

Affordances of Articulating Assessment Objectives in Research-based Assessment Development

Michael Vignal,^{1,2} Katherine D. Rainey,² Bethany R. Wilcox,² Marcos D. Caballero,^{3,4,5} and H. J. Lewandowski^{1,2}

¹*JILA, National Institute of Standards and Technology and the University of Colorado, Boulder, CO 80309, USA*

²*Department of Physics, University of Colorado, 390 UCB, Boulder, CO 80309, USA*

³*Department of Physics & Astronomy and CREATE for STEM Institute,
Michigan State University, East Lansing, Michigan 48824, USA*

⁴*Department of Computational Mathematics, Science, & Engineering,
Michigan State University, East Lansing, Michigan 48824, USA*

⁵*Department of Physics and Center for Computing in Science Education, University of Oslo, 0315 Oslo, Norway*

Research-based assessments have historically been developed based on teaching experience and/or course learning goals or objectives. However, using course learning goals for assessment development has limitations, including that the goals for a course are often broad and difficult or impossible to assess with an individualized, scalable assessment instrument. Thus, we propose articulating assessment objectives (AOs), which are concise and specific statements about concepts and practices that an assessment aims to measure, as a productive strategy for assessment development. While similar in many respects to learning goals, AOs are explicitly designed to aid in assessment development in numerous ways, including by helping researchers organize high-level assessment goals, providing an additional means for establishing content validity, operationalizing the goals of the assessment via targeted assessment items, and serving as a way to communicate the substance of an assessment to instructors and researchers interested in using the assessment in their course or research study. Here, we discuss these affordances of AOs in the development of two recent research-based assessments, and we present two detailed examples of AOs and how we progressed from initial assessment conception to AO articulation to finalized assessment items. We conclude by arguing that the articulation of AOs is a valuable step in the development of research-based assessments.

I. INTRODUCTION AND BACKGROUND

Since the early 1990's, research-based assessments have helped inform research in physics education [1], and, in recent years, both the pace of assessment development and use of assessments by instructors and researchers has grown [2–6]. Typically, these assessments are not intended to evaluate individual students for the purpose of assigning grades [7]; rather they aim to help course instructors and education researchers identify areas for, or evaluate the effectiveness of, reforms and interventions [8–11].

Assessment developers have employed a number of different assessment-design frameworks, including Evidence Centered Design [12], The Three-Dimensional Learning Assessment Protocol [13], and a framework described by Adams and Wieman [14], while others use no explicit assessment development framework. Independent of framework, a common early step in the assessment development process involves identifying content, concepts, and/or practices that are important to current instructors and content experts, which can be done through a variety of means such as interviews with faculty [15, 16], a review of common textbooks [17], faculty surveys [17], and the collection or creation of agreed-upon priorities for the assessment. This process informs the scope and content of the eventual assessment and guides the development of assessment questions (hereafter referred to as assessment items).

Exactly how assessment developers move from this step to then writing specific assessment items varies by theoretical frameworks and developer experience and expertise, and many assessment developers also employ or modify items from existing assessments [5, 9]. Additionally, assessment developers in PER have not always clearly articulated their framework or the details of the process used to go from this broader, desired scope of the assessment to specific items. However, both framework and process have important implications for the use and interpretation of the outcomes of the assessment, and this lack of clarity can result in assessment results being used inappropriately or being difficult to interpret. Here, we introduce the concept of *assessment objectives* (AOs) as a tool to standardize and clarify this process of developing specific assessment items. As examples of the critical role that AOs can play in developing assessment items, as well as in the design process more broadly, we discuss two recent research-based assessments that were developed using AOs. The two assessments discussed are a thermodynamics assessment, the Upper-level Statistical Mechanics and Thermodynamics Evaluation for Physics (U-STEP) [4] and the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE), which is intended for use in lower-division physics labs.

