application view of transfer with what they call preparation for future learning (PFL). From a PFL perspective, it is preferable to focus assessment on how well students can learn new information and relate that new learning to prior learning [11]. In our view this framing of transfer more closely aligns with the dynamics of learning in the structured sequences of courses common in higher education than does a SPS-style assessment. Thus, for the portion of our instrument set in a novel disciplinary context, students were provided with new information about the unfamiliar context which could be integrated with previous learning to answer the questions.

E. Cognitive resources

We have described cross-disciplinary learning as activating, transforming, and integrating multiple pieces of knowledge [2]. This view is based on the resources theoretical framework, developed by multiple researches over the past three decades [12–15]. In the resources framework, learners activate small-grained units of knowledge to build larger units or construct an explanation. The construct of cross-disciplinary learning assumes resource activation as a primary mechanism for learning, and contrasts with classic transfer studies [10,16], where "what gets applied" is more fully formed and coherent (e.g., a complete solution paradigm for a previously learned problem type). Examples of documented conceptual resources related to energy include associating an energy form with a physical indicator, and accounting for energy quantities before and after a process [17].

In accordance with a resources perspective (as well as the PFL perspective discussed earlier), we do not assume that students learn an explanatory conceptual framework based on energy all at once in a particular course and then apply that framework wholesale in subsequent courses. Instead, understanding of energy is developed over time as students are exposed to and acquire relevant resources and develop facility combining previously acquired resources with resources encountered in a current learning context. For example, students may learn about gravitational potential

energy in a physics course and then activate it as a resource for making sense of chemical potential energy when they encounter this topic in a later chemistry course. In this view, inconsistency in reasoning can be accounted for through the strong context dependence of resource activation [18]. The resources perspective has important implications for the design of assessment, which we discuss in Sec. IV F on construction and validation of the instrument.

F. Interdisciplinary reasoning and communication

Earlier we clarified that cross-disciplinary learning involves foregrounding of concept(s) that cut across several disciplines, rather than foregrounding a problem that requires knowledge and skills from different disciplines to address. However, there are concepts related to the latter, problem-based approach, that are applicable when thinking about cross-disciplinary learning. Specifically, Shen *et al.* put forward a framework called interdisciplinary reasoning and communication (IRC) that is useful for describing what happens in interdisciplinary learning contexts [19]. This framework includes four processes: integration, translation, transfer, and transformation. There is considerable overlap between the IRC framework and cross-disciplinary learning. For example, integration involves the bringing together of knowledge from multiple disciplines to understand or explain a phenomenon, something that forms part of our definition of cross-disciplinary learning as well. Both cross-disciplinary learning and IRC also involve transformation, which in cross-disciplinary learning facilitates integration and broadening of concepts to allow their utility in new disciplinary contexts. As noted above, we were interested in their ability to think within a single disciplinary context, while making use of concepts from earlier courses that may have been situated in a different disciplinary context.

In our view, perhaps the most important difference between IRC and cross-disciplinary learning is that IRC is a framework while cross-disciplinary learning is a construct. By that we mean that IRC provides a way to characterize and describe the various processes that occur when people and ideas drawn from different disciplinary contexts are brought together. Such a framework is very useful in qualitative research, which is how Shen *et al.* apply it [19]. Our goal, however, was to develop a quantitative measure of a specific kind of learning. Thus, although we made use of the concepts of integration and transformation from IDL in conceptualizing our construct for assessment, we were unable to adopt the assessment method of Shen *et al.* for our study [19]. Next, we turn to a consideration of existing quantitative measures related to cross-disciplinary learning.

III. EXISTING INSTRUMENTS

There are few existing instruments that measure students' ability to reason across disciplines. Concept inventories such as the Force Concept Inventory [20] and Chemical Concept Inventory [21] measure students' understanding of concepts in the discipline and tend to be diagnostic in nature, incorporating choices that relate to known misconceptions. Though these are useful for measuring students' ideas formatively to guide instruction in those disciplines, these inventories have no relationship to each other, do not use common vocabulary, and are not focused on energy concepts, so we were unable to adopt them for our study.

The Interdisciplinary Energy Concept Assessment (IDEA) measures interdisciplinary learning by analyzing correlations across four different forms, which contain questions written in physics, biology, chemistry, and environmental science contexts [22]. Although this instrument does focus on energy, it is not solely focused on the conservation principle and it relies on discipline-specific factual knowledge. It was developed to measure content knowledge in the various disciplines, somewhat like a combined and correlated set of concept inventories, rather than measuring the skills of activation and integration we are interested in. The Energy Concept Assessment [23] is an instrument meant to assess energy concepts without a particular disciplinary focus. While this instrument measures foundational energy ideas, it is not necessarily meant to be used to measure cross-disciplinary learning, as it is not organized by disciplinary focus. It also relies on mathematical formulations of energy, which we wanted to avoid so that we could make our measure accessible across many levels of experience with science instruction. We also wanted to focus our measurement on students' conceptual understanding and reasoning skills, rather than their interpretation and use of mathematical models.

Shen *et al.* developed a task for assessing interdisciplinary learning around osmosis [24]. This model is meant to measure integrative skills needed to put knowledge from different disciplines together to solve a particular problem, but does not focus on the kind of sequential, cross-disciplinary learning we are interested in. Further, this task is open ended in nature and therefore requires qualitative data analysis. We wanted an instrument that would be relatively easy to administer and analyze, so that it would be of practical use to instructors. Finally, Herrmann-Abell and DeBoer developed an item bank to assess energy ideas [25]. This bank was not explicitly developed to measure understanding energy concepts across disciplines, nor was it focused solely on energy conservation. However, we were able to adapt some items from this assessment to include in the physics form for our measure. Below we describe our process for adaptation and authoring of items and discuss several pieces of evidence for validity of our instrument.

IV. METHODS

A number of authors have outlined a general process that can be used for developing assessment instruments; see,

e.g., Refs. [26–28]. We have used their recommendations as a guide, while also adapting them to reflect the fact that our instrument differs in important ways from both concept inventories and measures of student attitudes and beliefs.

A. Clearly specifying what is to be measured

Given how cross-disciplinary learning is defined, an instrument designed to measure it must address a topic that is relevant and important across multiple disciplines. As mentioned in the Introduction, we chose to focus on energy. Within the broader topic of energy, we focused on the law of conservation of energy, which we express as follows: The change in the total energy of a system is equal to the net transfer of energy to the system from its surroundings, or symbolically

$$\Delta E_{\text{tot}} = E_{\text{in}} - E_{\text{out}}. \quad (1)$$

An important facet of LCE is that it can be stated and conceptualized in relatively discipline-neutral terms, without reference to factual knowledge and vocabulary specific to a single science discipline. We acknowledge that the application of LCE to specific disciplinary contexts does require discipline-specific knowledge. We discuss later how we attempt to "teach" some of this specialized knowledge in the instrument itself, to allow the instrument to assess the ability to apply an energy model, as a unifying construct that spans familiar and unfamiliar disciplinary contexts, independent of specific disciplinary knowledge.

