9-1-2020

From Theory to Practice: Gathering Evidence for the Validity of Data Collected with the Interdisciplinary Science Rubric (IDSR).

Brie Tripp
Portland State University

Erin E. Shortlidge
Portland State University, eshort@pdx.edu

Citation Details

Tripp, B., & Shortlidge, E. E. (2020). From Theory to Practice: Gathering Evidence for the Validity of Data Collected with the Interdisciplinary Science Rubric (IDSR). *CBE—Life Sciences Education*, *19*(3), ar33. https://doi.org/10.1187/cbe.20-02-0035

From Theory to Practice: Gathering Evidence for the Validity of Data Collected with the Interdisciplinary Science Rubric (IDSR)

Brie Tripp and Erin E. Shortlidge*
Portland State University, Department of Biology, Portland, OR 97201

ABSTRACT

In a world of burgeoning societal issues, future scientists must be equipped to work interdisciplinarily to address real-world problems. To train undergraduate students toward this end, practitioners must also have quality assessment tools to measure students' ability to think within an interdisciplinary system. There is, however, a dearth of instruments that accurately measure this competency. Using a theoretically and empirically based model, we developed an instrument, the Interdisciplinary Science Rubric (IDSR), to measure undergraduate students' interdisciplinary science thinking. An essay assignment was administered to 102 students across five courses at three different institutions. Students' work was scored with the newly developed rubric. Evidence of construct validity was established through novice and expert response processes via semistructured, think-aloud interviews with 29 students and four instructors to ensure the constructs and criteria within the instrument were operating as intended. Interrater reliability of essay scores was collected with the instructors of record ($\kappa = 0.67$). An expert panel of discipline-based education researchers ($n = 11$) were consulted to further refine the scoring metric of the rubric. Results indicate that the IDSR produces valid data to measure undergraduate students' ability to think interdisciplinarily in science.

INTRODUCTION

The acceleration in scientific advancement over the past few decades has been aided by scientists working collaboratively across a wide range of disciplines. Multilayer science initiatives have been launched to further support innovative workspaces that develop and promote interdisciplinary (ID) programs and collaborative research. The National Science Foundation (NSF), National Institutes of Health (NIH), and National Institute of General Medical Sciences have led various joint initiatives and specific grant proposal solicitations that encourage or require ID science collaborations in fields such as behavioral biomedicine, computational neuroscience, and mathematical biology (National Research Council [NRC], 2003). The National Institute on Drug Abuse has funded research that leverages cognitive science, neurobiology, and sociology to evaluate drug addiction and its impacts on society (NIH, n.d.).

Given the complexity of environmental, social, and public health problems, this surge of interest in ID collaborations is not only timely but also necessary for ameliorating these issues. In line with this, several funding agencies and stakeholders have called for ID science exposure at the undergraduate level (NRC, 2003, 2009; American Association for the Advancement of Science [AAAS], 2011), such that we can train future scientists to think and work interdisciplinarily to tackle these real-world challenges. One prominent recommendation for ID science in undergraduate education is in *Vision and Change in Undergraduate Biology Education: A Call for Action* (AAAS, 2011). This report identifies that undergraduate biology students should understand the ID nature of science, the role of science in society, and the ability to communicate

Ross Nehm, *Monitoring Editor*
Submitted Feb 27, 2020; Revised Jun 2, 2020; Accepted Jun 10, 2020
CBE Life Sci Educ September 1, 2020 19:ar33
DOI:10.1187/cbe.20-02-0035
*Address correspondence to: Erin E. Shortlidge (eshortlidge@pdx.edu).

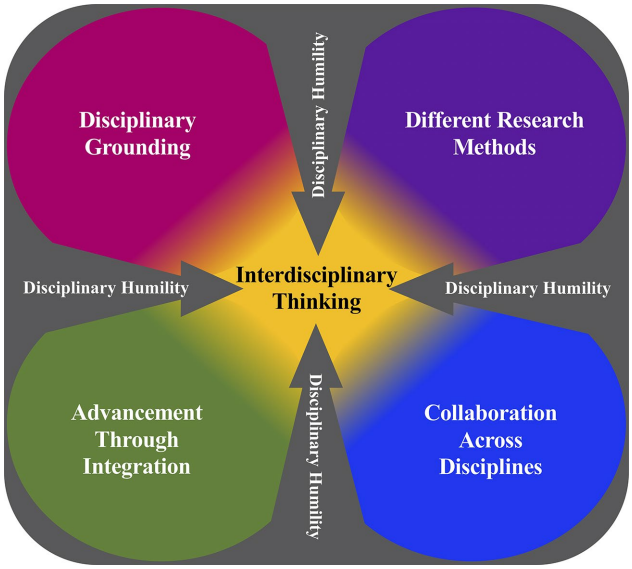

FIGURE 1. The Interdisciplinary Science Framework (IDSF). (Redrawn from Tripp and Shortlidge, 2019.)

and collaborate with other disciplines. Presumably, if undergraduate science students become adept at thinking and working interdisciplinarily, as these skills outlined in *Vision and Change* require, they will be better positioned to solve complex problems (NRC, 2003, 2009; AAAS, 2011).

Although the aforementioned initiatives play a key role in catalyzing ID science reform goals, they lack guidelines on how to create ID curricula and do not provide mechanisms by which to assess whether students are meeting these objectives. Thus, questions arise from educators such as: What does ID science mean? How do I foster an ID science environment in my classroom? Are students truly gaining ID science thinking skills, and if so, what tools are available to measure this competency? Our previous work made progress toward answering these questions through the development of a theoretical model, the Interdisciplinary Science Framework (IDSF; Figure 1). The IDSF is intended to guide instructors in what ID science looks like at the undergraduate level and, ideally, to provide a platform to create ID curricula and assessments (Tripp and Shortlidge, 2019). The IDSF was developed through expert feedback from 184 science faculty and literature related to ID science understanding (Klein, 1990, 1996, 2000, 2005, 2015; Newell, 1990; Boix Mansilla and Duraisingh, 2007; Borrego and Newswander, 2010; Öberg, 2009, Byrne *et al.*, 2016). This model then underwent testing for evidence of convergent validity in Tripp *et al.* (2020), establishing the IDSF as a valid framework by which to design curricula and inform instrument development.

To begin the process of assessing ID science comprehension, we first surveyed science faculty nationwide ($n = 68$) and asked how they assess this competency (Tripp *et al.*, 2020). We identified that writing assignments were the most common way that instructors assessed whether their students were meeting ID science learning goals. Based on these results, we tested one published rubric designed to measure students' ability to understand interdisciplinarity in the social sciences through a writing assignment (Boix Mansilla *et al.,* 2009). This rubric contained two constructs—disciplinary grounding and integration—that loosely aligned with the IDSF. These components of the rubric alongside its capability to assess students' writing provided justification for examining its functionality on science student populations. Therefore, we tested this tool's ability to produce valid data when implemented in natural and physical science courses. Results revealed that parts of the rubric were not operating on our population as the original designers intended, but rather, students conceptualized ID science more similarly to the constructs outlined in the IDSF (Tripp *et al.*, 2020). This called for the development of a new assessment tool guided by the evidence-based IDSF. Herein, we extend this work by gathering evidence from several sources of validity and reliability to develop an instrument based on the IDSF, the Interdisciplinary Science Rubric (IDSR), to measure students' ID science thinking in the context of real-world issues.

Learning Interdisciplinarity through Real-World Contexts

As science is conducted in a societal context, one purpose for "doing" science is to add knowledge to scientific (and nonscientific) domains that can be applied to real-world issues. When students can grasp and intertwine different pieces of knowledge from diverse disciplines to develop solutions to unresolved issues in society, they will likely be more prepared to enter the scientific workforce. A specific action item in *Vision and Change* states that instructors ought to "relate abstract concepts in biology to real-world examples on a regular basis, and make biology content relevant by presenting problems in a real-life context" (AAAS, 2011, p. 18). As new areas of research expand, future scientists will undoubtedly "need to contribute their expertise to research questions as they collaborate with people from other disciplines to address complex and increasingly interdisciplinary problems" (AAAS, 2011, p. 3). This call can be actualized by instructors implementing relevant activities in the classroom for students to apply their ID science knowledge and skills to unresolved issues.

There are several lines of evidence that support student writing as a means to generate greater literacy on current real-world problems (Connolly and Vilardi, 1989; Rivard, 1994; Keys, 1999; Balgopal and Wallace, 2009; Reynolds *et al.*, 2012; Balgopal *et al.*, 2012, 2017). A pedagogy known as "writing-to-learn" was adapted to move students from fact-based memorization and simplified connections to metacognitive skill development and scientific understanding through real-world applications (Connolly and Vilardi, 1989; Rivard, 1994). Writing activities guide students to reflect on what they know, what information they will need to gather, and their understanding of themselves in relation to the task and the strategies available to accomplish the task. Metacognitive theorists refer to these kinds of learning situations as "problem-solving" situations (Flavell, 1979). The ability to problem solve and make these connections is increasingly important when students are tasked with not only grasping deep disciplinary understanding but also integrating knowledge from multiple disciplines into a cohesive whole. Through this lens, writing in science can provide an outlet for students to compare and contrast methods, concepts, and ideas across disciplines toward novel solutions (Boix Mansilla *et al.,* 2009; Balgopal *et al.*, 2017).