We define *assessment objectives* (AOs) as *concise, specific articulations of measurable desired student performances regarding concepts and/or practices targeted by the assessment*. We note that the term *assessment objec-*

tive [18] and other similar terms and concepts (learning performances [6, 19], course learning goals [9, 20], educational objectives [21], instructional objectives [7, 21], and more generic “objectives” [7] have been discussed in the literature on both assessment development and curriculum development, though these terms are often not well defined and/or pertain to objectives developed for a course rather than an assessment.

AOs, while having varying degrees of conceptual overlap with these other terms and concepts, are specifically designed to inform assessment development, which distinguishes them in terms of how and why they are developed and the affordances they provide. In particular, we argue that articulated AOs: 1) help us process data on important concepts and practices collected from instructors and other experts; 2) aid in the establishment of content validity; 3) allow us to develop targeted assessment items, and; 4) can help communicate the purpose of the assessment to instructors and researchers interested in using these assessments in their courses and research (e.g., through comparing the AOs to course learning goals).

The primary goals of this paper are to clearly define AOs, present concrete examples of their use, and discuss their unique affordances at all stages of assessment development. These four affordances, as well as examples from U-STEP and SPRUCE, are discussed in the following four sections starting with how to create AOs that are well informed by an analysis of the domain of interest. Throughout these sections, we ground our discussion in example AOs from both the U-STEP and SPRUCE. In Sec. VI, we synthesize our recommendations for assessment developers and includes a discussion of future work.

II. CREATING ASSESSMENT OBJECTIVES

A common first step in assessment development is to “gather substantive information about the domain of interest that will have direct implications for assessment” and then qualitatively code that information [12]. We call this step the domain analysis, which is language borrowed from Evidence Centered Design (ECD) [12]. In this section, we discuss how AOs emerge from, and can interact with, the domain analysis.

Domain analyses might include a review of existing assessments [5, 17], a survey of relevant textbooks [17], and/or directly soliciting input from instructors and content experts [14–17, 23, 24]. Table I includes a summary of the domain analysis steps for U-STEP and SPRUCE (discussed in more detail in [17] and [16] respectively), as well as examples of data from the domain analysis that helped to inform several example AOs referenced in this and following sections.

As described in Ref. [4], AOs are written to “collectively span the space of content areas identified as important based on” the domain analysis and to be “directly

TABLE I. A subset of AOs from the U-STEP and SPRUCE with example data that informed articulation of those AOs and example items targeting those AOs.