Cross-disciplinary learning involves application of concepts that were learned within one disciplinary context to a problem framed within a different disciplinary context. Thus, in addition to selecting a topic to focus on, it was necessary to select two disciplines, one to serve as the context for original learning and one to serve as the context of application. We chose physics and chemistry, respectively. These choices were motivated by the fact that our team included disciplinary experts in these two areas, as well as by the fact that in the teacher education sequence mentioned above, the physics course is a prerequisite for the other courses.

B. The goals for the instrument

Although a primary goal of our instrument was to measure the cross-disciplinary understanding of energy across our teacher education course sequence, we also wanted our instrument to be useful and practical across a variety of teaching contexts. Our intent was not just to create an assessment tool for researchers, but an instrument useful to practitioners in guiding instruction at the course and program levels. To achieve this goal, several criteria needed to be met. First, the instrument should be usable across a range of students, varying from very little to a great deal of science background. Second, the instrument should be brief, to avoid placing excessive demands on student time, especially in cases where it is administered to the same students more than once. Third, it should be quick and easy both to administer and to score.

These criteria impacted the design of the instrument in several ways. To satisfy the first criterion, it was necessary to minimize technical vocabulary and avoid content that is tightly tied to specific courses. Regarding the second and third criteria, while interviews or think-aloud protocols are very helpful for obtaining a rich understanding of the thinking of individual students, such approaches are not practical for scaling up to the course or program level. Thus, we chose to use multiple-choice items, structured in a way that permits them to be administered in an online format.

C. Articulating a measurement model

Cross-disciplinary learning is a latent, or hidden, variable [29], which we must attempt to measure using observed variables. Earlier we noted that measuring application in a new discipline alone is not adequate for assessing cross-disciplinary learning. To estimate this variable, it is necessary to measure a student's success in applying a concept in both the discipline where they originally learned it (here, physics) as well as in a novel discipline (here, chemistry) and compare their performance in these two cases. The two assessments we created for this purpose are called Energy Resources in Physics (EResPhys) and Energy Applications in Chemistry (EAppChem). The use of the terms *resources* and *application* is intended as a reminder that one assessment is situated in the original learning context while the other involves applying knowledge in a novel discipline.

The extent of cross-disciplinary learning can then be estimated by comparing student performance in the original discipline and the novel discipline. A student who struggles to demonstrate mastery of a concept in the original discipline, but effectively leverages that limited knowledge in a new discipline, is exhibiting productive cross-disciplinary learning. In contrast, a student who demonstrates strong mastery in the original discipline, but struggles to apply the concept in a new discipline, is exhibiting less productive cross-disciplinary learning.

There are two main challenges to making such a comparison in practice. First, the assessments are in two different disciplines and use different formats and scales (as we will describe later), so the scores are not directly comparable. Second, applying knowledge in an unfamiliar discipline is a more difficult task. That is, even if the scores were directly comparable, it would be surprising if a student who scored 80% on the assessment in the original discipline also scored 80% on the assessment in the novel discipline. Rather, judgments of performance in the novel discipline should be based on the level of performance that could reasonably be expected.

Our approach to these challenges was to use a set of paired scores to calculate a regression equation that yields a predicted score for the second assessment, given the observed score on the first assessment:

$$s_{2_{\text{pred}}} = a + b s_{1_{\text{obs}}}. \tag{2}$$

In Eq. (2), $s_{1_{\text{obs}}}$ is the observed score on the first assessment, $s_{2_{\text{pred}}}$ is the predicted score on the second assessment, and a and b are coefficients that can be estimated from the data. We used this regression equation to generate a predicted score for the EAppChem, based on a student's EResPhys score. We could then subtract the predicted EAppChem score from the observed score to get a difference measure, which we call the cross-disciplinary learning index (CDLI). A positive CDLI score means that a student scored higher on the EAppChem than is typical for students with a comparable EResPhys score. Similarly, a negative CDLI score means the student scored lower on the EAppChem assessment than typical, given their EResPhys score.

These ideas are illustrated in Fig. 1, which shows a scatter plot of some hypothetical data from the two assessments, with the regression line superimposed on the data. The green dot and the red dot represent comparable scores on the EAppChem. However, relative to the score on the EResPhys, the green dot represents unexpectedly strong performance, and thus a positive CDLI, while the red dot represents unexpectedly weak performance, and thus a negative CDLI.

If the EResPhys and the EAppChem tapped into exactly the same knowledge and skills, then we would expect the CDLI scores to simply be random noise or measurement error. However, the EAppChem was designed to require students to do something that is not required with the EResPhys—namely, the transformation and integration steps from the definition of cross-disciplinary learning. Thus, the CDLI scores should be influenced by how well a student can do those things, relative to their peers.

Note that in using this approach we are using a particular dataset to define what is a reasonable expectation for EAppChem performance. For example, if a student scores near the group average on the EResPhys, we would predict that they would score near the group average on the EAppChem as well. However, with a different dataset, both of those averages could change. Thus, CDLI scores should always be interpreted in the context of the dataset that was used to create the regression equation.

D. Creating candidate items

In creating items for the EResPhys and EAppChem, we first examined relevant existing assessments [25,30] for items to adopt or adapt. In doing so, we found most existing items assessing energy concepts did not assess energy conservation applications, were not situated in the appropriate disciplinary contexts, and/or focused on quantitative applications of energy conservation rather than conceptual reasoning. Regarding the last issue, as noted earlier, we wanted our items to be approachable by a broad range of students, and therefore we did not want success in answering them to hinge on quantitative skills. In light of these considerations, we authored most of the items ourselves, using existing assessments from the teacher education physics and chemistry courses as a starting point. All items ask the student to compare quantities of energy before or after an event, between different systems, between different energy forms, or to compare energy inputs, outputs, and changes. All questions are worded in

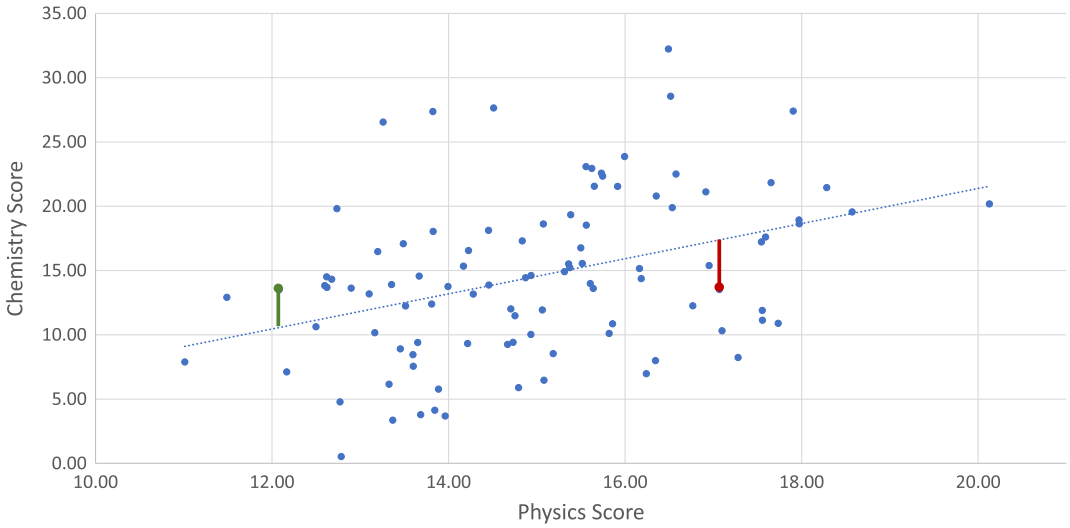


FIG. 1. Hypothetical scatter plot of EAppChem versus EResPhys scores. The green dot represents a student whose EAppChem score is higher than would be predicted based on their EResPhys score, resulting in a positive CDLI. The red dot represents a student whose EAppChem score is lower than predicted, resulting in a negative CDLI.

