Research-based assessment design in physics: including scientific practices and feedback for physics faculty

by

Amali Priyanka Jambuge

B.Sc., University of Colombo, Sri Lanka, 2014

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Physics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Abstract

Calls to transform introductory college physics courses to include scientific practices require assessments that can measure the extent to which these transformations are effective. Such assessments should be able to measure students' abilities to intertwine important concepts with practices in which scientists engage. In addition to evaluating student outcomes, another related goal of research-based assessments is to evaluate the efficacy of courses. To accomplish this goal, these assessments should have a mechanism to provide faculty concrete suggestions to modify their courses, beyond numerical scores. An approach to achieve this is lacking in the research literature. This motivates us to explore effective ways in which student outcomes can be reported to faculty to facilitate concrete suggestions to modify courses, i.e. actionable feedback.

Physics education research (PER) has a history of developing and disseminating research-based materials to faculty with the intention to improve student learning. However, lack of a consideration of what faculty want in the first place when developing these materials limits faculty to use these materials as developers intended. Even if these materials were adopted, faculty modify these materials to align with their needs and local contexts. There is a recent call to create partnerships with faculty when developing materials for them. In this dissertation, we provide a mechanism to develop assessment tasks that address scientific practices, provide feedback for faculty, and explore features of the external feedback that can be supportive of regular course modifications made by two physics faculty.

To design assessment tasks that can measure students' abilities to intertwine physics concepts with scientific practices, we leveraged Evidence-Centered Design and the Three-Dimensional Learning Assessment Protocol with the focal scientific practice of "Using Mathematics." We conducted video recorded one-on-one think-aloud interviews to explore how students interpreted these tasks. We articulate our design process and the analysis of students'

responses using the ACER (Activation-Construction-Execution-Reflection) framework. Our assessment tasks elicited students' abilities to intertwine concepts with mathematics and written solutions elicited evidence of their abilities to intertwine them most of the time.

We present a mechanism to design actionable feedback for faculty in parallel to developing a new research-based assessment: The Thermal and Statistical Physics Assessment (TaSPA). The feedback design mechanism is rooted in the student outcomes in response to assessment tasks in a coupled, multiple-response format. This assessment task format allows online test administration with streamlined evaluation of student work. We conducted semi-structured interviews with faculty to obtain their perspectives on the developed feedback. Thematic analysis was used to explore the nuance of the faculty perspectives on the generated feedback. We then discuss the process behind incorporating these perspectives into feedback for faculty.

We conducted two case studies of physics faculty to explore the nuance of experiences associated with their course modifications. These explorations can inform identification of features of the researcher-generated feedback that can be supportive of regular course modifications made by faculty. Two case studies of faculty revealed the features – content coverage of a course, time frame for course modifications, and typical enrollment – associated with modifying courses that can be incorporated when designing feedback for them.

This dissertation provides a mechanism to incorporate scientific practices into paper-based assessment design at the introductory college level, an approach to designing explicit feedback for faculty in the context of research-based assessments, and evidence supporting why partnership with faculty when developing research-based feedback for them is important.

Research-based assessment design in physics: including scientific practices and feedback for physics faculty

by

Amali Priyanka Jambuge

B.Sc., University of Colombo, Sri Lanka, 2014

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Physics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
James T. Laverty

Copyright

Abstract

Calls to transform introductory college physics courses to include scientific practices require assessments that can measure the extent to which these transformations are effective. Such assessments should be able to measure students' abilities to intertwine important concepts with practices in which scientists engage. In addition to evaluating student outcomes, another related goal of research-based assessments is to evaluate the efficacy of courses. To accomplish this goal, these assessments should have a mechanism to provide faculty concrete suggestions to modify their courses, beyond numerical scores. An approach to achieve this is lacking in the research literature. This motivates us to explore effective ways in which student outcomes can be reported to faculty to facilitate concrete suggestions to modify courses, i.e. actionable feedback.