Given our previous work revealing that writing assessments were the most prevalent way that science instructors assess this

competency in undergraduate classrooms (Tripp *et al.*, 2020), and the wealth of literature supporting writing in science, we decided to develop a quality instrument that assesses students' written work in relation to real-world issues.

The Role of Validity and Reliability in Instrument Development

Attention to assessment in higher education has increased since the 1980s, with a surge in research studies aimed at designing assessment tools to evaluate student learning gains, inform instruction practices, and improve curricula (Boix Mansilla and Duraisingh, 2007). With the growing array of assessment tools being designed by researchers, it is incumbent on instrument developers to ensure that the quality of these tools meet appropriate validity standards. The *Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards*) describes validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests," and is therefore "the most fundamental consideration in developing and evaluating tests" (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, and Psychological Testing [AERA], 2014, p. 11). In other words, validity is the centerpiece that assists developers in determining whether the instrument is measuring what it is intended to measure.

Research articles often refer to validity as the "validity of an instrument," which is inaccurate nomenclature according to psychometricians (Barbera and VandenPlas, 2011; Wren and Barbera, 2013; Arjoon *et al.*, 2013; AERA, 2014; Knekta *et al.*, 2019). An instrument cannot be valid or invalid, but rather it is the interpretation of the data collected from the tool that can be validated. These important interpretations are often in reference to specific concepts or theory-derived constructs that the instrument is designed to measure. The *Standards* use a contemporary view of construct validity as an overarching validity trait that all other validities could be used to establish (AERA, 2014). This conceptualization of construct validity has five main sources of evidence: test content, response process, internal structure, association with other variables, and consequence of use (Figure 2). A combination of these quantitative and qualitative sources is used in instrument development to increase the validity of the interpretations of data collected. For the purposes of this study, we will be describing and analyzing response process validity (please see AERA (2014) for a detailed description of other sources of validity).

Evidence of Validity Based on Novice and Expert Response Processes

It can be informative for researchers to gain evidence based on the cognitive processes students use to answer criteria or items within an instrument, as observed scores are inseparably linked to how students respond (Wren and Barbera, 2013). This process can assist instrument developers in ensuring that students are interpreting the criteria as intended and can provide a deeper lens into the cognitive processes that formulate a student's answer. This is formally known as *novice* response process validity (Arjoon *et al.,* 2013; AERA, 2014). Additional evidence can be collected from subject-matter experts on the appropriateness of the scoring scale, the accuracy of criteria within the constructs, and the extent to which the scorers are able to interpret the scores as intended (Wren and Barbera, 2013; AERA, 2014). These professional insights are referred to as *expert* response processes.

Evidence of Reliability

Reliability measurements are often obtained alongside validity to buttress the quality of an instrument. Reliability of a measure refers to consistency in the instruments' items and the extent to which it is free from error (Stangor, 2014; Arjoon *et al.*, 2013). The *Standards* describe two sources of reliability: temporal stability and internal consistency (Figure 2). These sources are discussed in the *Standards* strictly based on quantitative self-report scales (i.e., surveys). It is common practice, however, to apply the internal consistency approach to more qualitative sources of reliability, such as interrater reliability (IRR; Stangor, 2014). IRR measures the extent to which multiple judges' ratings on criteria correlate with each other, thus demonstrating that the judges are all measuring the true scores (or the same variables) rather than random error. Stangor (2014) provides a useful justification for taking this approach on work that is more interpretive (such as students' writing): "Just as any single item on [an instrument] is expected to have error, so the ratings of any one judge are more likely to contain error than the average rating across a group of

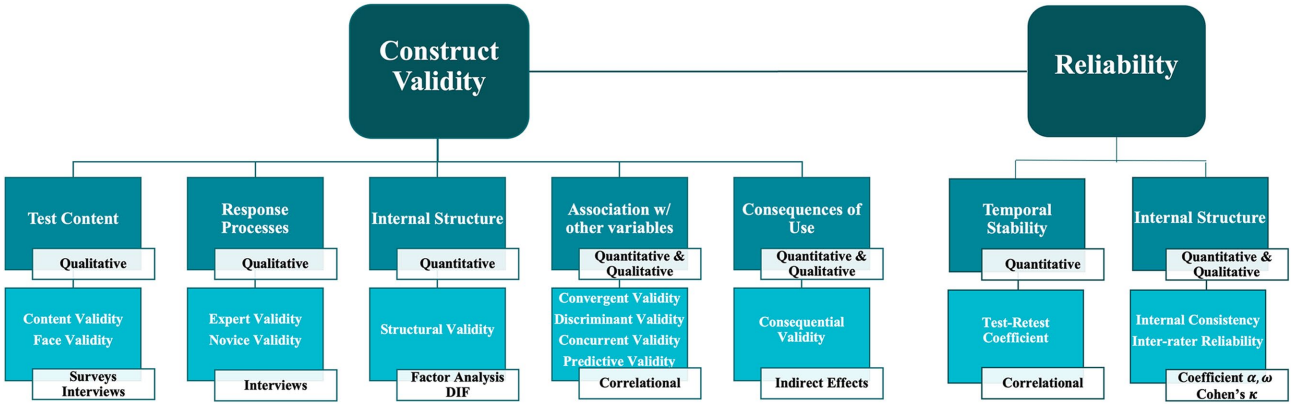

FIGURE 2. Schematic representation of the multiple sources of evidence for validity and reliability; DIF: differential item functioning.

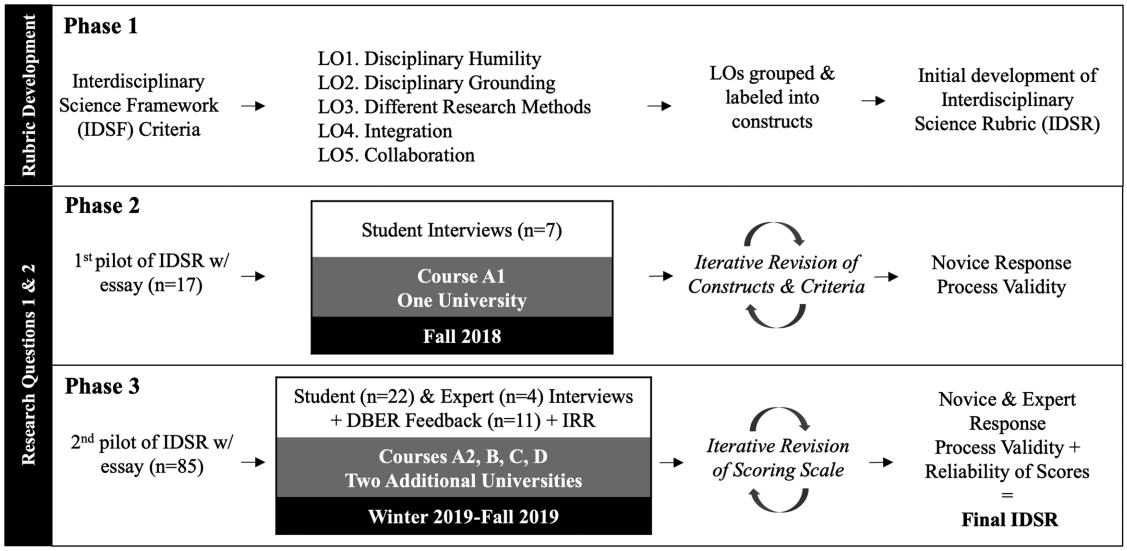

FIGURE 3. A three-phase outline for the development (phase 1) and testing (phases 2 and 3) of the Interdisciplinary Science Rubric (IDSR). LO, learning outcomes; DBER: discipline-based education researchers; IRR, interrater reliability.

raters" (p. 95). Hence, obtaining this form of reliability through the scoring of students' work (i.e., writing assignments) strengthens the precision of an instrument.

OVERVIEW OF THE STUDY

This research study focuses on testing for evidence of construct validity from novice and expert response processes to develop and iteratively revise a new instrument, the Interdisciplinary Science Rubric (IDSR). We additionally provide evidence of reliability through internal consistency (via IRR) of researchers and instructors scoring students' written work based on the constructs in the instrument. The intention of the IDSR is to accurately assess students' ID science thinking related to real-world problems. Herein, we address the following research question through pilot testing of the IDSR and semistructured student and faculty interviews: What evidence supports the constructs and criteria in the interdisciplinary science rubric as a quality assessment tool?

We have divided the evolution and development of the IDSR into three sequential phases: phase 1, rubric development; phase 2, first pilot of rubric; and phase 3, second pilot of rubric (Figure 3). We report the methods and results to each phase sequentially.