such a way as to force the choices into a finite number of possible energy comparison terms—e.g., "greater than," "less than," "the same as." Note that this is an important difference between our assessments and typical concept inventories: The distractors do not necessarily represent common misconceptions, but rather all the possibilities. Therefore, we did not use student interviews to generate the distractor options.

Consistent with the definition of cross-disciplinary learning, the EAppChem included some additional elements for the integration component. First, the questions were organized into four subcontexts: (1) atomic-level phenomena, (2) interatomic phenomena, (3) phase changes, and (4) chemical changes. Second, we inserted "instruction" in what energy forms are important for, and how they apply to, each subcontext, thus representing discipline-specific knowledge that must be combined with a more general understanding of LCE in order to successfully apply LCE across disciplines. This instruction takes the form of short narratives paired with 2–3 reading comprehension questions (RCQs) at the beginning of each subcontext. The RCQs contain feedback and are not included in the scoring of the instrument. Because this instruction does not contain explicit information about LCE, but essentially gives the student the tools to use to reason with LCE, the cross-disciplinary learning being measured by the EAppChem is the energy conservation principle applied to the new energy forms and context in which the participants were just "instructed."

E. Evaluating and revising the items

The candidate items underwent several rounds of piloting that included administration of a version of the assessment that had a written explanation field after each question. This version was given to students in the teacher education physics course and a traditional introductory physics course, as well as faculty members in physics (for the EResPhys) and chemistry (for the EAppChem). To further evaluate the construct validity of the assessments [31], the written explanations were then examined by at least one research assistant and one principal investigator for (a) incorrect explanations accompanying a correct choice, (b) correct explanations accompanying an incorrect choice, and (c) evidence for misunderstanding the statement of the question. Changes were made to questions with significant evidence of any of the above, and another round of piloting was done involving the same steps. Changes to wording to improve clarity were made throughout the piloting process.

After several iterations of piloting, students' written explanations matched the intention of the questions to a high degree, as evaluated by the observation that any remaining mismatches, which were few, were limited to one student and thus did not reveal generalizable patterns in student interpretation of the questions. Notably, this process was facilitated by the nature of the questions themselves, which involved proscribed answer choices describing a qualitative comparison of an energy quantity before and after an interaction or between two parts of a system. The process of assessing face validity thus depended almost entirely on students' interpretation of the question stems, not the choices. When we reached this stage, the content experts (faculty members) who took the assessments were also answering every item correctly. At this point, we administered the instruments without explanation fields to larger samples of students to generate quantitative measures of validity. For the EResPhys, the sample included students in the same two courses as before: teacher preparation and traditional physics ($N = 271$). For the EAppChem, the sample also included students from the geology, biology, and chemistry courses from the teacher education sequence, as well as more advanced students in both physics and chemistry ($N = 291$). There was substantial overlap between the samples, with 248 students being part of both. Although the majority of these students attended our own institution, we also included some students from a two-year college that was a partner on this project.

These samples were used to calculate item difficulty and a discrimination index, i.e., how well an item distinguished between low scorers and high scorers on the assessment.

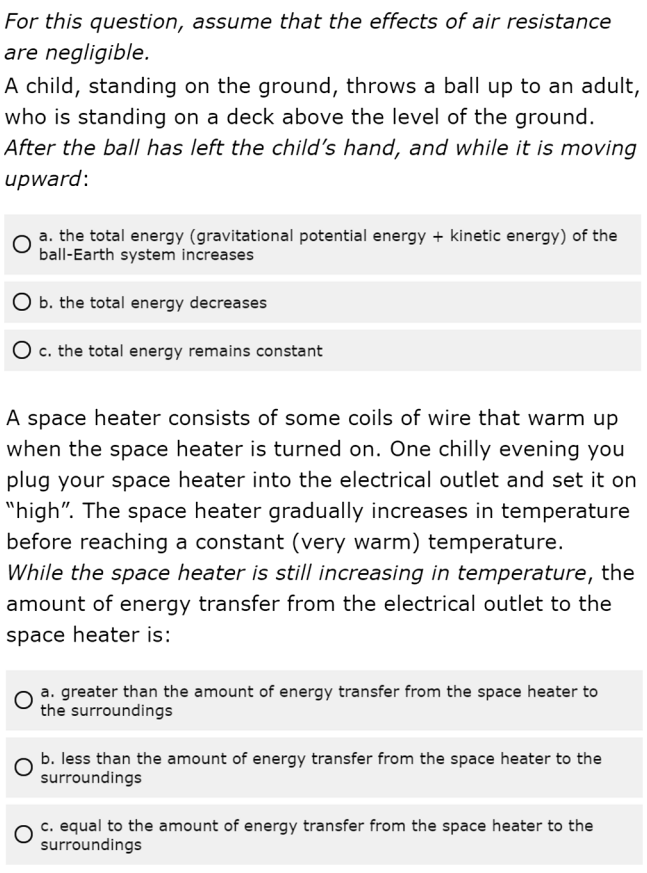


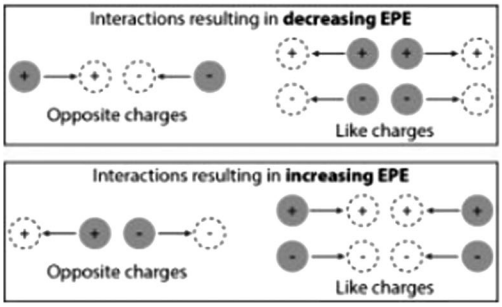
FIG. 2. Example items from the EResPhys instrument.

Percent correct responses was used as a measure of item difficulty, and point-biserial correlation was used for the discrimination index. None of the item difficulties had extreme values (all were between 0.2 and 0.8), suggesting that the items were neither too easy nor too difficult for this population. However, for the discrimination index, two of the items on the EResPhys and six of the items on the EAppChem had values below 0.2, suggesting that responses to these items were only slightly related to the overall score on the assessment. These eight items were excluded from the analyses below. The final versions of the EResPhys and EAppChem contain 14 and 10 questions (excluding RCQs), respectively.