U-STEP	SPRUCE
Content Domain Analysis	
• List of topics identified by consulting widely used thermodynamics textbooks • Focus group with instructors and researchers • Content survey distributed to instructors to identify most commonly covered topics	• Open-ended instructor interviews focusing on the teaching and assessment of measurement uncertainty concepts and practices • Content survey distributed to instructors to identify most important topics
Example Data Excerpts	
“Much of thermodynamics deals with three closely related concepts: temperature, energy, and heat. Much of students’ difficulty with thermodynamics comes from confusing these three concepts with each other” [22].	“I think that it’s important [students] understand how to use uncertainty as a way to compare two different measures to one another, and also how to compare their measurement to [an] expected value.” - An instructor during an interview
Example AOs derived (in part) from above data	
Students should be able to: • articulate differences between heat and temperature. • articulate that temperature is a property of a system and heat is not.	Students should be able to: • determine if two measurements (with uncertainty) agree with each other • *determine if a measured value (with uncertainty) agrees with an accepted/expected value • †determine if a single measured value with uncertainty agrees with a distribution of measurements
Example items targeting above AOs	
Example Item 1: Consider the following statement: <i>A thermodynamic system has a certain amount of heat, just like it has a certain temperature, pressure, and volume</i> This statement is... ☐ true ☐ false because...(select all that support your response above) ☐ the amount of heat contained in a system can be calculated from the system’s temperature, pressure, and volume ☐ the amount of heat contained in a system can be calculated from changes in the system’s temperature, pressure, and volume ☐ the amount of heat contained in a system can be calculated from a system’s heat capacity and temperature ☐ heat is a quantity exchanged between systems ☐ heat is a flow of thermal energy ☐ heat is a scalar, like temperature, pressure, and volume ☐ heat is not a state function (<i>i.e., heat is not process independent</i>)	Example Item 2: Using your [previously determined] values, you...calculate your spring constant and uncertainty, and you get the following value: $k = 3.62 \frac{\text{N}}{\text{m}} \pm 0.11 \frac{\text{N}}{\text{m}}.$ Several other lab groups took different approaches to calculating the spring constant. Their values (with estimated uncertainty) are shown below. Select all values you believe agree with your measured value. ☐ $3.71 \frac{\text{N}}{\text{m}} \pm 0.06 \frac{\text{N}}{\text{m}}$ ☐ $3.71 \frac{\text{N}}{\text{m}} \pm 0.17 \frac{\text{N}}{\text{m}}$ ☐ $3.76 \frac{\text{N}}{\text{m}} \pm 0.06 \frac{\text{N}}{\text{m}}$ ☐ $3.76 \frac{\text{N}}{\text{m}} \pm 0.17 \frac{\text{N}}{\text{m}}$ ☐ $3.91 \frac{\text{N}}{\text{m}} \pm 0.06 \frac{\text{N}}{\text{m}}$ ☐ $3.91 \frac{\text{N}}{\text{m}} \pm 0.17 \frac{\text{N}}{\text{m}}$ Example Item 3: You compare your group’s estimate of m with six other groups by sketching your results (gray circles) next to their results (blue triangles) on six different graphs. The error bars in the graphs represent one standard deviation (often referred to as “one sigma” or a “68% confidence interval”). Select all graphs that depict agreement between your data and data from other groups in your class.

*This objective was eventually removed from the assessment because it relates to modeling much more than to measurement uncertainty.

† This objective ultimately was re-framed in terms of identifying outliers and is not assessed by the given example items.

assessable” within the constraints of the assessment format (e.g., multiple-choice questions, multiple-response questions, etc.). The process of writing AOs can be ei-

ther linear or iterative. For example, with the U-STEP, we progressed linearly from completing the domain analysis to writing AOs based on the priorities for assessment

identified during the domain analysis and then to writing assessment items (discussed in the next section): writing the AOs did not impact the domain analysis, and writing items did not impact the AOs.

Alternatively, it is possible to incorporate the writing of AOs into the domain analysis. For SPRUCE, item drafts were developed without the use of AOs, but it was not always clear what concept or practice each item was intended to cover and the research team struggled to “span the space of content areas” with few enough items that the assessment would be a reasonable length. These issues motivated the research team to explicitly articulate AOs, which involved iteratively revisiting and updating domain analysis (i.e., our coding of instructor interviews). For example, several drafted items required students to make comparisons using data they had collected, but it was only after writing AOs for these items that the research team was able to identify several (meaningfully different) types of comparisons that students might make: between two measurements (each with uncertainty), between a measurement with uncertainty and an “accepted value,” and between a single measurement with estimated uncertainty and a distribution of measurements (see example AOs in Table I). Writing AOs helped us explicitly identify these different types of comparisons, and we can then look back at the domain analysis to identify other, similar ideas to make sure we effectively “span the space” of ideas important to experts.

III. USING ASSESSMENT OBJECTIVES FOR CONTENT VALIDITY

An assessment is said to have content validity when “the items adequately sample the domain,” and content validity is established in a review process with content experts [7]. Many assessments primarily or exclusively establish content validity by presenting experts with the finished items (e.g., [3, 5]), but presenting AOs to content experts in order to establish content validity provides the option for early feedback, potentially even before the items are written [7], to ensure that the scope of the assessment (as determined from the domain analysis) is appropriate and aligned with the values and needs of the eventual audience of the instrument. An added benefit is that the list of AOs also provides a simpler mechanism for instructors to review the scope and goals of the instrument without the need to carefully go through the individual items and make assumptions as to their intended purpose.