METHODS: PHASE 1

Rubric Development

In our previous work, we tested whether a pre-existing social science rubric (Boix Mansilla *et al.,* 2009), developed to score student essays related to ID understanding in the humanities, could effectively measure the ability of natural and physical science students to communicate ID science understanding (Tripp *et al.*, 2020). We established evidence of convergent validity for one of the original constructs, *disciplinary grounding*, while the remaining constructs, *integration* and *critical awareness*, failed validity tests (Tripp *et al.*, 2020). Therefore, we worked to revise the original rubric into a new, evidence-based rubric for natural science students, the IDSR. The five core criteria of the IDSF (Tripp and Shortlidge, 2019), *disciplinary humility*, *disciplinary grounding*, *different research methods*, *integration*, and *collaboration*, became the blueprint for the IDSR. We subsequently tested the IDSF for evidence of convergent validity and established that the constructs accurately represented students' and experts' understanding of this competency (Tripp *et al.*, 2020).

We would like to differentiate the use of the term "ID understanding" from that of "ID thinking": "ID understanding" was the phase used in the pre-existing rubric (Boix Mansilla *et al.,* 2009) that we tested (Tripp *et al.*, 2020) and was also used in the development of the IDSF (Tripp and Shortlidge, 2019). But in terms of assessment, "understanding" is a nebulous term that is extremely difficult, if not impossible, to measure. We also posit that several constructs (i.e., objective and broader awareness) in the IDSR do not necessarily measure a student's ID understanding, but rather a student's ability to *think* in an interdisciplinary way when considering how to address real-world problems. Thus, the measure of "students' thinking" is more accurate in describing the purpose of the IDSR.

In developing the initial dimensions of the IDSR, we followed Stevens and Levi's (2013) four basic stages for constructing rubrics: reflecting, listing learning objectives (LOs), grouping and labeling, and application (Figure 3). In the reflection stage, we took time to reflect on what we wanted from our students and what happened when we previously administered the original rubric published by Boix Mansilla *et al.* (2009). Using this knowledge, we stepped into stage 2: listing LOs for the assignment. We developed LOs to closely align with intended outcomes and criteria in the IDSF and labeled these as objectives for our rubric. In stage 3, we grouped and labeled similar objectives together (e.g., "different research methods" [LO 3] can be categorized under the larger construct "disciplinary grounding" [LO 2]) and identified the subcategories, or criteria, that would define each construct. Finally, stage 4 involved the application of the constructs and associated criteria from stage 3 into a grid format.

Scale

To measure how students address dimensions of a rubric, Stevens and Levi (2013) suggest confining rubrics to three levels of performance in the initial stage of construction, as it becomes more difficult to differentiate between student understanding on scales greater than five. Based on our previous findings (Tripp *et al.*, 2020), in which raters often found it challenging to clearly differentiate between the four discrete levels of understanding from the social science rubric (naïve, novice, apprentice, and mastery), four researchers (including B.T. and E.E.S.) created three levels to evaluate students' ID science thinking ability in the IDSR. We selected similar terms that were positive, active verb descriptions: novice, intermediate, and mastery. We iteratively revised these levels of understanding to provide succinct and direct instructions that practitioners could follow when scoring each construct. We then outlined specific criteria that would constitute students' ability to think interdisciplinarily in science for each construct domain, guided by the IDSF. Some constructs had more criteria than other constructs but were not meant to be interpreted as more or less important. Thus, we designed the rubric such that instructors average the scores from the criteria within each construct, resulting in one score per construct. Instructors then add the construct scores to provide the total earned point value for the assignment.

Writing Assignment

To pilot the IDSR, we collected samples of students' work that allowed them to exhibit ID knowledge. In our previous work, we developed course-specific essay assignments tasking students with integrating knowledge from multiple disciplines to effectively address a challenging real-world issue (Tripp *et al.*, 2020). Here, we used the same essay structure to create new, relevant prompts that we developed in conjunction with each course instructor represented in this study. Students were provided a "student version" of the rubric that outlined the expectations for the assignment. This version was identical to the full "practitioner version" provided in Table 2, minus the levels for scoring students (mastery, intermediate, novice; Table 2; see Supplemental Material A for example prompts and the shortened student version of the rubric). For a more detailed explanation and guidance on how to develop useful real-world prompts for this assignment, please see Tripp *et al.* (2020, p. 3).

RESULTS AND DISCUSSION: PHASE 1

Rubric Development

We infused elements of the IDSF into five learning outcomes for the writing assignment; students should: 1) display knowledge of one discipline while expressing provisional knowledge in other disciplines; 2) describe methods from multiple fields to accomplish said task; 3) integrate disciplines in a manner that results in a new discovery or idea; 4) display disciplinary humility through the inclusion and respect of team members and/or fields outside science, technology, engineering, and mathematics (STEM); and 5) describe how and why there is a need to create a collaborative team to accomplish the task. These LOs were transferred into constructs with several criteria aimed at measuring student thinking in ID science through the lens of the IDSF. This resulted in five constructs—objective, disciplinary grounding, integration, disciplinary humility, and collaboration. We created an objective construct to guide students with framing the issue and outlining an approach to tackle the problem in the essay prompt. Although this construct is not included in the IDSF, it is an essential component in essay development (Boix Mansilla *et al.,* 2009). Within this construct, we included criteria designed to assess students' ability to synthesize background information through credible sources and to evaluate the quality of students' approach to the essay prompt.

The next step in students thinking interdisciplinarily is the acquisition of deep disciplinary knowledge. The IDSF outlines disciplinary grounding and different research methods as exhibiting disciplinary knowledge through the inclusion of information, concepts, and research methods from each contributing discipline. Thus, we determined disciplinary grounding would be a construct, and embedded disciplinary reasoning and different research methods as criteria within the construct.

After students grasped the foundational pieces within disciplines, we tasked students with integrating their knowledge from different fields. In developing the integration construct, we paid close attention to the organization and wording of its criteria, as there is evidence that integration of knowledge across disciplines is the central factor separating interdisciplinarity from cross- and multidisciplinarity (Boix Mansilla *et al.,* 2009; Repko and Szostak, 2020; Borrego and Newswander 2010). Furthermore, this construct as defined in the social science rubric tested in our previous work did not accurately represent how students operationalized ID science integration (Tripp *et al.*, 2020). The IDSF alternatively defines integration as "not only collecting the appropriate disciplinary pieces of information and placing them in a central repository, but also proficiency in *integrating*—mixing, connecting, and applying them to discover new insights or ideas" (Tripp and Shortlidge, 2019, p. 6). Thus, we initially created criteria within this construct to assess students' ability to integrate different disciplines to further the project/solution in a way that one discipline could not.

Disciplinary humility was a criterion in the IDSF that was particularly challenging to initially develop. The IDSF describes this criterion as an affective measure that calls for respect, appreciation, and acknowledgment of other disciplinary perspectives and epistemologies, as well as the inclusion of science and non-science disciplines (Tripp and Shortlidge, 2019). We contemplated not only ways for students to understand this construct, but also how to measure this affective dimension in a writing sample. We started with a disciplinary humility criterion embedded within the integration construct that required students to explain why they must rely on disciplines/experts to address the problem.

Finally, a collaboration construct was developed to encourage the creation of common ground among team players and the inclusion of diverse disciplinary perspectives. This remained congruent with how the IDSF describes collaboration across disciplines.

Scoring and Scale

We scored students' ability to think interdisciplinarily on a three-point scale: 0, novice; 1, intermediate; and 2, mastery. We used a 0 value to represent novice thinking skills based on our previous findings, where 1 was overcrediting students who did not address any aspect of a particular criterion. We assigned

TABLE 1. Summary of five universities and associated upper-division course format (ID, interdisciplinary; D, disciplinary) and sample sizes of essays and interviews collected over the course of one academic calendar year.

University: Carnegie classification	Course department listing: Format	Essays (*n*)	Student interviews (*n*)	Instructor interviews (*n*)
A1: Public High Research Activity[a,b]	Biology: ID	17	7	N/A
A2: Public High Research Activity[a,b]	Biology: ID	15	7	1
B: Public High Research Activity[a]	Biology: D	23	5	1
C: Public Master's Colleges and Universities: Small Programs	Biology: ID	34	6	1
D: Public High Research Activity	Honors: ID	13	4	1
Total		102	29	4

[a]Denotes same university.
[b]Denotes same course taught at two separate time points.

values of 1 and 2 to intermediate and mastery, respectively, as we did not want there to be ambiguity *between* levels of thinking. For instance, if intermediate thinking had a value of 5 and mastery a value of 10, there would be a large range of numbers in between levels with no description of how to apply those values to students' essay responses (Table 2).

METHODS: PHASES 2 AND 3

Phase 2: First Pilot of Rubric

Recruitment and Data Collection. We piloted the first version of the IDSR and associated assignment to an upper-division ID science course, which was largely project based (course A1), at a public, research-intensive northwestern university in Fall 2018 (Table 1). The instructor of record (author E.E.S.) announced the assignment to the class. Students in this course had consented to any course work being used for research purposes, thereby allowing us to use their responses to the essay assignment for this study (PSU IRB no. 174450).

