A New Paradigm for Research-Based Assessment Development

James T. Laverty,[1] Amogh Sirnoorkar,[1] Amali Priyanka Jambuge,[1] Katherine D. Rainey,[2] Joshua Weaver,[1] Alexander Adamson,[1] and Bethany R. Wilcox[2]

[1]*Department of Physics, Kansas State University, Manhattan, Kansas 66506*
[2]*Department of Physics, University of Colorado Boulder, 390 UCB, Boulder, CO 80309*

Research based assessments have a productive and storied history in PER. While useful for conducting research on student learning, their utility is limited for instructors interested in improving their own courses. We have developed a new assessment design process that leverages three-dimensional learning, evidence-centered design, and self-regulated learning to deliver actionable feedback to instructors about supporting their students' learning. We are using this approach to design the Thermal and Statistical Physics Assessment (TaSPA), which also allows instructors to choose learning goals that align with their teaching. Perhaps more importantly, this system will be completely automated when it is completed, making the assessment scalable with minimal burden on instructors and researchers. This work represents an advancement in how we assess physics learning at a large scale and how the PER community can better support physics instructors and students.

2022 PERC Proceedings edited by Frank, Jones, and Ryan; Peer-reviewed, doi.org/10.1119/perc.2022.pr.Laverty

I. INTRODUCTION

Physics Education Researchers have extensively focused on improving student learning by designing and disseminating new teaching environments, materials, and pedagogical approaches. One of the most common ways to generate evidence that these new approaches are productive has been through the use of research-based assessments (RBAs), often also referred to as concept inventories. This "design and disseminate" paradigm has been shown to be relatively ineffective at producing change[1]. While instructors are frequently aware of such innovations, many have not tried them and others have tried and discontinued their use[2]. Other approaches that are more effective at promoting change include developing reflective teachers and developing shared visions. Within these approaches, the role of RBAs becomes unclear, particularly because instructors are often not certain how to translate scores into concrete changes in the classroom[3].

This paper presents a new paradigm in RBA development that focuses on providing instructors with actionable feedback about their own courses. Instead of designing RBAs to support researchers to find more productive approaches to teaching and learning (which they can then disseminate)[4], we want to design an RBA that directly supports instructors trying to improve their students' learning. This represents a shift in how we think about improving student learning; instead of researchers focusing on fixing how instructors teach, this new approach to RBAs can promote researchers and instructors focusing on what students learn. By shifting the focus of change from teaching to learning, we shift from telling instructors how they should teach to empowering instructors to ground their instruction based on their own students' needs.

While many instructors have used existing RBAs, many have also expressed their limitations, often feeling that they do not measure what the instructor cares about, or that they are not sure how to interpret the results (e.g. scores or effect size) to improve their teaching[3]. This new paradigm, which focuses on delivering actionable feedback to instructors, is currently being used to design the Thermal and Statistical Physics Assessment (TaSPA), from which we draw examples in this paper. At its core, we aim to develop an RBA that supports instructors by taking information from the assessment and using it to support their students' learning. This paper is important to both RBA designers and researchers interested in promoting instructor change.

In this work, we address several concerns about existing RBAs: 1) Instructors must use a static assessment, which may or may not align with what they feel is important; 2) It is not clear to most instructors (or even researchers) how scores from RBAs can be used to modify teaching in an individual course; and 3) The community is not measuring what it cares about, that is whether or not students can 'do physics'[5]. In particular, this paper aims to answer the following guiding question: How can we adopt and adapt existing assessment theories to design an RBA that directly supports faculty/instructors in improving their students' learning? We then provide some evidence that the approach is reasonable and promising, drawing on interviews we have conducted with faculty.

II. TASPA - INSTRUCTOR VIEW

To guide the reader, we present a quick summary of our vision for what using TaSPA will look like from the instructor perspective. We ask the reader to take the perspective of an instructor teaching an undergraduate thermal and/or statistical physics course who is interested in seeing how effective their course has been at helping students learn particular ideas. With this in mind, the instructor goes online to the TaSPA portal and chooses *learning performances* (defined in Section III B) that align with their course learning goals. This addresses concern 1 listed in the previous section. For example, the instructor might choose to assess if their students can "Construct an argument justifying or refuting claims about the changes to internal energy of a thermodynamic system given information about the energy flow into and out of the system." After choosing this (and perhaps other learning performances), the instructor is then asked for contact information for the students and the time frame they would like the assessment to be available to the students.