With the U-STEP, the AOs were used in establishing content validity in multiple ways: they were developed based off interview data with faculty, a follow-up faculty survey (listing topical areas and asking instructors to report if those topics were taught in their course) that received more than 70 responses from instructors nation-wide, and feedback on the full list of AOs was solicited from PER researchers and instructors with experi-

ence teaching upper-division thermal physics. Feedback helped to identify AOs that needed to be modified for clarity or removed from the list due to overly narrow focus. For SPRUCE, a survey asking instructors to rate how important each AO is to them was sent out: we received 19 responses, and these responses helped inform some small changes to our AOs. As SPRUCE is still in development, additional efforts to establish content validity will be forthcoming. As the process of receiving and processing feedback on AOs from instructors can take several weeks, we recommend assessment developers schedule this into their development timeline.

IV. USING ASSESSMENT OBJECTIVES FOR ITEM CREATION

For both the U-STEP and SPRUCE, the primary reason for articulating AOs was to aid in writing assessment items. In this section, we outline several ways one might target a specific AO with an item.

The first AO we discuss for the U-STEP is: “*Articulate differences between heat and temperature.*” This AO is one of several targeted by the example item shown below this AO in Table I. This example item asks students to consider the statement that “A thermodynamic system has a certain amount of heat, just like it has a certain temperature, pressure, and volume.” The conflation of heat and temperature is a well documented student difficulty [22, 25] that was identified during the domain analysis, prompting the articulation of this AO. This item was thus designed to target a specific and fundamental aspect of the difference between heat and temperature: temperature is a function of state and heat is not.

For SPRUCE, the example AO “*Determine if two measurements (with uncertainty) agree with each other*” was targeted with multiple items. There are various reasons to have multiple items target an AO, including to assess a concept or practice at multiple levels of difficulty or across different representations. The two example items given in Table I ask students to evaluate 6 comparisons and determine which depict agreement. The data is presented using two different representations (numerically for example item 2 and graphically for example item 3) in order to identify the impact (if any) of representation on such comparisons. In fact, while these items are presented to students as comparisons in two entirely different experiments, if one were to graph the 6 numeric answer-options in example item 2, one would obtain exactly the 6 graphs in example item 3.

In addition to informing the creation of individual items, articulating AOs can help developers make decisions about the scope of the assessment. By examining our complete list of AOs, we identified four content areas: sources of uncertainty, handling and propagation of uncertainty, uncertainty in distributions of data, and modeling. We concluded that “*Determine if a measured value (with uncertainty) agrees with an accepted/expected*

value” was more about modeling than about measurement uncertainty, and so when we chose to remove modeling from the scope of SPRUCE, this AO was removed and items with this AO were reexamined. Additionally, the AO *Determine if a single measured value with uncertainty agrees with a distribution of measurements* was re-conceptualized as an AO about removing outliers from a distribution of data and this new form of the AO is not targeted by the given example items.

While Engelhardt recommends items target only a single objective [7], for both the U-STEP and SPRUCE, we often found it impractical to do so, and thus many of the items in both assessments target multiple AOs.

V. USING ASSESSMENT OBJECTIVES TO COMMUNICATE PURPOSE AND SCOPE OF ASSESSMENT

Research-based assessments are designed to be used by instructors and researchers not involved in the assessment development process, and so it is important for developers to communicate what the assessment claims to measure, which is conveniently well-articulated by AOs. For this reason, we posit that a list of AOs could be an effective method of helping potential users of the assessment determine if the assessment is appropriate for their specific context. This is particularly valuable in content areas where there is significant variation in what is covered during a particular course. Laboratory courses and certain upper-division theory courses (e.g., thermal physics or quantum mechanics) are good examples of such courses [17]. There is an increasing desire amongst instructors and researchers to have greater alignment between learning goals and assessments [26], and having a set of AOs that can quickly be compared to a course’s learning goals can provide an efficient and effective means for instructors to ensure that assessments that they employ align with their courses goals.