The instructor administered the essay assignment, which included the shortened student version of the rubric, to students at the end of the course. Students were given approximately 7 days to complete the individual assignment using any outside resources available to them. The assignment was attached to course points and factored into the overall course grade. Following completion of the assignment, we emailed students enrolled in the course for a follow-up interview and scheduled interviews with respondents. Before the interview commenced, students provided written consent for their interview responses to be used for research purposes. The students were provided $20 gift cards for participating in interviews. Student participation in this study remained anonymous from the instructor, and each student received a unique numerical identifier. Audio files were transcribed verbatim (Rev.com).

Student Think-Aloud Interviews. To establish evidence of novice response process validity, we conducted semistructured think-aloud interviews with students to better understand how they were interpreting each construct and associated criteria in the IDSR. Interview questions were designed and iteratively revised by three researchers (including B.T. and E.E.S.) to 1) investigate how wording of the constructs and criteria affected student responses, and whether students understood them as we had intended; 2) identify other ways in which students may understand ID science outside the rubric; and 3) gain insight on students' perceptions of the value of the assignment and rubric (see Supplemental Material B for student interview questions). We analyzed interview transcripts deductively (Patton, 1990) for evidence that students understood the criteria in the IDSR as we intended. We did not interview the instructor of this course for feedback on the rubric and assignment, as E.E.S. is an author of the paper and was fully involved in the development and revision of the IDSR.

Scoring and Interrater Reliability. Two researchers (including B.T.) scored essays to consensus with the IDSR. E.E.S. independently graded 100% of the essays with the rubric and compared scores with these researchers via IRR ($\kappa = 83$) calculated through R Studio (R Studio Team, 2020). We then collaboratively discussed and revised aspects of the rubric based on essay responses and student interviews to better reflect the intentions of each criteria and provide clarity to areas that were confusing and/or misleading students. English is not the first language of one of the researchers involved in this work; therefore, this researcher was able to provide a unique perspective on the syntax and word choice of the rubric. This is especially important regarding students whose first language is also not English. This provided a small, but important, element of inclusivity in the development of this instrument. This study was conducted under exempt status at Portland State University (IRB no. 163998).

Phase 3: Second Pilot of Rubric

Recruitment and Data Collection. After revising the rubric based on phase 2 results, we piloted the IDSR and essay assignments in phase 3 on four additional upper-division courses at varying institutions to examine the utility of the rubric on different populations: two courses from the same university (courses A2 and B) and two courses from two separate universities (courses C and D). Because courses A1 and A2 were the same course (taught in Fall 2018 and 2019), this allowed us to examine whether our revisions to the IDSR from phase 2 had an effect on a very similar student population in phase 3. Course B was an upper-division, non-ID, lecture-based science course from the same university (Winter 2019). Course C was an ID science, lecture-based course located at a public northwestern master's-granting university. Course D was an ID small group–based course located at an eastern research-intensive university (Table 1). A schematic outline of dissemination of the rubric across phases, institutions, and courses is visualized in Figure 3.

Most courses across the universities were listed under a biology department, except for course D—it was listed as an upper-division ID honors course that was not assigned to any specific discipline and was open to all upper-division majors (Table 1). However, the instructor was a biology faculty member

and science content was covered throughout the course, with the majority of enrolled students declaring science majors (77%). The administration of the IDSR on an array of disciplinary to ID formats from varying institution types allowed us to examine the functionality of the IDSR regardless of course content, format, student major, and population.

Instructors showed a recruitment video that requested students' consent for their responses to the essay assignment to be used for research purposes. The instructors then provided students with an online (Qualtrics) link to accept or decline for their essay responses to be used in this study. The survey also asked students if they would be interested in participating in a follow-up interview. Students who agreed to participate in an interview received another link requesting their consent for their interview responses to be used. We conducted interviews via an online service platform. The students were provided $20 gift cards for participating in interviews. We followed the same semistructured think-aloud student interview process and used the same interview questions as in phase 2.

Faculty Interviews. After the courses concluded, we obtained evidence of expert response process validity through semistructured interviews with the instructors of record to gather insight and feedback on the rubric and assignment. We emailed the instructors requesting their participation and obtained consent for their interview responses to be used through a survey in Qualtrics. Instructors were given a $50 gift card for their participation. Interviews were held in person or via Skype depending on the instructor's location. Interview questions were designed and iteratively revised by two researchers (B.T. and E.E.S.) to solicit information on the functionality of the instructor version of the rubric and to gain knowledge on the accuracy of each construct and criteria (see Supplemental Material C for instructor interview questions). Audio files from instructor interviews were transcribed verbatim by Rev.com.

Scoring and Interrater Reliability. To obtain evidence of reliability, B.T. emailed individual instructors 1 week before their interviews and had them score 20% of essays from their courses with the IDSR; B.T. independently scored the same 20% of essays. During each instructor's interview, the participant and B.T. obtained IRR on these documents. These analyses of reliability were conducted in R Studio (R Studio Team, 2020).

Discipline-Based Education Research Panel Feedback. To collect additional expert feedback on the IDSR, we consulted a group of discipline-based education researchers at a research-intensive university unrelated to the institutions in this study (n = 11). The researchers collectively read one essay assignment and then split into three groups to evaluate constructs and associated criteria in the IDSR, as well as the three levels of thinking in the "full practitioner" version. We then reconvened to discuss the researchers' perspectives and gain insight into the usability and clarity of the rubric. We subsequently modified these levels within the IDSR based on the feedback.

Statistical Analyses

To assess the utility of the revised rubric across different populations from phase 3, we conducted a one-way analysis of variance (ANOVA) to detect differences in mean student essay scores across courses A2–D. A Levene test for unequal variances and a Kruskal-Wallis nonparametric test for normality were run to confirm that ANOVA assumptions were met. We hypothesized that there would be no difference in overall essay scores between courses, as we iteratively revised the rubric to be broadly applicable to any discipline and real-world problem. ANOVA analyses were performed in R Studio (R Studio Team, 2020).

RESULTS AND DISCUSSION: PHASES 2 AND 3

We have combined the results and discussion from phases 2 and 3 to illuminate our process and decision-making for changes between the two versions of the IDSR.

Changes in Dimensions of IDSR

Examining interview transcripts from students (n = 25) in phase 2 resulted in the final rubric containing four constructs and associated criteria by which to measure students' ID science thinking related to real-world issues: objective, disciplinary grounding, integration, and broader awareness (Table 2). Thus, the rubric was modified from five constructs in phase 2 to four constructs in phase 3, which involved rearrangement and revisions to many dimensions of the rubric. We have labeled these constructs as "categories" in Table 2, so instructors who are unfamiliar with the word "construct" have a better understanding of this particular dimension of the rubric. A "format" category was included as an optional element for instructors to score basic requirements they deem necessary for complete written assignments (e.g., APA format, spelling, grammar). To reduce construct irrelevance variance (AERA, 2014)—the inclusion of extraneous and/or confounding variables that skew assessment outcomes—we did not include the scores from this format category when scoring essay assignments for research purposes.

Evidence of Novice and Expert Response Process Validity

We have provided student essay and interview responses and faculty interview responses for each criterion in the IDSR in Table 3. The data indicate a high level of consistency between novice and expert understanding of the rubric. Below, we outline additional evidence for the inclusion or exclusion of each criterion in the piloting of our instrument.

Objective Construct

There were several constructs that were relatively straightforward to both students and faculty in phases 2 and 3, including the objective construct (Table 3). Students expressed that this construct assisted them in collecting pertinent information and structuring their essays by providing a launch point to start the writing process:

> "The requirements for [objective] just sounds like a lead in, understanding what you're supposed to be doing, and then how you're going to apply that. It is design, think, build, or whatever the engineering workflow will be for the paper."
> —Student Interview, Course A1

Faculty similarly attested to the accuracy of the objective construct in helping students frame big issues:

TABLE 2. The instructor version of the IDSR to measure students' understanding of ID science[a]