Given the nature and purpose of our assessments, we did not use two reliability and validity criteria that are commonly used for concept inventories. First, on many instruments, distractors are chosen to reflect common misconceptions. Our assessments were not intended to identify or quantify misconceptions related to energy; rather, they focus on application of a principle (LCE) in the original disciplinary context, and then again in a novel disciplinary context. Because LCE involves a kind of energy accounting, appropriate answer options generally are specific quantities of energy, or a relationship between quantities of energy, e.g., less than, same as, more than. For most items this created a natural, symmetric set of answer options representing all possible choices, making it inappropriate to include only those answer options corresponding to specific misconceptions.

Second, it is very common for authors to report a measure of internal consistency for their instrument, such as Cronbach's alpha or KR20 [32]. These measures are

When an atom has a different number of electrons than protons, it is said to be an *ion.* An ion has an overall charge because of "extra" or "missing" electrons. The overall charges of ions govern how they will interact: two ions with opposite types of electric charge attract, while two ions with the same type of charge repel. As two ions interact, changes in *Electrostatic Potential Energy* (EPE) may occur. Attracting ions increase in EPE when they move *away* from each other, while repelling ions increase in EPE when they move *closer* to each other. Changes in EPE are shown in the diagram below.



When ions of the same charge (+ and + or - and -) move closer together the two-ion system:

- ◯ a) Increases in EPE
- ◯ b) Decreases in EPE
- ◯ c) Retains the same amount of EPE

A system consisting of two oppositely charged (+ and -) ions increases in EPE. They must be:

- ◯ a) Moving closer together
- ◯ b) Moving farther apart
- ◯ c) Impossible to tell from the information given

FIG. 3. Example of a teaching narrative and reading comprehension questions from the EAppChem instrument.

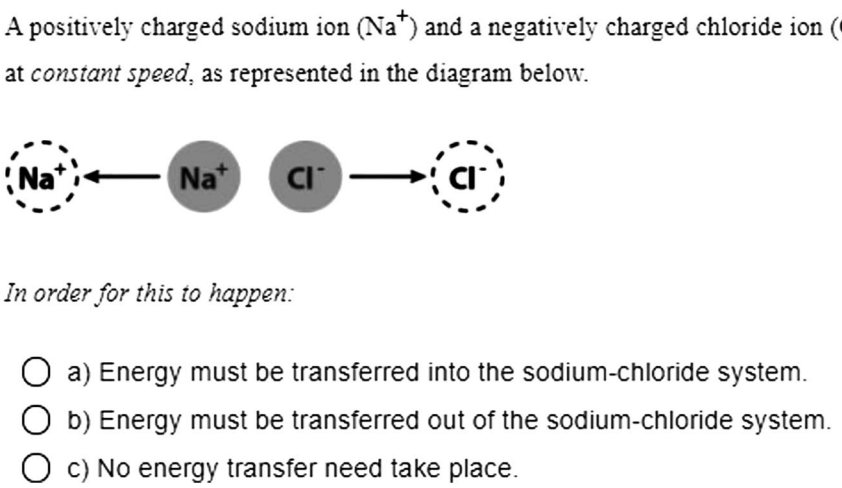


FIG. 4. Example item from the EAppChem instrument.

based on the assumption that each item in the measure is tapping into the exact same construct; i.e., in the absence of measurement error, one would expect the score for each item to be the same (what statisticians call tau equivalence [33]). For an assessment instrument, this assumption is very often violated, as such instruments generally include questions on multiple topics. Furthermore, as noted in the Introduction, we would not necessarily even expect strong correlations between student responses to questions that ask about the same energy concepts in superficially different ways. In fact, to the extent we are comparing performance in near-transfer and far-transfer contexts, we are deliberately setting up superficially different presentations of the same ideas with the expectation that student performance in the two contexts can and will diverge.

For that reason, measures of internal consistency can be misleading in this context. Taber provided a detailed review of appropriate and less appropriate ways that Cronbach's alpha has been applied in science education research, and recommended that "In developing instruments of this kind—tests, diagnostic instruments, concept inventories—researchers should carefully consider whether seeking a high value of internal consistency in the sense measured by alpha (i.e., equivalence across the set of items) is actually desirable in terms of their research aims" (Ref. [32], p. 1292). Beyond addressing the same general theme (LCE), our items were not designed to be equivalent, and so we did not expect or seek a particularly high alpha value.

F. Example items

To illustrate the structure and logic of the instrument pair, here we share example items from both instruments. Figure 2 illustrates two items from the EResPhys. Figure 3 shows one of the teaching narratives from the EAppChem, along with the corresponding reading comprehension questions. Note that real-time feedback was provided for these questions, but the questions were not included in the scoring of the instrument. Finally, Fig. 4 presents one of the scored items from the EAppChem. Note that although the instruments are intended to be a paired set, the individual items on one instrument are not directly paired with individual items on the other. The complete instruments are available upon request.

G. Increasing the sample size

Once we had settled on a set of items that had suitable psychometric properties, we sought further evidence regarding the validity of our assessments, such as whether the assessments correlate in sensible ways with other measures (convergent validity), and whether the assessment can distinguish groups that one would expect to perform differently (known groups validity). For this purpose a larger set of data was needed. In expanding our dataset, we continued to focus on students in teacher education and traditional physics courses. However, we also gathered data from students in an introductory psychology course, to check that the assessments are in fact usable with a broad range of students. At this point in the process we had begun to shift from an instrument development process to actually applying the instrument, so we report on these data in the next section.

V. RESULTS

A. Descriptive statistics

Currently, one or both of the assessments have been administered nearly 3000 times, including the initial samples described above. Tables I–III provide a general overview of these administrations. In the teacher education science courses (labeled SCED in Table I) as well as traditional physics, both the EResPhys and EAppChem were administered at the end of the term. For the physics courses only (teacher education or traditional), the EResPhys was also administered at the beginning of the term. This additional administration was not critical for our

TABLE I. Summary of unique EResPhys and EAppChem assessments completed, by course.

	Sample size	
Course	EResPhys	EAppChem
SCED Physics	1137	560
SCED Physics (2YC)[a]	203	95
SCED Biology	67	67
SCED Geology	54	54
SCED Chemistry	23	23
Introductory Psychology	56	60
Traditional Introductory Physics	1111	578
Upper-division Physics	0	17
Honors Introductory Chemistry	0	24
Upper-division Chemistry	34	34
Other	7	5

[a]2YC designates data collected from students at the two-year college.

project, but it was useful for the physics instructors, as a pretest to posttest comparison could be used to evaluate the effect of instruction.

As we transitioned from instrument validation toward collecting usable data, we added several demographics questions to the survey. Since we only have this information for a subset of the sample, the characteristics of the entire sample may differ somewhat from the numbers reported here. For each demographic statistic, we also indicate the number of students we have that information for. These numbers vary because different questions were added at different points in the process, and some students did not answer some of the questions.