Physics education research (PER) has a history of developing and disseminating research-based materials to faculty with the intention to improve student learning. However, lack of a consideration of what faculty want in the first place when developing these materials limits faculty to use these materials as developers intended. Even if these materials were adopted, faculty modify these materials to align with their needs and local contexts. There is a recent call to create partnerships with faculty when developing materials for them. In this dissertation, we provide a mechanism to develop assessment tasks that address scientific practices, provide feedback for faculty, and explore features of the external feedback that can be supportive of regular course modifications made by two physics faculty.

To design assessment tasks that can measure students' abilities to intertwine physics concepts with scientific practices, we leveraged Evidence-Centered Design and the Three-Dimensional Learning Assessment Protocol with the focal scientific practice of "Using Mathematics." We conducted video recorded one-on-one think-aloud interviews to explore how students interpreted these tasks. We articulate our design process and the analysis of students'

responses using the ACER (Activation-Construction-Execution-Reflection) framework. Our assessment tasks elicited students' abilities to intertwine concepts with mathematics and written solutions elicited evidence of their abilities to intertwine them most of the time.

We present a mechanism to design actionable feedback for faculty in parallel to developing a new research-based assessment: The Thermal and Statistical Physics Assessment (TaSPA). The feedback design mechanism is rooted in the student outcomes in response to assessment tasks in a coupled, multiple-response format. This assessment task format allows online test administration with streamlined evaluation of student work. We conducted semi-structured interviews with faculty to obtain their perspectives on the developed feedback. Thematic analysis was used to explore the nuance of the faculty perspectives on the generated feedback. We then discuss the process behind incorporating these perspectives into feedback for faculty.

We conducted two case studies of physics faculty to explore the nuance of experiences associated with their course modifications. These explorations can inform identification of features of the researcher-generated feedback that can be supportive of regular course modifications made by faculty. Two case studies of faculty revealed the features – content coverage of a course, time frame for course modifications, and typical enrollment – associated with modifying courses that can be incorporated when designing feedback for them.

This dissertation provides a mechanism to incorporate scientific practices into paper-based assessment design at the introductory college level, an approach to designing explicit feedback for faculty in the context of research-based assessments, and evidence supporting why partnership with faculty when developing research-based feedback for them is important.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to convey my sincere gratitude to my advisor, Dr. James T. (J.T.) Laverty for his continued support, guidance, and encouragement throughout the process of completing this dissertation. I am appreciative of his patience in providing me time and opportunities to grow as a researcher. I am thankful for him for giving me the research projects that I really enjoyed and excited to work on. I appreciate his attention to graduate students' mental well-being, while they go through graduate school. I am grateful for the opportunity received to work with him.

I am thankful and appreciative of the time and constructive feedback received from of my committee members, Dr. Bethany R. Wilcox, Dr. Glenn Horton-Smith, Dr. Frederick Burrack, and Dr. Jacqueline Spears.

I would like to thank Dr. Eleanor C. Sayre for giving me the opportunity to collaboratively work on the "Equity in Labs" research project. Working on this project has helped me to grow as a researcher. I appreciate her for providing me the opportunity to participate and learn in Professional-Development for Emerging Education Researchers (PEER), and the interview intensive course. I appreciate Dr. Dean Zollman for his valuable feedback and help on my research work.

I am grateful for the opportunity received to work with Dr. Bethany R. Wilcox on the Thermal and Statistical Physics Assessment (TaSPA) research project, and that experience has helped me to grow as a researcher. I am thankful for your continued support and encouragement throughout the time I was working on that project. I appreciate you for encouraging me during the times of both successes and failures.

I would like to thank Dr. Katherine D. Rainey (Katie) for creating an amazing collaborative work environment for the TaSPA project and providing constructive feedback on my work. Thank you for managing student data collection at University of Colorado Boulder,

and making those data available for me promptly. I would like to thank Dr. Paul Bergeron for providing me constructive feedback and helping me with the inter-rater reliability process for the "Using Math" project. I thank Dr. Michael Vignal for providing me with a compiled list of emails of physics faculty in the US, which I directly used to recruit interview participants for the TaSPA research project. Thank you also for helping me with the pilot interview.