Category	Criteria	Mastery (3)	Intermediate (2)	Novice (1)	Naïve (0)	Score**
OBJECTIVE	**1.1 Purpose:** What is the problem and task? Provide background information to introduce and frame the problem/ task.	Includes background information that *sufficiently* reviews the subject matter and *accurately* reports any historical and/or current material to frame the problem	Includes background information that *moderately* reviews the subject matter and *accurately* reports any historical and/or current material to frame the problem	Includes background information that *moderately* reviews the subject but *inaccurately* reports material OR includes *minimal* background information but *accurately* reports any historical and/or current material to frame the problem	Does *not* include background information	/3
	1.2 Approach: How will you approach the problem/task? Formulate a plan that *clearly* outlines your approach (steps/ procedures).	Includes a plan that *clearly* outlines steps/procedures that will be accomplished to address the problem/ at hand	Includes a plan that *moderately* outlines steps/procedures that will be accomplished but *excludes* steps/procedures that are discussed further in essay.	Includes a plan that *unclearly* and/or *insufficiently* addresses the problem/task at hand	Does *not* include a plan	/3
	1.3 Credibility: What sources will you include? Use peer-reviewed articles and/or other supporting information that are relevant to the problem/task.	Includes peer-reviewed articles and/or information that are *relevant* to the problem	Includes peer-reviewed articles and/or information that are *tangential* to the problem	Includes peer-reviewed articles and/or information but are *irrelevant* to the problem	Does *not* include peer-reviewed articles and/or information	/3
					Average *Objective* Score	/3
DISCIPLINARY GROUNDING	**2.1 Disciplines/Experts:** What disciplines and/or experts will be involved? Include *two or more disciplines and/or experts* in your approach to the problem/task.	Includes *two or more* disciplines and/or experts *relevant* to student's approach	Includes *two or more* disciplines and/or experts but some are *irrelevant* to student's approach.	Includes only *one* discipline and/ or expert *relevant* to student's approach.	Does *not* include disciplines and/or experts	/3
	2.2 Disciplinary Reasoning: Why are you including *each* discipline and/ or expert? *Meaningfully* explain the reasoning behind the use of each discipline and/or expert.	Includes *accurate, informed reasoning* behind all contributing disciplines and/ or experts	Includes *accurate, informed* reasoning behind one or multiple but *not all* contributing disciplines and/or experts	Includes *inaccurate* or *oversimplified* reasoning for all contributing disciplines and/or experts	Does *not* include reasoning behind using disciplines and/or experts	/3
	2.3 Methods & Tools: What methods will each discipline and/or expert use? Include techniques/ procedures/tools from contributing disciplines and/or experts.	Includes *accurate* methods (techniques/procedures/ tools) from *all* contributing disciplines	Includes *accurate* methods (techniques/procedures/ tools) from one or multiple but *not all* contributing disciplines and/or experts	Includes *inaccurate* methods (techniques/procedures/ tools) for all contributing disciplines and/or experts	Does not include methods (techniques/ procedures/ tools)	/3
					Average *Disciplinary Grounding* Score	/3

(*Continues*)

TABLE 2. Continued

Category	Criteria	Mastery (3)	Intermediate (2)	Novice (1)	Naïve (0)	Score**
INTEGRATION	**3.1 Leveraging Disciplines/Experts:** How will each contributing discipline and/or expert *build off* one another to effectively address the problem/task in a way that one contributor cannot? Specifically address how each discipline's and/or expert's contribution (*knowledge/methods*) will be useful for the other disciplines and/or experts.	Leverages contributing disciplines and/or experts by *building off* knowledge/methods to effectively address the problem/task in a way that one contributor cannot	Leverages contributing disciplines and/or experts by building off knowledge/methods to effectively address the problem/task in a way that one contributor cannot but *does not* consider all disciplines involved	Lists disciplines/experts knowledge/methods contribution *without* building off the knowledge/methods from each contributor	Does *not* include ways to leverage the disciplines and/or experts	/3
	3.2 Collaboration: How will you foster successful partnerships? Include and *explain* two or more ways to build community and respect among different disciplinary team members (e.g., establishing common ground and language, overcoming different perspectives, etc.).	Includes and *sufficiently* explains *two or more* logical ways to build community and respect among different disciplinary team members	Includes and *sufficiently* explains *one* logical way to build community and respect among different disciplinary team members	Lists *two or more* logical ways but *does not* explain how to build community and respect among different disciplinary team members	*Does not* include *nor* explain ways to build community and respect among different disciplinary team members	/3
					Average *Integration* Score	/3
BROADER AWARENESS	**4.1 Societal Impact:** How does your proposed solution impact society? Include *why* your solution is locally and more broadly relevant to society and what/who will be affected (e.g., economics, politics, social, health, etc.).	Includes local and broader societal impacts and *sufficiently* explains what/who will be affected	Includes local and broader societal impacts and *moderately* explains what/who will be affected OR Sufficiently explains what/who will be affected but includes *only* local *or* broader (not both) societal impacts	Includes *only* local *or* broader (not both) societal impacts and *does not* sufficiently explain what/who will be affected	Does *not* include local or broader societal impacts *nor* (what/who) will be affected	/3
	4.2 Limitations: What are the potential limitations to your plan and how will you overcome these barriers? Forecast possible limitations of your plan and provide resolutions.	Includes potential limitations of plan and *sufficiently* explains resolutions to overcome these barriers	Includes potential limitations of plan and *moderately* explains resolutions to overcome these barriers	Includes potential limitations but does *not* explains resolutions to overcome these barriers	Does *not* include limitations *nor* resolutions	/3
					Average *Broader Awareness* Score	/3
FORMAT*	4.1 Format, Grammar, Structure: Have you followed all formatting guidelines? Does your proposal have an introduction, body, and conclusion?					/3
					Total Score**	/15

[a]The "Category" and "Criteria" columns were provided to students in essay assignments (i.e., student version of rubric).

*Optional construct.

**Adjust each category's score to align with your course's point needs (i.e., if course assignment is worth 50 points, adjust category criteria accordingly). Average each category's criteria to obtain a category score. Keep each criterion equal in value to not weight one criteria and/or category over another.

TABLE 3. Examples of student essay and interview responses and faculty interview responses supporting the criteria in the IDSR

Construct	Criteria	Example of student essays from course C	Student interviews from all five courses	Instructor interviews
OBJECTIVE	**1.1. Purpose:** Provide background information to introduce and frame the problem.	"Louisiana wishes to build a park w/ a garden for human consumption, but the soil is filled w/ hydrocarbons & alkyl halides from a BP oil spill."	"General scope of research in the current field, what's going on—the problem. So, with that I focused on Department of Education information and statistics … how many people is this currently affecting." —Course A2	"What I see students answering for purpose is what the state of things are & how to address the problem."
	1.2. Approach: Formulate a plan that clearly outlines your approach.	"The best solution for reducing contaminates in the soil would be to use bioremediation methods with bacterial species that have the ability to use hydrocarbons as their source of energy by inoculating the soil on a mineral medium in the presence of sweet crude oil. Subsequent bacterial colonies would then be grown to produce more bacteria and reintroduced back into the contaminated site."	"How you go about solving the problem & the different necessary steps; explain what you're going to do about the problem."—Course D	"This will help students organize how to attack the issue and give step-wise direction."
	1.3. Credibility: Use peer-reviewed articles and other supporting information that are relevant to the problem/task.	"Our team will use data collected from this project to craft an application to the Environmental Protection Agency's brownfields, superfund, or emergency response cleanup programs, in order to offset the economic burden the Remediation Plan will place on the local communities (EPA, 2013)."	"I just tried to include as many articles from peer-reviewed journals as possible. I also looked for previous credible authors who had multiple articles under the same subject. I tried to avoid Wikipedia too." —Course B	"This is a given … always have students use credible sources."
DISCIPLINARY GROUNDING	**2.1. Disciplines/ experts:** Include two or more disciplines and/or experts in your approach to the problem/task.	"An important role in the recovery process is that of a public policy administrator, as well as a grant writer, a chemist, and a microbiologist."	"For me, when I was looking at disciplines and experts, how I interpret it is needing people from very specialized fields. The sheer complexity of the problem requires people with various specific skillsets coming together." —Course D	"2.1 means 'I'm going to use these disciplinarians and experts to do X, Y, and Z.'"
	2.2. Disciplinary reasoning: Meaningfully explain the reasoning behind the use of each discipline and/or expert.	"The policy administrator will ensure that the community is aware of all of the steps taken by scientific experts to restore the land while the grant writer familiar with the U.S. Environmental Protection Agency and its guidelines will be necessary to offset the costs of the cleanup and request more funds to sustain the newly developed garden. The microbiologist will take soil and water samples while the analytical chemist assesses the level of toxicity in these samples."	"An immunologist may not have the background or experience to address public opinion. Someone in public health may have a better tool set to do so."—Course B	"And then 2.2 builds off of 2.1 by having students go on to explain what those things are."
	2.3. Methods and tools: Include techniques/ procedures/tools from contributing disciplines and/or experts.	"Organic & inorganic chemists will extract, separate, & examine the pollutants w/ GC, NMR imaging, & Mass Spec."	"What is the direct action of what your experts will be doing and how they will accomplish that—what tools."—Course A2	"What techniques will be implemented from each discipline to accomplish the task."