During that time frame, the assessment is available to the students and invites them to take it. The instructor then receives a report that provides information about the students' performance and provides actionable feedback to the instructors. Continuing the example, the instructor receives:

> *Students were expected to relate changes in internal energy of a system to both heat and work as forms of energy that flow into and out of that system. Many of the students were able to relate one, but not both of these quantities to the internal energy. We recommend providing more opportunities for students to relate both heat and work to changes in internal energy of real-world scenarios.*

This feedback addresses concern 2 from the previous section. After receiving the feedback, it is up to the instructor to decide what to do next.

III. THEORETICAL FRAMEWORKS

Development of an RBA should start by identifying its purpose and potential use, then validating its design against those stated goals[6, 7]. In this approach, we conceptualize assessments as instruments that support instructors interested in improving their students' learning (arguably, all instructors). We do this by articulating a theory-of-action for our assessment and its associated feedback [8–12].

A theory-of-action includes intended long-term and short-term effects, mechanisms that lead to those effects, the components of the assessment, and its measurement argument. In this paper, we focus on connecting our short-term effect (providing actionable feedback to course instructors), with the measurement argument. To do this, we outline a novel approach that leverages three-dimensional learning, evidence-centered, design, and self-regulated learning.

A. Three-Dimensional Learning

To determine what we should assess, we draw on the assessment design framework laid out by the Next Generation Science Standards; namely, three-dimensional learning (3DL). 3DL divides what we want students to learn into three "dimensions" of learning. A brief description of each of the three dimensions is given below. More details can be found in Ref. [13].

Scientific practices are components of the process of science. They focus on using scientific knowledge to model, predict, and explain phenomena. *Crosscutting concepts* are "ways of thinking" that bridge scientific disciplines. They are used to explore phenomena in distinct ways. *Disciplinary core ideas* are concepts fundamental to each scientific discipline. These concepts are required to explain a wide range of phenomena and provide a way to generate new ideas and predictions.

3DL emphasizes blending all three dimensions together into instruction, curriculum, and (most notably here) assessment (sometimes referred to as "knowledge-in-use"). While 3DL was written for the K-12 education system, it has been argued that these ideas are relevant to higher education[14–16] and instructors seem supportive of this (see Section V).

B. Evidence-Centered Design

We take the position that assessment is about building an evidence-based argument of what students can do. Thus, assessments are designed to collect the evidence needed to support that argument[17]. Evidence-Centered Design is one approach to assessment development intended to insure coherent assessment task design. It has also specifically been highlighted as productive for assessing 'knowledge-in-use'[18].

We draw on the work of Harris et al. who divided task design into a series of theory-based steps that build on each other to construct an evidentiary argument[19]. The first step in task design focuses on identifying *learning performances* which are assessable statements about what students should know and be able to do with their knowledge. The learning performance is ultimately the claim that assessment developers want to make regarding the students' abilities. Thus, the assessment must collect evidence to support (or refute) the learning performance.

The next step is to identify the *Knowledge, Skills, and Abilities (KSAs)* that students must have in order to accomplish the learning performance. These ultimately serve as the targets of the assessment. That is, if students demonstrate these KSAs, we can make the claim that the students are meeting the learning performance. *Evidence Statements* are the observable features of a student's response that would allow us to claim that the student has the KSAs we are trying to assess. In turn, the evidence statements articulate the observable features of student responses needed to support the claim that they meet the learning performance. *Task Features* are the design elements that tasks need in order to elicit observable aspects of the evidence statements. These features should be present in any task that could be developed for a particular learning performance. Lastly, the *Rubric* builds from the evidence statements. The rubric defines what kinds of responses from students are considered to provide strong, weak, or no evidence that the student has met the learning performance. Ultimately, the rubric provides a systematic way to evaluate the evidence provided by student responses.

All of the features above, as well as the tasks themselves, must be validated with student responses. In practice, this validation process typically leads to iterative modifications of each of these components. Taken together, this theoretical approach allows assessment developers to construct arguments that students are providing evidence that they meet the learning performance. However, we note that the opposite is not true. That is, we cannot construct an argument that students are incapable of meeting the learning performance; we only can state that they did not provide evidence that they meet the learning performance.

C. Self-Regulated Learning

We use a model of self-regulated learning to support our approach to include explicit instructor feedback in the assessment design. We treat the instructor as the 'learner" identified within self-regulated learning. This learner undergoes the learning process while engaging in the activity of developing, teaching, and modifying their course. This enables us to identify features of external feedback that can support the internal processes of an individual engaged in the activity of teaching a course. [20, 21]

Combining Self-Regulated Learning with the idea of formative assessment, Nicol and Macfarlane highlight seven principles of good feedback[22]. We focus on the three most relevant for our purposes: 1) Clarify what good performance is, 2) Provide the current state of the performance, and 3) Provide opportunities to close the gap between the current and desired performance.