We previously reported item difficulty and discrimination indices for our initial samples. To ensure that these values continued to fall in an acceptable range even with a larger and more diverse sample, we calculated them again for the expanded dataset. For the EResPhys, item difficulty ranged from 0.25 to 0.79, and the discrimination index ranged from 0.19 to 0.40. For the EAppChem, item difficulty ranged from 0.22 to 0.63, and the discrimination index ranged from 0.24 to 0.37. The data used to calculate these values included the initial samples discussed in Sec. IV, but excluded any data from items that were not part of the final measures. These values suggest that the instrument is likely to have acceptable psychometrics for a fairly broad range of populations.

TABLE III. Descriptive statistics for the two assessments and the cross-disciplinary learning index. Note that EResPhys and EAppChem statistics are in terms of percent correct, while the CDLI is a difference score, which can be either positive or negative and is expected to have a mean near zero.

Measure	N	Mean	Standard deviation
EResPhys	2661	49.8%	21.9%
EAppChem	1499	40.2%	23.1%
CDLI	1240	−0.02	18.6

B. Assessing validity

The construct of cross-disciplinary learning is more abstract than what typical assessment tools aim to measure. Although the questions on our assessments focus on energy, we were not trying to tap into student understanding of energy ideas *per se*, but rather the ability to reason with these ideas across multiple disciplinary contexts. This makes establishing the validity of the instrument more complex. However, the design of our assessments and the measurement model we adopted lead to several predicted patterns in the scores. Confirming that these predicted patterns can be observed in the actual data would support the interpretation of CDLI scores that we offered earlier. Here we step through each prediction in turn. Note that we used a significance level of 0.05 for all statistical tests reported in this paper.

1. Correlation between the EResPhys and EAppChem

Both the assessments involved questions about the same topic, namely, LCE. Thus, we would expect that a student who does well on one assessment would tend to do well on the other as well; i.e., the scores should be correlated. Our dataset included 1240 instances where both assessments were completed at the same time. 86% of these instances occurred with end-of-term administrations of the assessments in either traditional or teacher education physics courses, with the remainder spread across several other

TABLE II. Demographics of the sample. Note that most of our sample was currently in a physics course and may have varied in whether they decided to consider or not consider that course in making their response. In the latter part of data collection we clarified the question wording so that the appropriate response would be "no" for a pretest but "yes" for a posttest.

Gender ($N = 1134$)	Class standing ($N = 1428$)	Has taken college physics ($N = 1431$)	Has taken college chemistry ($N = 1437$)	Total college science courses ($N = 1364$)
74.8% female	15.5% freshmen	39.0% yes	42.9% yes	Mean of 4.2
23.5% male	33.3% sophomores	61.0% no	57.1% no	
1.7% nonbinary	34.5% juniors			
	16.7% seniors			

science courses. Using these paired data points, we regressed the EAppChem scores against the EResPhys scores. We found a significant positive correlation, $r = 0.57$, $F(1, 1239) = 594.77$, $p < 0.001$, $R^2 = 0.32$. To provide some context for interpreting this correlation, the math section of the SAT and grades in first year college math courses have a positive correlation of $r = 0.52$, quite similar to what we found. In contrast, the correlation between SAT math scores and grades in first year English courses is $r = 0.29$. [34]. Thus, the correlation we observed is consistent with what would be expected for two different measures of the same ability, and is higher than what we would expect if it was only due to the shared impact of more general factors like test-taking skills or motivation.

2. Impact of instruction

Although the primary use for the instrument pair is to measure cross-disciplinary learning, the EResPhys alone can be interpreted to assess students' understanding of LCE within a physics context. Thus, one would expect scores on the EResPhys to be impacted by physics instruction. As noted above, when data were collected from physics courses, whether teacher education or traditional, the EResPhys was administered as both a pretest and a posttest. Examination of class average pretest and posttest scores revealed that students performed significantly better on the posttest (52.2%) compared to the pretest (47.0%), $F(1, 2423) = 35.0$, $p < 0.001$, although the effect size was modest ($d = 0.24$). By way of comparison, a Cohen's d of 0.42 has been reported for engineering students on the Heat and Energy Concept Inventory for learning gain from beginning to end of a course addressing those topics [35]. It is worth nothing that students in teacher education physics courses exhibited much larger gains (from 43.7% to 51.3%, $d = 0.37$) than students in the traditional physics course (50.8% to 53.5%, $d = 0.12$). We do not know for certain why this occurred. Two possibilities are that the teacher education course has a heavier emphasis on the LCE, and that it utilizes a more conceptual teaching approach that matches well with the type of questions asked on the assessment.

3. Tapping into transformation and integration

Although the EResPhys and EAppChem assessments both focus on LCE, the items on the EAppChem ask students to combine prior learning with new content presented in the context of the assessment itself. This is an extra step that is not needed for the physics items, and was intended to correspond to the transformation and integration aspects of cross-disciplinary learning. Thus, the EAppChem scores should in part reflect students' facility with these higher-order thinking skills, above and beyond content knowledge, while the EResPhys scores should be less impacted by those skills.

To test this, we conducted a pair of regression analyses, one using EResPhys and EAppChem scores to predict year in college, the other using them to predict number of science courses completed. These demographic variables served as a general and a science-specific measure of how much academic experience the student has. If the EAppChem taps into the cross-disciplinary learning skills of integration and transformation more than the EResPhys, we would expect it to predict more of the variance in these variables. Note that for year in college, we coded freshman as 1, sophomore as 2, junior as 3, and senior as 4. In doing so, we assumed that the difference between a freshman and a sophomore is equivalent to the difference between a junior and a senior, which is probably not the case. However, the logic of the analysis still works so long as some variance can be predicted.

The results of the year-in-college analysis are shown in Table IV. EResPhys scores were a very weak but statistically significant predictor of year in college. By comparison, EAppChem scores were a much better predictor, though they were still a weak predictor in an absolute sense. This is hardly surprising, since year in college is an extremely broad and coarse-grained measure of academic accomplishment. The most important result comes when both scores were used as predictors simultaneously. EAppChem scores explained unique variance beyond that explained by the EResPhys scores. In contrast, EResPhys scores did not explain a statistically significant amount of variance beyond what EAppChem scores did.

The results of the science courses analysis are shown in Table V. Both EResPhys scores and EAppChem scores were significant predictors of number of science courses taken. The correlations were much larger than in the previous analysis, which makes sense given that there is a more obvious and direct relationship between the variables. Again, the most important result is that the EAppChem scores continued to have predictive power even when both scores were used as predictors simultaneously. However, the EResPhys scores added no additional predictive power beyond the EAppChem scores.

These analyses support the claim that the EAppChem measures something beyond what the EResPhys does, and

TABLE IV. Predicting year in college.

Each assessment on its own			
Measure	r	Significance test	Effect size
EResPhys	0.07	$F(1, 1367) = 5.72$, $p = 0.017$	$R^2 = 0.003$
EAppChem	0.18	$F(1, 769) = 26.82$, $p < 0.001$	$R^2 = 0.03$

After controlling for the other assessment		
Measure	R^2_{change}	Significance test
EResPhys	0.004	$F(1, 709) = 3.23$, $p = 0.073$
EAppChem	0.03	$F(1, 709) = 22.26$, $p < 0.001$

TABLE V. Predicting number of science courses.