I would like to thank my past and current KSUPER colleagues and friends for their feedback and encouragement during my studies. Thank you, Hien Khong, Christopher Hass (Chris), Brandi Lohman, Jessy Changstrom, Amogh Sirnoorkar, Shams El-Adawy, Tyler Garcia, and Bill Bridges for creating a collaborative working environment for everyone in KSUPER. Thank you, Ginny Coghlan and Chris, for helping me with the pilot interview for the "Using Math" project. Thank you, Lydia Bender for being a lovely colleague and friend, and helping me with recruiting student volunteers for the "Using Math" project. Thank you Katherine C. (K.C.) Ventura for helping me with recruiting student volunteers for the "Using Math" project. Thank you K.C. for designing, administering, and recording student responses to two assessment tasks that I adopted for my analysis. I would like to thank Dr. Bahar Modir for her kindness in sharing the interview tips when I needed them the most, and helping in finding video recording equipment.

I am grateful for the students who volunteered for my projects. I am thankful for the instructors in the classes in which we recruited students for interviews, for allowing us to make announcements about recruiting students. I am grateful for the faculty members who voluntarily participated in the interviews.

Thank you Kim Coy for your kind gestures towards all the graduate students. Thank you for being there for us during both ups and downs. Thank you for creating the Department of Physics at Kansas State University, a lovely and welcoming place to work with. Thank you also for managing the gift cards for the interviews of the "Using Math" project. Thank you, Peter Nelson, for willing to provide me with video recording equipment, upon request.

I am thankful for the travel scholarships received from the Graduate School and College of Arts and Science at Kansas State University, which gave me the opportunity to present

and receive feedback on my work. I am thankful for the professional development activities organized by the Graduate School at Kansas State University, which helped me grow as a professional. I am thankful for the writing center at Kansas State University for providing me help on academic writing. I thank Ruth Newton, and Boglarka Davies at International Student and Scholar Services (ISSS) at Kansas State University for helping me with paperwork relating to maintaining visa status.

I thank the Sri Lankan Students' Association and Powercat Masters Toastmasters organization at Kansas State University for providing me opportunities to improve my communication and leadership skills. Thank you, Shed Mayberry and Maslyn Prosper-Mayberry for mentoring me with speech preparations for meetings, and it has helped me improve my academic presentations as well.

I am grateful for the funding provided on the work presented on this dissertation by Department of Physics at Kansas State University and National Science Foundation (Grant Numbers: 1726360, 2013332, and 2013339).

I am grateful for my friends in Manhattan, Kansas for welcoming and making me feel I am at home. Last but not the least, I am grateful for my family, the pillar of strength for everything I do in my life.

This dissertation would not have been possible without all of you.

Dedication

To my beloved parents for their unconditional love, encouragement, and guidance to pursue a career of my interest.

To the government of Sri Lanka for offering free education for everyone, which never let me stop dreaming big.

To my beloved teacher, Mrs. Korala, for her commitment in teaching me English language, which I never thought would help me reach this far.

To my beloved brother and sister for looking up to me, which never let me give up easily.

To my beloved husband for his unconditional love, and patience.

Chapter 1

Introduction

There are recent calls to include scientific practices into college classrooms that underscore the importance of bringing student knowledge closer to its usage[7–11]. This is in part to expose college students to the same learning environment as they have been exposed at K-12, where not just knowledge but application of knowledge is also emphasized[8]. As Cooper *et al.*[8] mentions,

> *"It would be a disservice to throw these students back into typical introductory courses. . . "*

Scientific practices constitute generalizable actions that scientists engage-in on a daily basis (such as develop and use models, analyze and interpret data, use mathematics, and plan and carry out investigation). Intertwining these practices with concepts core to physics (Core Ideas) promotes deeper learning[12;13].