IV. TASPA - DEVELOPER VIEW

In this section, we combine these theoretical frameworks to connect identifying what to assess, with our measurement argument, and subsequent feedback for instructors. Figure 1 shows a summary of how the theories and development stages

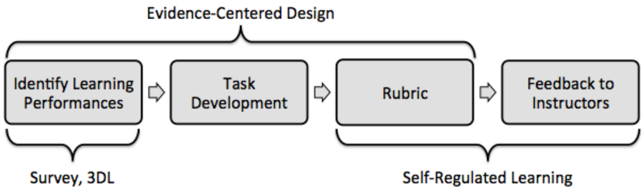

FIG. 1. Summary of how the theories (brackets) inform the different stages (boxes) in the development of the TaSPA.

fit together. Table I shows examples taken from the development of a single task for the TaSPA.

Identifying Learning Performances: As developers, the first challenge for developing tasks comes from deciding what to assess. For the TaSPA, we addressed this problem using 3DL, and by conducting a survey of faculty who recently taught a thermal and/or statistical physics course. The survey was populated with topics identified via common textbooks and a list of the scientific practices from 3DL. In total, 73 people responded to the survey. We then used these survey results to guide the development of our learning performances, by focusing on topics that the majority of respondents selected as important. For more details about the survey, see Ref. [23].

In addition to the survey, We use 3DL to articulate specific assessable learning performances, by combining a scientific practice, a crosscutting concept, and a disciplinary core idea. We draw on the full list of scientific practices and crosscutting concepts from the NGSS, but modify the list of core ideas slightly for this upper-level undergraduate context. Specifically, we consider the core ideas most relevant for this assessment to be Energy, Entropy, and Properties of Matter. For the example shown in Table I, we have combined the scientific practice of *engaging in argument from evidence*, the crosscutting concept of *scale, proportion, and quantity*, and the core idea of *energy*. We used the survey results to guide which combinations are likely relevant to many instructors.

TABLE I. Examples of each step in the development of a single task for TaSPA. All examples are for the same Learning Performance.

Step	Example
Learning Performance	Construct an argument justifying or refuting claims about the changes to internal energy of a thermodynamic system given information about the energy flow into and out of the system.
Knowledge, Skills, and Abilities	KSA2: Use the unpacked relations [from KSA1] to generate an explanation about the change in internal energy of the system.
Evidence Statements	ES1: Relations that connect change in internal energy to heat and work.
Task Features	Question asks student to make a claim based on the given event, and justify the claim using appropriate physics principles.
Rubric	Partial Evidence (Identifies heat or work, not both)
Feedback (Goal)	Students should relate changes in internal energy of a system to both heat and work as forms of energy flow into and out of that system.
Feedback (Current State)	Students related changes in internal energy of a system to either heat or work as forms of energy flow into and out of that system, but not both.
Feedback (Achieving Goal)	Give more opportunities to relate both heat and work to changes in internal energy of real-world scenarios.

We note that these Learning Performances are larger in grain-size than what many existing RBAs assess. This is an intentional choice that aligns with existing work on assessing 3DL[13–16, 24]. This also supports our argument that the assessment can be used to determine if students can 'do physics'[5].

Task Development: Having identified the Learning Performances, we follow Harris et al. by identifying the KSAs students would need in order to meet the learning performance][19]. In practice, there are usually 2-4 of these per learning performance. An example of one (of 3) can be found in Table I. We then identify the evidence statements (typically 2-4), often building them directly from the KSAs. This means that there are typically 2-4 evidence statements for each learning performance. The task features are guided by a combination of the evidence statements, and the Three-Dimensional Learning Assessment Protocol (3D-LAP) [24]. The 3D-LAP consists of criteria for tasks to elicit evidence of students engaging in scientific practices, crosscutting concepts, and core ideas.

For the TaSPA, the task needs to align with each of the components above, which often includes choosing a real-world context. Finding such a context is usually the starting point for developing the task, and oftentimes the hardest part. After developing an open-ended version of the item, we collect student responses to the task, develop a rubric based on the responses, and then revise the item into a coupled, multiple-response task.

Rubric: The rubric is strongly informed by the evidence statements. We analyze student responses to the open-ended task and use the rubric to categorize the results into one of three categories: (1) The student's answer provides evidence that they have the knowledge, skill, or ability (Evidence Met); (2) the student's answer provides some, but not enough evidence of their knowledge, skill, or ability (Partial Evidence); or (3) the students answer provides no evidence (No Evidence). This rubric then informs the development of the coupled, multiple-response task. Knowing what evidence is needed for certain rubric ratings, and what students actually answered helps us determine how to split the original question into several smaller ones, and what answer options should be available[25]. For more details on this task development, see Ref. [26].