AOs also offer a potential tool for communicating the results of an assessment to instructors. Where appropriate and valid, assessment results could be presented in such a way that instructors get their student performance broken down across particular AOs rather than, or in addition to, across individual items. Such a breakdown may offer a more clear and actionable summary of student performance that can help guide instructors in making concrete changes to their classroom instruction, which is one of the primary goals of research-based assessments. Additionally, using AOs rather than items to communicate the intent and results of an assessment can improve test security [7].

VI. SYNTHESIS AND CONCLUSION

In this paper, we outlined some of the affordances of articulating *assessment objectives* (AOs) as part of the

process of developing research-based assessments. Articulating AOs can: help us to understand and categorize the domain of interest, provide an additional tool for establishing content validity, operationalize the domain analysis to help in the development of assessment items, and communicate the intent, and structure the results, of the assessment instrument for researchers and instructors. We contextualized these affordances within the development of two research-based assessments, the U-STEP and SPRUCE.

We posit that the articulation of explicit AOs can be a productive step in the development of any research-based assessment. We would recommend that assessment developers articulate AOs during a domain analysis (rather than at a later stage of assessment development) to ensure that the AOs span the domain of interest: one could even use AOs as codes during a qualitative coding process. These AOs can then be refined while establishing preliminary content validity where the AOs are themselves the articulation of content to be validated. The next step, developing individual assessment items to target AOs, can then proceed from a well-informed, well-articulated understanding of the purpose of the assessment. Then, after the development of assessment items using AOs, researchers can use the AOs to communicate the intent and scope of the assessment items to researchers and instructors.

In addition to the benefits described above, AOs have potential applications to the next generation of research-based assessments. Content variation within physics courses presents a consistent challenge for assessment developers who must balance spanning the domain of interest with making sure the assessment is broadly applicable across courses and institutions. One method for addressing content variation is the creation of flexible assessments that can be customized to match the local learning goals. Such flexible assessments might take the form of a test bank of items or modular sub-tests. The use of AOs provide a natural method that would allow instructors to create their customized assessment quickly and efficiently. For example, instructors could review a list of AOs and select those most relevant to their course, which could then be used to automatically generate a version of the assessment targeting these specific topics. Ongoing work in the area of thermal physics aims to test this model for customizable assessment [6, 24].

ACKNOWLEDGEMENTS

This work is supported by NSF DUE 1914840, DUE 2013332, and PHY 1734006, and by the University of Colorado Boulder’s Department of Physics and Center for STEM Learning. We would also like to thank Rachel Henderson, Gayle Geschwind, and the students, instructors, and/or researchers who have contributed to this study.