(*Continues*)

TABLE 2. Continued

Construct	Criteria	Example of student essays from course C	Student interviews from all five courses	Instructor interviews
INTEGRATION	**3.1. Leveraging disciplines/ experts:** Address how each discipline's and/ or expert's contribution (knowledge/ methods) will be useful for the other disciplines and/or experts.	"It would vastly benefit the Remediation Plan to consult with local Department of Health agents and medical personnel in order to craft a Public Health bulletin to address community and health concerns about introducing a robust bacterial species to local properties. Assuming that we are not violating policies and can abate community concerns, the remediation strategy to remove the alkyl halides depends [on] chemists' toxicity analysis and on the microbiologists' soil samples.	"The idea that even when you break the problem up into chunks, it's not independent chunks, each chunk contributing to a whole, so they need to really work together. So, this gets at the idea that one part has to feed into another." —Course C	"This is providing logic behind how each puzzle piece fits into the whole & formulating a solution that is not possible without them."
	3.2. Collaboration: Include two or more ways to build community and respect among different disciplinary team members.	"Regular public meetings and private team building activities between participating parties should be facilitated through the entire process in order to maintain regular communication, active community involvement, accessible public education, and trust building."	"Going through typical municipal processes of community involvement, having team building exercises, being transparent." —Course D	"This is requiring [students] to think about collaborating in effective ways. More of a social skill to prepare them for real life problem-solving."
BROADER AWARENESS	**4.1. Societal impact:** Include what/ who will be affected (e.g., economics, politics, social, health).	The cleanup of this area and creation of a community garden would create a social, common space accessible to local people, as well as a source of healthy food for low income families. Phase four will require significant community input, and will be most successful if the Remediation Team supports local leadership rather than spearheading the restoration. This park belongs to the city and its people and should be treated as such."	"At the end of the day I'm doing this because if this pipeline does fail, it's going to lead to massive water contamination and potential illness and death among our communities. But if we succeed, our communities are going to have a better quality of life because their water will be safe."—Course A1	"Possible outcomes in the event of not receiving a vaccine—implications on public health, health insurance, economy, etc."
	4.2. Limitations: Forecast possible limitations of your plan and provide resolutions.	"A limitation of this plan is the timeline—completely renovating this area will take longer than the projected opening of the community garden date. This is due to the high toxicity of PAH compounds and alkyl halide waste. We recommend waiting at least fifteen years after remediation to develop an in-ground vegetable garden, pending toxicity reports from the analytical and organic chemistry teams and EPA guidelines. In the interim, the Remediation Plan suggests the use of above-ground planter beds as community garden plots."	"Limitations … that just means if my plan failed, people may get sick. There would be detectable levels of alkaloids & bromine in the squash and oil saturating the food." —Course C	"This is great because it will force students to be metacognitive and see inherent holes in their plan. Also how to mitigate those potential issues ahead of time."

"[Objective] feels like a really, really important step. So, how someone frames a problem determines everything else, right? And one of the things that I try to do in this class is really focus on framing the problem earlier on and how you go about framing that problem—the approach. Because if framing is narrow then so are the solutions."—Faculty Interview, Course B

Disciplinary Grounding Construct

Within the disciplinary grounding construct, most criteria were clear to students throughout phases 2 and 3 (Table 3); however, several students in both phases found disciplinary research methods (criterion 2.3) to be challenging based on their lack of knowledge across disciplines:

> "I found talking about the specific methodology of each discipline to be really difficult, because I mean, even if you were a PhD Microbiologist, it would be difficult to understand the depth of knowledge that you need for each of these issues."—Student Interview, Course A1

This mirrors ideas from the IDSF, in that students may need to focus on methods from one particular discipline in which they have developed knowledge, while providing provisional knowledge of methods from other disciplines (Tripp and Shortlidge, 2019). Faculty assisted in providing justification for why students were typically providing simplistic laundry lists of methods:

> "Because [students have] never solved big problems literally like this, they're just thinking. They're not actually getting to the nuts and bolts [of methods]. It's interdisciplinary thinking versus execution. [Methods] is important though and in my opinion, since I've been probably doing interdisciplinary type teaching for ~20 years, I would say that methods are often removed from interdisciplinary thinking. So I agree the next step is how well can they execute a plan and potentially be a project manager by explicitly stating what their mode of action is and what methods they'll use to do it."—Faculty Interview, Course C

Given that students were not misinterpreting this criterion and faculty saw the importance in students including methods from more unfamiliar disciplines, we changed the language in the levels of thinking to more explicitly help instructors score this criterion and better reflect the tiered thinking skills that students commonly expressed (Table 2).

Integration and Collaboration Constructs

We initially had several criteria within integration that led many students to list disciplines needed without an intentional effort to integrate knowledge in their essays in phase 2:

> "I recommend a team of surveyors to monitor the erosion in the immediate area, soil specialists and engineers to make an updated survey of the immediate topography, as well as researchers to study innovative erosion remediation techniques."—Student Essay, Course A1

To assist students in the rather advanced task of truly integrating different disciplinary pieces of information, we reworked this construct to include specific instructions to *leverage* each discipline's knowledge/methods in a way that will be useful to the other disciplines involved (criterion 3.1). Rewording integration in this manner also inherently required a collaboration component for students to address (Table 2). This resulted in essays with much richer integration between disciplines evidenced by students providing ways to synthesize information into a cohesive whole through involvement and reliance on other disciplinary fields and/or expertise (Table 3).

With this inherent interplay between integration and collaboration, we chose to collapse the collaboration construct into a criterion within integration (Table 3). Several students in phase 3 then reflected on the integration construct as the heart of what makes the assignment "interdisciplinary science," such as this student from course A2:

> Student: "The integration construct helped me use interdisciplinary science to tackle this problem on this assignment."
>
> Interviewer: "So this specific construct [integration] was the one that made you realize that you have to be interdisciplinary in answering the essay?"
>
> Student: "Yeah. Because you needed to have your group work together and build the different areas in their respective fields to solve this problem better."

Conversely, a different proportion of our data from phase 3 reflected a difficulty for students to address collaboration in their essays but for reasons that were quite unexpected; students were not misinterpreting this criterion, but rather were afraid to include evidence of collaboration for fear of the instructor's stringent grading on such a requirement:

> "I did not include anything about collaboration because [this essay] is going to be a thing that I'm graded on by my professor. They're pretty far down the rabbit hole of, 'You're in a science class. You are not in an intentional community class.' I felt like any attention I gave to this would be counted against me."—Student Interview, Course C

> "[Collaboration]is just difficult to address because then you get into the whole discipline of psychology and sociology and that's not a part of science, or this class … definitely not. I think the structuring, the whole information gathering, that's kind of how I addressed it; information gathering aspect; that everybody played a critical role. Yeah. I mean it's a clear question. It's just hard to answer because of personality differences, personality disorders, and this class subject isn't about people."—Student Interview, Course C

Because we only observed this kind of feedback from course C, we suspect that the nature of instruction and/or more purely disciplinary content had an impact on if and how students addressed collaboration. Given the fundamental importance of collaboration in ID science work expressed by content experts in this study (Table 3) and the literature (NRC, 2003, 2009; Borrego and Newswander, 2010; AAAS, 2011), we posit that this criterion is foundational in acquiring real-world problem-solving skills. This serves as an important lesson for science instructors to embed ID activities and assessments that span beyond STEM disciplines. On the other side of the coin, this feeds into disciplinary humility and the necessity to foster respect, appreciation, and inclusion of social sciences and humanities into science courses, as collaboration within these fields is how real-world problems are best addressed. We are not arguing that deep disciplinary knowledge in scientific content is unimportant, but rather that it should be accompanied with an application aspect to help students broaden their awareness outside science.

To help mold students' ID thinking toward these ends, in phase 3 we created a broader awareness construct with a criterion of "social impacts" (4.1), as our previous study revealed a deficit of students including disciplines outside STEM (politics, business, economics, sociology, etc.; Tripp *et al.*, 2020). We also decided to change language in the rubric from "disciplines" in phase 2 to "disciplines and/or experts" in phase 3 (Table 2) to

broaden students' ability to include individuals who may or may not be strictly within a discipline (e.g., native peoples, community groups). This simple rephrasing resulted in an overwhelming shift in students incorporating "nontraditional" disciplines and individuals as part of the equation to address the problem.

Despite the hesitation from some students in course C to incorporate collaboration, a few students from this class were able to extend their thinking not only to address collaboration, but also incorporated nontraditional and non–STEM fields:

> "The most crucial but often overlooked step towards the remediation of such a level of contamination is clear and accessible communication with the local community. This includes not only the locals, but the indigenous community and tribal leaders. Indigenous people are often the most severely affected by such disasters, as they are more connected to and reliant on the environment; an environmental disaster such as the BP oil spill proves mentally, physically and spiritually harmful to those people. The Louisiana coast is traditionally the land of the Houma and Choctaw Nations. Therefore, it is crucial to communicate with the tribal council of these nations the nature and magnitude of the contamination. This is mutually beneficial, as indigenous people are holders of traditional ecological knowledge, a vast but often untapped body of knowledge developed and refined over the course of centuries."—Student Essay, Course C

When we asked this student about crafting the essay in this manner, the student indicated the social impacts criterion as influential:

> "In my own essay, I talked about the safety of doing a community garden with the uptake problems of toxins like heavy metal toxins in a former industrial area. And you have to think about the people that it will impact. So I think that 'social impacts' is a good thing to have in the rubric and I think that it's a … social awareness thing. All of these [criteria] are weighted equally, but that one's pretty important in my opinion."—Student Interview, Course C

This indicates that, regardless of the content taught in courses, the social impacts criterion may help students broaden their approach to the issue by exhibiting disciplinary humility and collaboration through the inclusion of STEM and non–STEM disciplines and community members.