Feedback for Instructors: Each rubric rating, described above, includes feedback designed to help instructors modify their instruction. These suggested modifications are focused on specific recommendations that could help students meet

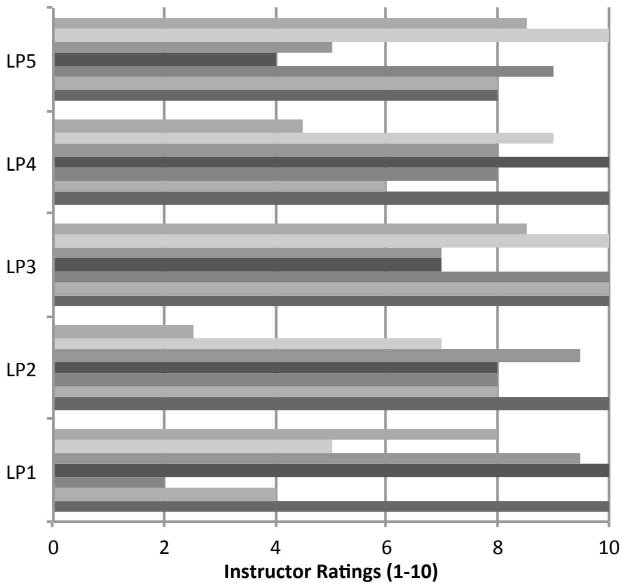

FIG. 2. Ratings each of the seven instructors assigned to each learning performance in the interview. Ratings were averaged if the instructor gave a range (e.g., "9 or 10" became 9.5).

the learning performance. Feedback to the instructor is based on the relative frequency of rubric ratings among all of the students in the course. For example, if most students reach the 'Proficiency Met' category, the feedback states that "no modifications are suggested."

The feedback for each rubric rating contains three distinct pieces, to align with the theory of self-regulated learning highlighted in Sec. III C. Thus, for each rubric rating, there is a listed goal, information about the current state of the students' performance, and suggestions for changes that could help the students meet the goal. Table I shows an example of the feedback, broken into these three pieces. The development of the feedback is described in more detail in Ref. [27].

V. INTERVIEWS

The goal of this paper is to articulate the theoretical structure for a new paradigm of assessment development, not to analyze specific data. However, to provide some evidence that the approach is reasonable and that faculty are receptive to it, we conducted interviews with seven faculty members from across the United States who have recently taught an undergraduate thermal and/or statistical physics course. The interview protocol included questions about the faculty member's course, their impressions of five example learning performances, and their thoughts on several examples of feedback.

Specifically, we asked each faculty member to rate each learning performance on a scale from 1 (least likely to assess) to 10 (most likely to assess). The results can be found in Fig. 2. The faculty had an overall positive reaction to the learning performances, which is noteworthy because it demonstrates that using 3DL to design learning performances for upper-division physics courses is viable. Additionally, every faculty member rated some of the learning performances highly. This suggests that every faculty member would find something they believe is important to assess in the TaSPA. Also, every faculty member found different learning performances important. This highlights and validates our design choice to allow faculty to decide which learning performances they want to assess.

Finally, in response to the interview questions about feedback, we highlight a single quote which supports our approach. In response to the example feedback, one instructor responded,

> *I like...this is really easy to understand. I have used some assessments before and they're often like, here are lots of graphs and things, and I don't quite know how to interpret this. This is very clear, and not jargon-y as saying like, here is something they seem to conceptually have trouble with.*

though we recognize a single quote is not strong evidence.

VI. DISCUSSION & CONCLUSIONS

We argue that the TaSPA's theory-driven design represents a new paradigm for RBAs; one that directly supports faculty trying to become better teachers as determined by student learning. This new approach allows instructors to select learning performances that they feel align with their courses, and the feedback provides actionable ideas about how to better support their students' learning. Additionally, the TaSPA aligns with contemporary ideas that we should be assessing students' doing what physicists do with physics concepts.

While the data presented here is minimal, given the strong positive response from instructors in our initial interviews, we feel confident that this approach will appeal to, and be useful for, instructors more generally. Moving forward, we plan to develop and validate many more learning performances, tasks, and instructor feedback. Additionally, we are currently developing the official TaSPA portal and automating the instructor feedback generation. We anticipate this assessment being available to the public in 2024.

ACKNOWLEDGMENTS

We would like to thank the Kansas State University Physics Department for their support. Also thanks to KSUPER for their valuable insights. This material is based upon work supported by the National Science Foundation under Grant #2013339 and #2013332.