- [1] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, *The physics teacher* **30**, 141 (1992), publisher: American Association of Physics Teachers.
- [2] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource Letter RBAI-1: Research-Based Assessment Instruments in Physics and Astronomy, *American Journal of Physics* **85**, 245 (2017), publisher: American Association of Physics Teachers.
- [3] C. Walsh, K. N. Quinn, C. Wieman, and N. Holmes, Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking, *Physical Review Physics Education Research* **15**, 010135 (2019), publisher: American Physical Society.
- [4] K. D. Rainey, M. Vignal, and B. R. Wilcox, Designing upper-division thermal physics assessment items informed by faculty perspectives of key content coverage, *Physical Review Physics Education Research* **16**, 020113 (2020).
- [5] S. White Brahmia, A. Olsho, T. I. Smith, A. Boudreaux, P. Eaton, and C. Zimmerman, Physics Inventory of Quantitative Literacy: A tool for assessing mathematical reasoning in introductory physics, *Physical Review Physics Education Research* **17**, 020129 (2021), publisher: American Physical Society.
- [6] A. P. Jambuge, *Research-based assessment design in physics: including scientific practices and feedback for physics faculty*, PhD Thesis (2021).
- [7] P. V. Engelhardt, An Introduction to Classical Test Theory as Applied to Conceptual Multiple-choice Tests (2009).
- [8] T. S. Volkwyn, S. Allie, A. Buffler, and F. Lubben, Impact of a conventional introductory laboratory course on the understanding of measurement, *Physical Review Special Topics - Physics Education Research* **4**, 010108 (2008), publisher: American Physical Society.
- [9] S. Goldhaber, S. Pollock, M. Dubson, P. Beale, K. Perkins, M. Sabella, C. Henderson, and C. Singh, Transforming Upper-Division Quantum Mechanics: Learning Goals and Assessment, in *Physics Education Research Conference 2009* (Ann Arbor (MI), 2009) pp. 145–148.
- [10] M. Eblen-Zayas, The impact of metacognitive activities on student attitudes towards experimental physics (2016) pp. 104–107, iISSN: 2377-2379.
- [11] B. Pollard, A. Werth, R. Hobbs, and H. Lewandowski, Impact of a course transformation on students' reasoning about measurement uncertainty, *Physical Review Physics Education Research* **16**, 020160 (2020), publisher: American Physical Society.
- [12] R. J. Mislevy and M. M. Riconscente, *Evidence-Centered Assessment Design: Layers, Structures, and Terminology*, Tech. Rep. (2005).
- [13] 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).
- [14] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, *International journal of science education* **33**, 1289 (2011), publisher: Taylor & Francis.
- [15] D. R. Dounas-Frazer, L. Ríos, B. Pollard, J. T. Stanley, and H. Lewandowski, Characterizing lab instructors' self-reported learning goals to inform development of an experimental modeling skills assessment, *Physical Review Physics Education Research* **14**, 020118 (2018), publisher: American Physical Society.
- [16] B. Pollard, R. Hobbs, R. Henderson, M. D. Caballero, and H. Lewandowski, Introductory physics lab instructors' perspectives on measurement uncertainty, *Physical Review Physics Education Research* **17**, 010133 (2021), publisher: American Physical Society.
- [17] K. D. Rainey and B. R. Wilcox, Faculty survey on upper-division thermal physics content coverage, in *Physics Education Research Conference Proceedings 2019* (2019).
- [18] E. Dempster, E. Cilliers, K. Johnson, P. Pillay, and S. Wiese, *What's in the CAPS package? Natural Sciences* (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), eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/emip.12253>.
- [20] B. Campbell, F. Lubben, A. Buffler, and A. Sallih, *Teaching Scientific Measurement at University: Understanding Student's Ideas and Laboratory Curriculum Reform* (Southern African Association for Research in Mathematics, Science and Technology Education, 2005).
- [21] L. W. Anderson and D. R. Krathwohl, *A taxonomy for learning, teaching, and assessing : a revision of Bloom's taxonomy of educational objectives* (Longman,, 2001) accepted: 2015-11-06T15:44:27Z Artwork Medium: Ressource physique Interview Medium: Ressource physique.
- [22] D. V. Schroeder, *An introduction to thermal physics* (Pearson, 1999).
- [23] M. Dubson, S. Goldhaber, S. Pollock, K. Perkins, M. Sabella, C. Henderson, and C. Singh, Faculty Disagreement about the Teaching of Quantum Mechanics, in *Physics Education Research Conference 2009* (Ann Arbor (MI), 2009) pp. 137–140.
- [24] 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, iISSN: 2377-2379.
- [25] A. A. Alwan, Misconception of heat and temperature Among physics students, *Procedia - Social and Behavioral Sciences International Conference on Education and Educational Psychology*, **12**, 600 (2011).
- [26] D. T. Brookes and E. Etкина, In search of alignment: Matching learning goals and class assessments (2012) pp. 11–14, iISSN: 1551-7616.