Each assessment on its own			
Measure	r	Significance test	Effect size
EResPhys	0.21	$F(1, 1305) = 60.1,\ p < 0.001$	$R^2 = 0.04$
EAppChem	0.38	$F(1, 760) = 131.4,\ p < 0.001$	$R^2 = 0.15$

After controlling for the other assessment		
Measure	R^2_{change}	Significance test
EResPhys	0.001	$F(1, 702) = 0.64,\ p = 0.43$
EAppChem	0.11	$F(1, 702) = 85.6,\ p < 0.001$

that what is being measured is related to academic experience but not closely tied to content knowledge. We cannot know for certain that this something is transformation and integration ability; the nature of a latent variable limits us to inference rather than deduction. However, the main difference in the design of the two assessments is whether transformation and integration is required or not. Thus, the most plausible inference is that what is being captured by the EAppChem scores relates to the need to engage in cross-disciplinary learning.

C. Assessing utility

It was our goal was to create an instrument that not only measured what it is intended to measure, but also to maximize the utility of the instrument. Two criteria were considered in examining the utility: that the assessment pair be usable with a broad range of students, and that it be short enough that it could be incorporated into classes without creating a significant burden for students. Usability with a range of students was assessed by examining the range of scores for different student populations and analyzing the item difficulty. Group mean scores on the EResPhys range from 43.6% for students in teacher education physics taking the assessment as a pretest to 68.1% for students in teacher education chemistry taking the assessment as a posttest. Thus, this assessment is sensitive enough to detect meaningful differences between groups, while still leaving room for both lower and higher scores if used with more "extreme" groups than we collected data from (e.g., high school physics students, graduate-level physics students). Group mean scores on the EAppChem range from 27.0% for psychology students (typically novices for the content on the measures) to 80.0% for upper division physics undergraduates. Thus it appears the EAppChem is a somewhat more sensitive assessment than the EResPhys. However, the EAppChem might not be able to detect differences between college undergraduates with minimal science instruction and younger groups, since the psychology students were effectively at chance on the assessment. In principle, there is room for more advanced students (e.g., graduate students in physics or chemistry) to score higher on the scale.

Item difficulty was already mentioned above, but to reiterate, for the EResPhys it ranged from 0.25 to 0.79, while for the EAppChem it ranged from 0.22 to 0.63. Item difficulty also gives us information to assess the range of populations that the assessments can give useful information for. Varying item difficulties are desirable for an assessment intended to be used with a broad range of students, as it allows the assessment to continue to differentiate between groups even at the low and high ends of the range.

Brevity is also an important aspect of the utility of an instrument. Any instrument that is too long will not be used by instructors, and is at risk of generating assessment fatigue, and thus untrustworthy data, from students. As discussed earlier, we removed poorly performing items early in the process of developing the assessments. We also examined completion time once we moved toward online administration of the assessments. Completion times varied widely. However, most students spent between 10 and 15 minutes on individual assessments when administered in isolation but 20 minutes when the two assessments were administered as a package. Thus, there is a time savings when the assessments are given together, and even the combined assessments place only a modest demand on student time. Although there is less control over administration of the assessments in an online environment, we have been able to find numerous meaningful patterns in the data. The disadvantages of online administration probably do add more "noise" to the scores, but the ability to collect data from large numbers of students offsets this. However, caution should be exercised in using the scores for individual students, such as for placement purposes, or as a basis for differentiated instruction. The fact that the assessments produce useful data at the group level does not mean they are appropriate for use at the individual level.

D. Preliminary application of the CDLI: Relation to amount of science education

The previous section presented several analyses that tested predictions emerging from the design of our assessments and the measurement model used for the CDLI. The general pattern in the data is consistent with the proposed interpretation of CDLI scores as a measure of cross-disciplinary learning. This remains a tentative conclusion and more work is needed to continue the validation process. However, we believe that to motivate that additional work it is helpful to present an illustration of how the CDLI can be applied to address practical issues.

As mentioned earlier, CDLI scores must be interpreted in the context of the sample used to create the regression equation. That is, they are norm referenced, with the score of a given student resulting from comparison with other students. This is in contrast to a criterion-referenced

instrument, where students are compared to an external benchmark such as a set of learning objectives. Thus, the most natural way to use CDLI scores is for comparing different subgroups of students.

Building on our earlier analysis using raw assessment scores, we examined the relationship between number of college science courses taken and a student's CDLI score, which combines information from the two assessments. For this purpose, we calculated CDLI scores based on the 1240 instances in our dataset where we had scores on both assessments from the same administration (this is the same set of data points we used for examining the correlation between EResPhys and EAppChem scores). Of those 1240 instances, we had data on number of college science courses for 705 of them.

There was a significant correlation between courses taken and CDLI score, $r = 0.32$, $F(1, 703) = 81.54$, $p < 0.001$, $R^2 = 0.10$. Since the CDLI effectively controls for amount of content knowledge, this suggests that taking multiple science courses improves thinking and reasoning skills that support productive application of that knowledge.

This conclusion is neither novel nor surprising. However, it helps illustrate the value of the approach used to generate the CDLI scores. In particular, we are able to make meaningful comparisons between students with different educational histories by factoring out the role of content knowledge. This can be useful in situations where variability in background cannot be easily controlled, as occurs with transfer students or students pursuing different specializations within a major.

VI. DISCUSSION

College students majoring in the natural sciences are frequently asked to take foundational courses from multiple science disciplines. Realizing the full benefits of this investment requires that students are able to productively draw on concepts first learned in one disciplinary context in order to understand new concepts introduced in a different disciplinary context. We have labeled this cross-disciplinary learning. Despite its practical significance, there have been few efforts to assess this kind of learning, relative to the substantial number of measures designed for use within a single disciplinary context. In this paper we have described the creation of a unique pair of assessments built around the law of conservation of energy that can be used together to generate a cross-disciplinary learning index. The EResPhys measures original learning of LCE in a physics context, while the EAppChem measures an individual's ability to apply this knowledge in chemistry.

Results from administration of the assessments with several different populations provide preliminary evidence that the CDLI is a valid metric of cross-disciplinary learning. Cross-disciplinary learning is a more abstract construct than what is targeted by a typical concept inventory, and fully establishing the validity of the metric will require additional data collection. However, based on the results of our initial analyses we believe both the specific instrument and the general approach show promise. We are sharing our work at this time in order to make the instrument known and available to a broader audience and to help stimulate further research and discussion on what cross-disciplinary learning is and how best to measure it.

A. Potential applications

Although the CDLI could be used within the context of a single course, we see the most natural application being at the program level. The pedagogical methods that maximize mastery within the context of a specific course are not necessarily the same as those that help students broadly apply what they learn. A distinction is often drawn between learning disconnected facts and learning a coherent conceptual framework [36]. The latter is more effective at promoting generalization, but is slower than superficial learning of facts. Consequently, if the assessment tools we use measure narrow, context-specific mastery, pedagogical methods that emphasize the ability to generalize may appear to produce poorer results. This is likely to discourage instructors from persisting with such methods. Thus, if we wish for students to be able to integrate ideas across courses and disciplines, it is important that we have assessment tools focusing on that outcome. The CDLI is one such tool.