Disciplinary Humility Mindset

Disciplinary humility posits that students will likely need to gain respect and open-mindedness toward other disciplinary perspectives, both within and outside STEM disciplines (Tripp and Shortlidge, 2019). We included it as a criterion for students to meet in phase 2 of our rubric development. However, it became clear that the majority of students were exhibiting levels of disciplinary humility by addressing criterion 2.2—disciplinary reasoning:

> "I guess ['disciplinary reasoning' and 'disciplinary humility'] seem like the same thing to a large extent. Because they are both just pressing for the need to explain why it is necessary to have an interdisciplinary approach—they are similar in context. And you can't have successful collaboration without disciplinary humility so you really have to have humility to even answer a question like this."—Student Interview, Course A1

The IDSF also reflects this overlap by stating disciplinary humility is the thread that runs throughout ID science understanding and will likely increase as students have an opportunity to think and work interdisciplinarily (Tripp and Shortlidge, 2019). It logically follows that, as students go through the process of thinking in this way, they will inherently gain a level of disciplinary humility without having to describe the process of acquiring this mindset in their essays. As such, we excluded this requirement from the final rubric provided in this paper.

Modifications for Broad Use in STEM and non–STEM Disciplines

The final, fundamental aspect of disciplinary humility, and arguably at the core of ID science, is one of inclusion—the IDSF specifically highlights the importance of integration and collaboration across a spectrum of STEM, social sciences, arts, and humanities fields (Tripp and Shortlidge, 2019). Throughout the development of the IDSR, we deliberately modified the rubric to be broadly useful across disciplines, both within and outside STEM fields. As mentioned, students in our study ranged in major, background, course, and institution. Thus, amendments were made to the IDSR between phases 2 and 3 to be more inclusive of the spectrum of undergraduate populations that may be scored with this instrument. For instance, the word "hypothesis" was initially used in tasking students to formulate a plan. Based on novice and expert feedback, we removed language such as this, as students may have felt confined to the natural and physical science fields. To be civic leaders and contributors to the challenging real-world problems we continually face, students will undoubtedly interface with disciplines outside STEM and may need to act as stewards to lower the hierarchical barriers between competing ideologies across disciplines (Garibay, 2015; Tripp and Shortlidge, 2019). This assignment and rubric may be a small step toward cultivating a mindset that prepares them for these challenges.

Scoring and Scale

We analyzed the levels of ID science thinking (i.e., scale) for the IDSR through instructor interviews ($n = 4$) across the four courses, as well as a panel of discipline-based education researchers ($n = 11$) in phase 3 (Figure 3). Three instructors suggested expanding the three-point scale, as they found some students' responses falling in between the three levels of thinking (e.g., students were exhibiting thinking that fell between mastery and apprentice). Similarly, the discipline-based education researcher group suggested adding a fourth level to more fully represent the array of knowledge that students were directly exhibiting in their essays. For instance, we observed that some students communicated leveraging disciplines/expert contributions, but did not specifically address how to leverage the knowledge and/or methods from these individuals in a way that would be useful to the project (criterion 3.1). However, the initial three-point scale did not include this as a scoring outcome. Thus, we added another level to the IDSR and redefined the existing

levels based on what students were actually exhibiting in their essays. We ensured that the spectrum of ways students described (or did not describe) each aspect of interdisciplinary science was represented in these tiers of scored thinking. This resulted in the measurement of ID science thinking across a four-point spectrum: mastery (3), intermediate (2), novice (1), and naïve (0; Table 2). The numerical values assigned to these levels can be changed based on assignment point value and instructor preference.

Interrater Reliability and Statistical Findings

The instructors of the courses who were previously unfamiliar with our instrument and had no training in using it, scored essays from their courses with high reliability in score interpretations with one of the rubric designers (B.T.; $\kappa = 0.67$). This indicates that practitioners can use the IDSR without being trained in how to interpret or grade students' work and additionally attests to the reliability of data collected from the IDSR.

In examining differences in students' thinking across populations, we found no significant differences between essay scores in phase 3 across courses A2, B, C, and D ($F = 0.72$, $p = 0.79$, $n = 85$; see Supplemental Material D for descriptive statistics). We did not include course A1 in our statistical analysis, as this group received the unrevised version of the IDSR (in phase 2) and thus is statistically incompatible with the other courses. As mentioned previously, we did not include scores from the format category, as this element is not directly related to ID science thinking and would likely skew the data based on construct irrelevance variance (AERA, 2014).

Based on the low sample sizes at the course level (Table 1), we performed a post hoc power analysis (R Studio Team, 2020), which indicated the need for larger sample sizes to effectively make claims about differences in students' ID science thinking across populations and institutions. Nonetheless, as evidenced through the validity and reliability tests, the IDSR can indeed accurately and reliably detect students' ability to think interdisciplinarily in science in a variety of course environments.

Student Perceptions of Rubric and Assignment

One way to measure the outcomes of an activity is to gauge student perceptions of the activity, and how it impacts their interest and learning of the subject (Shortlidge *et al.,* 2018). To evaluate whether the rubric and assignment were assisting students in thinking interdisciplinarily in a valuable way, we inquired about student perceptions of both of these tools in student interviews in phases 2 and 3. Many students expressed that the rubric helped narrow the scope and expectations of the assignment, and moreover, the combination of the assignment and rubric together significantly improved students' perceptions and ability to think interdisciplinarily in science:

> "I felt like in a way, there was so much freedom at first in the prompt. That's where I was having a little bit of a hard time, just looking at the prompt. I guess the rubric helped me to narrow the scope. At first, it was hard just looking off the prompt itself. But the rubric really helped for me to be like, "This is how I develop an interdisciplinary approach." So yeah, I think having the rubric there and using it as a checklist. It helped to answer a lot of questions."—Student Interview, Course A1

> "I think this assignment was kind of profound. When you're tasked with something that is as far reaching as this is, it's really … It's taxing. It makes you think outside the box. It makes you take a step back and figure out what you actually know is effective for approaching the problem, and what sort of things you actually need to do and who to involve to help you get there."—Student Interview, Course B

> "[The rubric] was helpful because I think it's hard for, I don't know what you call them, hardcore science students, to pay mind to a lot of these things that seem also focused on social stuff than the hard sciences, but I think that it's important and the rubric is what got me to expand my thinking in that way."—Student Interview, Course C

> "I think the assignment really helped me understand what it is like to be a real scientist. I don't feel like I'm necessarily there yet obviously, like how to ask the questions or propose how to solve problems. So I guess this assignment forced me to think that way. It's the first real assignment where I've had to think about it from the start and not necessarily guided like I am in labs."—Student Interview, Course D

Students also verbalized a greater understanding and necessity for interdisciplinarity to solve real-world issues based on the assignment:

> "The types of questions that are most interesting don't just stay in one category, as [evidenced] by this assignment. They aren't under a single discipline. They are far reaching, and you need a lot of different background knowledge to understand some of those more critical things. I don't know if I can think of any real question in science that doesn't require that you understand at several different levels what's actually going on. So, interdisciplinary studies really breaks down a lot of walls that are created when we make those boxes for different fields, you know? When we say, 'Okay, this is the chemistry … We're in chemistry class, just open up the chemistry box.' Or, 'We're in biology, just open up the biology box.' Or, 'We're in physics, just open up the physics box.' Interdisciplinary assignments like this allow for a lot more bridging of those gaps."—Student Interview, Course A2

Overall, these data indicate that students appear to recognize the significance of ID science based on their experience with the essay assignment and rubric. This indicates that the assignment and the IDSR not only measure conceptualization of ID science, but likely foster a more integrated way of knowing. When students can appreciate the value of learning from fields with different perspectives from their own, they may be able to see the benefits gained from working across disciplines.

LIMITATIONS

Through the development of this instrument, there have been a few noteworthy limitations. Our decision to include a student version of the rubric may have led students to think about interdisciplinarity in science from a narrower perspective. We based this decision on a pilot study in which we withheld the student rubric from the assignment, resulting in diffuse, tangential, and unstructured essay responses (Tripp *et al.*, 2020). Thus, we resorted to providing students with the student rubric in subsequent pilots. The outcome from this modification revealed that

students were still able to think holistically and creatively about how to arrive at plausible solutions to real-world issues, with the rubric actually assisting in the expansion of their mindset and ability to think "outside the box."

Another limitation may be related to the context of the real-world problem that instructors chose to embed in the assignment, as well as the content that is taught within the course. Several current real-world problems inherently have higher levels of ID science, such as climate change, while others may be more strictly focused on more discrete issues. This could potentially change the level of integration or ID thinking that students are able or required, to exhibit. However, the prompts used in this study varied from tasking students to develop an education plan for sexually transmitted infections to mitigating damaging effects from the construction of the Keystone pipeline. The rubric was applicable to each type of problem regardless of the content covered in the courses.

Finally, we were unable to draw conclusions about the efficacy of the IDSR related to student demographics (age, gender, race, English-language learners, etc.). We did not collect demographic information, nor was it included in our IRB protocols, as the intention was to first validate the data across general populations. Collecting English composition or literature course scores and correlating them with students who completed the assignment would have been informative in understanding the influence of student reading and writing ability on essay scores. Disaggregating influences such as language ability compared with scientific knowledge would prove an interesting next step. We did, however, gauge students' ability to explain interdisciplinary science in interviews compared with their essay scores and used these data to inform the effectiveness and accuracy of the IDSR constructs and criteria. However, no convergent validity was formally established, as the interview questions were not designed to capture this type of validity evidence. We highly encourage practitioners and researchers to extend this work with larger sample sizes to examine if and how demographic factors may impact the accuracy of the rubric in capturing ID science thinking.