[1] C. Henderson and others, Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature, JRST **48**, 952 (2011).

[2] C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj, Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process?, Physical Review Special Topics - Physics Education Research **8**, 020104 (2012).

[3] A. Madsen, S. B. McKagan, M. S. Martinuk, A. Bell, and E. C. Sayre, Research-based assessment affordances and constraints: Perceptions of physics faculty, Physical Review Physics Education Research **12**, 010115 (2016).

[4] W. K. Adams and C. E. Wieman, Development and Validation of Instruments to Measure Learning of ExpertâLike Thinking, International Journal of Science Education **33**, 1289 (2011).

[5] J. T. Laverty and M. D. Caballero, Analysis of the most common concept inventories in physics: What are we assessing?, Physical Review Physics Education Research **14**, 010123 (2018).

[6] M. T. Kane, Validation, Educational measurement **4**, 17 (2006).

[7] M. T. Kane, Validating the interpretations and uses of test scores, Journal of Educational Measurement **50**, 1 (2013).

[8] R. E. Bennett, Cognitively based assessment of, for, and as learning (cbal): A preliminary theory of action for summative and formative assessment, Measurement **8**, 70 (2010).

[9] R. Bennett, M. Kane, and B. Bridgeman, Theory of action and validity argument in the context of through-course summative assessment, Princeton, NJ: Educational Testing Service (2011).

[10] R. E. Bennett, Formative assessment: A critical review, Assessment in education: principles, policy & practice **18**, 5 (2011).

[11] R. E. Bennett and M. von Davier, *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (Springer Nature, 2017).

[12] D. L. Reinholz and T. C. Andrews, Change theory and theory of change: whatâs the difference anyway?, International Journal of STEM Education **7**, 2 (2020).

[13] N. R. Council, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (2011).

[14] National Research Council, *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering* (National Academies Press, 2012).

[15] M. M. Cooper, M. D. Caballero, D. Ebert-May, C. L. Fata-Hartley, S. E. Jardeleza, J. S. Krajcik, J. T. Laverty, R. L. Matz, L. A. Posey, and S. M. Underwood, Challenge faculty to transform STEM learning, Science **350**, 281 (2015).

[16] J. McDonald, The Next Generation Science Standards: Impact on College Science Teaching, Journal of College Science Teaching **45**, 13 (2015).

[17] R. J. Mislevy, R. G. Almond, and J. F. Lukas, A Brief Introduction to Evidence-Centered Design, ETS Research Report Series **2003**, i (2003), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.2003.tb01908.x.

[18] J. Pellegrino, L. DiBello, and S. Brophy, The Science and Design of Assessment in Engineering Education, in *Cambridge Handbook of Engineering Education Research* (2014).

[19] C. J. Harris, J. S. Krajcik, J. W. Pellegrino, and A. H. DeBarger, Designing Knowledge-In-Use Assessments to Promote Deeper Learning, Educational Measurement: Issues and Practice **38**, 53 (2019).

[20] D. L. Butler and P. H. Winne, Feedback and Self-Regulated Learning: A Theoretical Synthesis, Review of Educational Research **65**, 245 (1995), publisher: American Educational Research Association.

[21] C. SIN, Epistemology, Sociology, and Learning and Teaching in Physics, Science Education **98**, 342 (2014).

[22] D. Nicol and D. Macfarlane, Formative Assessment and Self-Regulated Learning: A Model and Seven Principles of Good Feedback Practice, Studies in Higher Education **31**, 199 (2006).

[23] K. D. Rainey and B. R. Wilcox, Faculty survey on upper-division thermal physics content coverage (2020) pp. 494–499, iSSN: 2377-2379.

[24] J. T. Laverty, S. M. Underwood, R. L. Matz, L. A. Posey, J. H. Carmel, M. D. Caballero, C. L. Fata-Hartley, D. Ebert-May, S. E. Jardeleza, and M. M. Cooper, Characterizing College Science Assessments: The Three-Dimensional Learning Assessment Protocol, PLOS ONE **11**, e0162333 (2016).

[25] B. R. Wilcox and S. J. Pollock, Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics, Physical Review Special Topics - Physics Education Research **10**, 020124 (2014).

[26] K. D. Rainey, A. P. Jambuge, J. T. Laverty, and B. R. Wilcox, Developing coupled, multiple-response assessment items addressing scientific practices (2020) pp. 418–423, iSSN: 2377-2379.

[27] A. P. Jambuge, K. D. Rainey, B. R. Wilcox, and J. T. Laverty, Assessment feedback: A tool to promote scientific practices in upper-division (2020) pp. 234–239, iSSN: 2377-2379.