In carrying out program assessment, there are two methodological approaches that can be used. Programs sometimes state learning objectives involving a specific set of content or skills. For example, the chemistry program at our institution identifies atomic theory and thermodynamics as two such content areas. One could then evaluate success by developing an assessment testing mastery of these topics. This would be a criterion-referenced assessment. In physics, the Force Concept Inventory is an instrument that adopts this approach [20].

However, for many program goals, it may be very difficult to determine an appropriate criterion in this way. For example, six-year-graduation rate is a metric often used in program assessment, but there is no particular value of this metric that can be said to indicate success. Thus, programs typically compare their rate with other programs, or may compare the rates within their program for different subgroups of students. Similarly, our chemistry program also has an objective for students to be able to "Effectively communicate scientific information in written and oral forms." It is again not clear what absolute level of performance should count as success with respect to this objective. However, the program could evaluate success by examining the improvement in student communication skills between introductory courses and capstone courses. These examples illustrate a norm-referenced approach.

The CDLI also uses a norm-referenced approach. Accordingly, it cannot be used to determine whether or

not students have "achieved" a learning objective of demonstrating cross-disciplinary learning. However, it may be quite useful for comparing the performance of different groups of students, or tracking student performance over time. In the PFL framework discussed in the Introduction [11], one of the core ideas is that students can experience learning that is not immediately evident on a traditional test, but becomes evident in quicker mastery of a related topic down the road. Focusing on student learning within the context of a specific course would obscure this kind of learning, and perhaps lead to the conclusion that instruction had been unsuccessful. Thus, it is important to be able to track the development of student understanding over a longer time span than a single course. Because the questions on our assessments do not assume specific technical knowledge or utilize discipline-specific jargon, the assessments are well suited to such longitudinal assessment. The fact that the assessments are appropriate for a broad range of students also makes them useful for cross-sectional comparisons across different types of courses. For example, since they require minimal mathematical reasoning, they could be used to compare a calculus-based physics course to an algebra-based physics course to a conceptual physics course.

B. Limitations

Although we believe the CDLI can be a valuable tool for instructors and programs, there are several limitations to the existing dataset and the assessments themselves that are important to keep in mind.

First, interpretation of CDLI scores is heavily context dependent. For example, if students have already taken a college-level chemistry course that touches on the application of LCE to chemistry, they will most likely respond to the EAppChem items by drawing on their learning from that course, rather than a previous physics course. This issue is not specific to our instrument, but it is more important to consider when examining learning across courses, where the context is arguably a student's overall educational history. We believe this context dependence is not a bug but a feature, in that it is both desirable and useful to consider student learning in a larger context than individual courses. At the same time, it is important to not blindly compare CDLI scores collected in different contexts.

Second, the assessments are not intended to be a comprehensive inventory of student understanding of energy ideas. They are meant specifically to probe students' ability to apply the energy conservation principle. Both assessments do correlate with total number of science courses taken, suggesting that they are sensitive to a student's increasing level of scientific sophistication as they complete more coursework. However, these assessments on their own are not going to be appropriate as an overall measure of how student thinking about energy concepts develops across courses.

Third, both assessments use a multiple-choice format and yield a quantitative score. The multiple-choice format makes the assessments convenient to administer but does not provide deep insights into student thinking. Consequently, the assessments are good for answering "how much" questions, but less useful for answering "how" or "why" questions. This is especially true given that the distracters were not chosen to target particular lines of student reasoning. In our previous paper describing our overall measurement approach [2], we discussed the value of think-aloud interviews to characterize students' thinking during cross-disciplinary learning. We plan to further describe results from an interview component of our project in an upcoming paper.

Fourth, our initial analyses using the CDLI have been based on a cross-sectional dataset. Thus, the relationships between the CDLI and other factors may not be causal in nature. For example, students who have taken fewer versus more science courses may differ in ways other than amount of science coursework. Students taking more science may have a greater interest in science, or may have more confidence in their ability to perform well in science. Consequently, students may self-select into different groups in a way that concentrates students "good" at generalization into certain groups, rather than students learning to generalize through instruction. To test for this possibility, a longitudinal study must be conducted. We have a modest amount of longitudinal data in our dataset, but not enough for a meaningful analysis.

Finally, we only have data from two institutions, and most of those data come from a single university. Collecting data from a range of institutions will be necessary to establish that the assessments truly are usable with a broad range of students, and that the patterns we have identified are general rather than being idiosyncratic to our specific population.

C. Future directions

Moving forward, we see several next steps for this line of work. First, as mentioned in the previous section, we currently have a rather homogeneous sample. By partnering with other institutions, we can augment our sample with students representing different types of institutions, different educational histories, and different relationships with science. To facilitate these partnerships, we have begun constructing a Web-based portal that will allow instructors to distribute the assessment pair to their students and review the resulting data. We intend this to be an easy-to-use tool that instructors or assessment coordinators can deploy to get some quick feedback on particular interventions, and which will serve as one part of a larger portfolio of assessment measures for a program.

Second, and also mentioned above, although there are suggestive relationships in our data, without following the same students over time we cannot definitively establish a causal relationship between particular coursework and their

ability to integrate concepts across disciplines. To partially address this limitation, we have been collecting data across the sequence of teacher education courses mentioned in the Introduction. At this point we do have some data from multiple courses for some students, but the numbers are too small to permit a meaningful analysis. Further longitudinal data collection will help address this deficit. We hope that as other institutions start using these assessments they will collect data across multiple courses. In particular, it would be useful to follow students through a more traditional science course sequence, so the effects of that kind of instruction can be compared with the effects of the pedagogy used in the teacher education courses.

Third, with larger samples, it should become possible to more fully explore which students are demonstrating a strong ability to integrate ideas across disciplines and which students find this more challenging. Identifying the various factors that are related to integration ability will help researchers develop more sophisticated theoretical models. In turn, more sophisticated models can help instructors design more effective pedagogical interventions and target those interventions more effectively.

VII. CONCLUSION

The primary goal of this project was to create an easy-to-use instrument for measuring the construct of cross-disciplinary learning. In this paper we have described the development and initial validation of two assessments, EResPhys and EAppChem, as well as how the scores from these two assessments can be used to calculate an index—the CDLI—that appears to be largely independent of content knowledge, but is sensitive to a student's ability to integrate ideas across disciplinary boundaries. As our assessments are fairly narrow in scope, this instrument would not be appropriate as a stand-alone approach to assessing student learning, whether within or across courses. However, we believe it can be a useful complement to other types of assessment, especially at the program level, because it taps into an aspect of student learning that few other quantitative measures seek to capture. Furthermore, we believe our overall approach may provide a useful model for other researchers and educators who seek ways to go beyond measuring what a student knows to explore how that student is able to generalize and apply that knowledge in novel situations. As we have noted previously, placing the focus on cross-disciplinary learning requires thinking of ourselves as STEM educators first and disciplinary experts second [2]. Although we certainly hope that others find our assessments useful, the surest indicator that our project has been successful is if it stimulates productive conversations about what aspects of student learning are most important to measure and spurs innovative approaches to doing that measuring.