CONCLUSION AND FUTURE DIRECTIONS

Research-based disciplines have a fundamental duty to pose and answer research questions that are often contingent on valid data collected from highly calibrated instruments. In this study, we developed an instrument, the Interdisciplinary Science Rubric, to assess undergraduate students' ID science thinking related to real-world issues. Through an iterative process based on novice and expert response processes and internal consistency of rubric scores, we have provided evidence that the IDSR is a quality assessment tool to measure ID science thinking in undergraduate education.

Results also revealed that the IDSR and a real-world, problem-based writing assignment can be widely used across institutional and course platforms to measure *Vision and Change*'s ID science competency (AAAS, 2011). We encourage faculty to use these tools to gauge students' ability to think interdisciplinarily and challenge them to broaden their mindset toward more disciplinary inclusion and humility. Furthermore, this activity represents a relatively easy way for instructors to encourage students to practice creating outward-facing solutions to big issues. There are a number of ways that the IDSR and IDSF could be used in various classrooms. For example, a shortened version of the essay assignment could be used as a pre/post examination of ID science thinking to assess the impact of a course on students' conceptualization of this competency. The practitioner/researcher must first, however, maintain validity and reliability measures by examining validity and reliability of data for the given population both before and after administering a modified version of this instrument. This tool could also potentially be used to inform the development of a survey instrument to quantitatively assess students' ID science thinking; however, researchers would first want to establish evidence of validity and reliability of the data collected with the rubric for their student populations.

We contend that, although this assignment and the IDSR have the ability to foster and accurately measure students' ID science thinking skills, this is one brush stroke in the holistic picture of interdisciplinary training and assessment. Interdisciplinary thinking and practice must be intentionally infused into curricula and pedagogy to assist students in working across fields and collaborating with teams to address complex issues. This instrument provides a way to measure a complicated competency, and we look forward to researchers building on the work. We hope this effort inspires educators to use the IDSF and IDSR as guides to create group activities and research projects that engage students in interdisciplinary collaboration. Through these efforts, we can better prepare students to enter the workforce with tools and skills to optimize a fluctuating world of burgeoning societal issues.

ACKNOWLEDGMENTS

We thank the student and expert interviewee participants, as well as instructors for their willingness to incorporate the assignment into their courses. Undergraduate researchers made important contributions to the project; in particular, we thank Sophia A. Vornoff and Analee Pham. We would also like to thank the Biology Education Research Group at the University of Washington for providing insight on the project, and the *LSE* reviewers who provided thoughtful and insightful suggestions for improving the article. And finally, we are grateful to the Biology Leadership Community's Catalytic Grant awarded to E.E.S. (2019) for funding this research.

REFERENCES

American Association for the Advancement of Science. (2011). *Vision and change: A call to action, final report*. Washington, DC, Retrieved July 3, 2018, from http://visionandchange.org/finalreport

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arjoon, J. A., Xu, X., & Lewis, J. E. (2013). Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence. *Journal of Chemical Education*, *90*(5), 536–545.

Balgopal, M. M., & Wallace, A. M. (2009). Decisions and dilemmas: Using writing to learn activities to increase ecological literacy. *Journal of Environmental Education*, *40*(3), 13–26.

Balgopal, M. M., Wallace, A. M., & Dahlberg, S. (2012). Writing to learn ecology: A study of three populations of college students. *Environmental Education Research*, *18*(1), 67–90. https://doi.org/10.1080/13504622.2011.576316

Balgopal, M. M., Wallace, A. M., & Dahlberg, S. (2017). Writing from different cultural contexts: How college students frame an environmental SSI through written arguments. *Journal of Research in Science Teaching*, *54*(2), 195–218.

Barbera, J., & VandenPlas, J. R. (2011). All assessment materials are not created equal: The myths about instrument development, validity, and reliability. *Investigating Classroom Myths through Research on Teaching and Learning*, *1074*, 177–193.

Boix Mansilla, V., & Duraisingh, E. D. (2007). Targeted assessment of students' interdisciplinary work: An empirically grounded framework proposed. *Journal of Higher Education*, *78*(2), 215–237.

Boix Mansilla, V., Duraisingh, E. D., Wolfe, C. R., & Haynes, C. (2009). Targeted Assessment Rubric: An empirically grounded rubric for interdisciplinary writing. *Journal of Higher Education*, *80*(3), 334–353. doi: 10.1353/jhe.0.0044

Borrego, M., & Newswander, L. K. (2010). Definitions of interdisciplinary research: Toward graduate-level interdisciplinary learning outcomes. *Review of Higher Education*, *34*(1), 61–84.

Byrne, E., Sage, C., & Mullally, G. (2016). *Transdisciplinarity perspectives on transitions to sustainability*. London: Routledge.

Connolly, P., & Vilardi, T. (1989). *Writing to learn mathematics and science*. New York: Teachers College Press.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new era cognitive development inquiry. *American Psychologist*, *34*, 906–911.

Garibay, J. C. (2015). STEM students' social agency and views on working for social change: Are STEM disciplines developing socially and civically responsible students? *Journal of Research in Science Teaching*, *52*(5), 610–632. https://doi.org/10.1002/tea.21203

Keys, C. W. (1999). Revitalizing instruction in scientific genres: Connecting knowledge production with writing to learn in science. *Science Education*, *83*(2), 115–130.

Klein, J. T. (1990). *Interdisciplinarity: History, theory, and practice*. Detroit, MI: Wayne State University Press.

Klein, J. T. (1996). *Crossing boundaries: Knowledge, disciplinarities, and interdisciplinarities*. Charlottesville: University of Virginia Press.

Klein, J. T. (2000). A conceptual vocabulary of interdisciplinary science. In Stehr, N., & Weingart, P. (Eds.), *Practising interdisciplinarity* (pp. 3–24). Toronto: University of Toronto Press.

Klein, J. T. (2005). Integrative learning and interdisciplinary studies. *Peer Review*, *7*(4), 8–10.

Klein, J. T. (2015). Reprint of "Discourses of Transdisciplinarity: Looking Back to the Future." *Futures*, *65*, 10–16.

Knekta, E., Runyon, C., & Eddy, S. (2019). One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research. *CBE—Life Sciences Education*, *18*(1), rm1. https://doi.org/10.1187/cbe.18-04-0064

National Institutes of Health. (n.d.) *National Institute on Drug Abuse home page*. Retrieved September 12, 2019, from www.nih.gov/about-nih/what-we-do/nih-almanac/national-institute-drug-abuse-nida .

National Research Council (NRC). (2003). *BIO2010: Transforming undergraduate education for future research biologists*. Washington, DC: National Academies Press.

NRC. (2009). *A new biology for the 21st century*. Washington, DC: National Academies Press.

Newell, W. H. (1990). Interdisciplinary curriculum. Issues in Integrative Studies. *Newell*, *8*, 69–86.

Newell, W. H., & Green, W. J. (1982). Defining and teaching interdisciplinary studies. *Improving College and University Teaching*, *30*(1), 23–30.

Öberg, G. (2009). Facilitating interdisciplinary work: Using quality assessment to create common ground. *Higher Education*, *57*(4), 405–415.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. Newbury Park, CA: Sage.

Repko, A. F., & Szostak, R. (2020). *Interdisciplinary research: Process and theory*. Thousand Oaks, CA: Sage.

R Studio Team. (2020). *Integrated develolpment for R. RStudio PBC*. Boston, MA. http://www.rstudio.com/

Reynolds, J. A., Thaiss, C., Katkin, W., & Thompson, R. J. (2012). Writing-to-learn in undergraduate science education: A community-based, conceptually driven approach. *CBE—Life Sciences Education*, *11*(1), 17–25. https://doi.org/10.1187/cbe.11-08-0064

Rivard, L. O. P. (1994). A review of writing to learn in science: Implications for practice and research. *Journal of Research in Science Teaching*, *31*(9), 969–983.

Shortlidge, E. E.; Rain-Griffith, L., Shelby, C., Barbera, J., & Shusterman, G. P. (2018). Despite similar perceptions and attitudes, postbaccalaureate students outperform in introductory biology and chemistry courses. *CBE—Life Sciences Education*, *18*(1), ar1. doi: 10.1187/cbe.17-12-0289

Stangor, C. (2014). *Research methods for the behavioral sciences* (5th ed.). New York: Houghton Mifflin.

Stevens, D. D., & Levi, A. J. (2013). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. Sterling, VA: Stylus.

Tripp, B., & Shortlidge, E. E. (2019). A framework to guide undergraduate education in interdisciplinary science. *CBE—Life Sciences Education*, *18*(2), es3.

Tripp, B., Vornoff, S. A., & Shortlidge, E. E. (2020). Crossing boundaries: Steps toward measuring undergraduate interdisicplinary science understanding. *CBE—Life Sciences Education*, *19*(1), ar8.

Wren, D., & Barbera, J. (2013). Gathering evidence for validity during the design, development, and qualitative evaluation of thermochemistry concept inventory items. *Journal of Chemical Education*, *90*(12), 1590–1601. https://doi.org/10.1021/ed400384g