ACKNOWLEDGMENTS

We are grateful to co-principal investigator Sarah Julin for helping develop interview protocols and methods, and to advisory board members Dr. Jack Barbera, Dr. Amy Robertson, and Dr. Ben Geller for feedback, advice, and guidance in conceptualizing cross-disciplinary learning and its measurement. We would also like to thank Dr. Deborah Donovan and Dr. Susan DeBari, who gave feedback on initial drafts of instruments and protocols, and Dr. Daniel Hanley, who provided formative evaluation in early stages of this project. This work was supported by the National Science Foundation Grants No. DUE-1612055 and No. DUE-1612251. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

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

[2] E. Borda, T. Haskell, and A. Boudreaux, Cross-disciplinary learning: A framework for assessing application of concepts across STEM disciplines, *J. Coll. Sci. Teach.* (to be published).

[3] A. Eisenkraft, J. Nordine, B. Chen, D. Fortus, J. Krajcik, K. Neumann, and A. Scheff, Introduction: Why focus on energy instruction?, in *Teaching and Learning of Energy in K–12 Education*, edited by R. Chen, A. Eisenkraft, D. Fortus, J. Krajcik, K. Neumann, J. Nordine, and A. Scheff (Springer, Dordrecht, 2014), pp. 1–11.

[4] NGSS Lead States, *Next Generation Science Standards: For States, By States* (The National Academies Press, Washington, DC, 2013).

[5] F. Burrack and T. McKenzie, Enhanced student learning through cross-disciplinary projects, Music Educ. J. **91**, 45 (2005).

[6] E. P. Galloway, W. Qin, P. Uccelli, and C. D. Barr, The role of cross-disciplinary academic language skills in disciplinary, source-based writing: Investigating the role of core academic language skills in science summarization for middle grade writers, Read. Writ. **33**, 13 (2020).

[7] S. M. Barnett and S. J. Ceci, When and where do we apply what we learn? A taxonomy for far transfer, Psychol. Bull. **128**, 612 (2002).

[8] Y. J. Dori and I. Sasson, A three-attribute transfer skills framework—Part I: Establishing the model and its relation to chemical education, Chem. Educ. Res. Pract. **14**, 363 (2013).
[9] I. Sasson and Y. J. Dori, A three-attribute transfer skills framework—Part II: Applying and assessing the model in science education, Chem. Educ. Res. Pract. **16**, 154 (2015).
[10] M. L. Gick and K. J. Holyoak, Schema induction and analogical transfer, Cogn. Psychol. **15**, 1 (1983).
[11] J. D. Bransford and D. L. Schwartz, Rethinking transfer: A simple proposal with multiple implications, Rev. Res. Educ. **24**, 61 (1999).
[12] A. diSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).
[13] A. diSessa and B. Sherin, What changes in conceptual change?, Int. J. Sci. Educ. **20**, 1155 (1998).
[14] D. Hammer, Student resources for learning introductory physics, Phys. Educ. Res. Am. J. Phys. Suppl. **68**, S52 (2000).
[15] R. Scherr, Modeling student thinking: An example from special relativity, Am. J. Phys. **75**, 272 (2007).
[16] M. L. Gick and K. J. Holyoak, Analogical problem solving, Cogn. Psychol. **12**, 306 (1980).
[17] H. C. Sabo, L. M. Goodhew, and A. D. Robertson, University student conceptual resources for understanding energy, Phys. Rev. Phys. Educ. Res. **12**, 010126 (2016).
[18] D. Hammer, The variability of student reasoning, in *Proceedings of the Enrico Fermi summer school, course CLVI*, edited by E. Redish and M. Vicentini (Italian Physical Society, Bologna, 2004), pp. 279–340.
[19] J. Shen, S. Sung, and D. Zhang, Toward an analytic framework of interdisciplinary reasoning and communication (IRC) processes in science, Int. J. Sci. Educ. **37**, 2809 (2015).
[20] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[21] D. R. Mulford and W. R. Robinson, An inventory for alternate conceptions among first-semester general chemistry students, J. Chem. Educ. **79**, 739 (2002).
[22] M. Park and X. Liu, Assessing understanding of the energy concept in different science disciplines, Sci. Educ. **100**, 483 (2016).
[23] L. Ding, R. Chabay, and B. Sherwood, How do students in an innovative principle-based mechanics course understand energy concepts?, J. Res. Sci. Teach. **50**, 722 (2013).
[24] J. Shen, O. L. Liu, and S. Sung, Designing interdisciplinary assessments in sciences for college students: An example on osmosis, Int. J. Sci. Educ. **36**, 1773 (2014).
[25] C. F. Herrmann-Abell and G. E. DeBoer, Developing and using distractor-driven multiple-choice assessments aligned to ideas about energy forms, transformation, transfer, and conservation, in *Teaching and Learning of Energy in K-12 Education*, edited by R. F. Chen, A. Eisenkraft, D. Fortus, J. Krajcik, K. Neumann, J. C. Nordine, and A. Scheff (Springer International Publishing, New York, NY, 2014), pp. 103–133.
[26] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, Int. J. Sci. Educ. **33**, 1289 (2011).
[27] A. Barlow, S. Brown, B. Lutz, N. Pitterson, N. Hunsu, and O. Adesope, Development of the student course cognitive engagement instrument (SCCEI) for college engineering courses, Int. J. STEM Educ. **7**, 22 (2020).
[28] A. Madsen, S. B. McKagan, and E. C. Sayre, Research-based assessment instruments in physics and astronomy, Am. J. Phys. **85**, 245 (2017).
[29] K. Oberauer and S. Lewandowsky, Simple measurement models for complex working-memory tasks, Psychol. Rev. **126**, 880 (2019).
[30] L. Ding, Designing an energy assessment to evaluate student understanding of energy topics, Ph.D. thesis, North Carolina State University, 2007.
[31] G. J. Privitera, *Research Methods for the Behavioral Sciences* (Sage Publications, Los Angeles, 2014).
[32] K. S. Taber, The use of Cronbach's alpha when developing and reporting research instruments in science education, Res. Sci. Educ. **48**, 1273 (2018).
[33] *Introduction to Measurement Theory*, edited by M. J. Allen and W. M. Yen (Waveland Press, Long Grove, IL, 2002).
[34] K. D. Mattern, B. F. Patterson, and J. L. Kobrin, *The Validity of SAT Scores in Predicting First-Year Mathematics and English Grades* (The College Board, New York, NY, 2012).
[35] N. Prince, M. Vigeant, and K. Nottis, Development of the Heat and Energy Concept Inventory: Preliminary results on the prevalence and persistence of engineering students' misconceptions, J. Eng. Educ. **101**, 412 (2012).
[36] *How People Learn: Bridging Research and Practice*, edited by M. S. Donovan, J. D. Bransford, and J. W. Pellegrino (National Academies Press, Washington, DC, 1999).