Incorporating scientific practices into college courses and evaluating the extent this transformation is effective requires assessments that have the ability to measure not just what students know, but how they use and apply their knowledge to new situations[14]. As Cooper[15] mentions,

> *"If we know what we are looking for, it is easier to recognize and assess it when we see it."*

The K-12 framework for science education[12] well describes the scientific practices, allowing us to assess these practices[15]. However, developing assessments that address scientific practices is identified as an arduous and time consuming process[8;16–18]. Thus, how can we assess students' work products to reason about their abilities to engage in scientific practices along with concepts?

In Chapter 3, we articulate a process for developing assessment tasks by focusing on the scientific practice of "Using Mathematics" and the concept of "force." Our focus here is to assess how students use mathematics (hereinafter, math) to do physics rather than just math. We use data from interviews of students solving a paper-based exam that address the scientific practice of "Using Math," while simulating a summative assessment environment. The interview participants are introductory-level students who were not specifically instructed with learning goals associated with the scientific practice of "Using Math."

We build on work by Harris *et al.*, and Stephenson *et al.* to design assessment tasks that address students' use of math by leveraging principles of Evidence-Centered Design and to validate them for their potential to elicit expected evidence[3;19–24]. The existing work on designing assessment tasks to assess scientific practices using Evidence-Centered Design covers middle school science students[3;16;17;19;25;26], introductory-level chemistry students[21], and upper-division physics students[27;28]. We fill the gap in the literature by introducing a theory-driven methodology adopting Evidence-Centered Design to assess students' use of math in physics paper-based assessments at introductory-level. Having a systematic, theory-driven approach to assess students' ability to engage in the scientific practice of "Using Math" would facilitate extending our understanding of assessing students' ability to engage in the rest of the scientific practices as well.

In addition to assessing student learning, another related utility of research-based assessments is to provide information to faculty about the extent to which students achieve intended performance thresholds. Current research-based assessments provide insights to faculty about student learning through numerical scores (e.g., pre- and post-test percentage of scores provided by the force concept inventory[29]). Numerical scores have been a useful and productive source of information for faculty when aggregated over years. The numeri-

cal scores aggregated among similar courses or unique courses over time enabled faculty to make informed decisions[30–32]. Faculty member's decision to implement more active learning environments for students than traditional, lecture-based learning environments was one such example. This implementation was based upon the aggregated data of students' scores to research-based assessments among similar courses, that showed increased students' scores when they were exposed to active learning environments in comparison to traditional, lecture-based learning environments. However, aggregating students' scores over time is arduous and time consuming, thus limiting immediate actions that can be taken by an individual faculty to inform course modifications.

When an individual faculty member conducted available research-based assessments in their classroom, it was not always clear to them what students' scores in response to that assessments communicated about their instruction[33]. Madsen *et al.*[33] conducted faculty interviews to obtain perspectives of faculty about the affordances and limitations of available research-based assessments. One of their key findings is:

> *"Faculty also want a deeper understanding of what the results mean, for example, a better understanding of what a specific numerical score tells them about their teaching."*

To illustrate this further, assume Dr. X conducted a standardized assessment in their classroom and received an average pre-test percentage score of 76% with 6% standard deviation. Assume, they also conducted the same assessment and obtained an average post-test percentage score of 89% with 5% standard deviation. What information do these increased average scores communicate to Dr. X about their instruction? How does Dr. X know what worked well for students and what needs to be modified in their instruction to better facilitate student learning?

To address this need raised by faculty, we need a better approach to communicate with them about the information relating to student learning, beyond providing scores. In Chapter 4, we provide a methodology to translate evidence of student learning rooted in a new research-based assessment under development (more information on this new assessment will

be provided later), to information that can guide faculty towards explicit course modifications. We simply refer to this as providing a methodology to develop "actionable feedback" for faculty.

"Feedback" is the translated information that will be provided to faculty based on the evidence of student learning. We add the term "actionable" before the term "feedback", simply because the feedback should be able to operationalize instruction, rather than just being mere information to faculty.

Introducing actionable feedback (hereinafter simply "feedback") to research-based assessments is a novel approach. In Chapter 4, we inspect the impact the addition of actionable feedback makes on a research-based assessment. We design feedback and conduct interviews with faculty to explore how they react to this novel approach and provide perspectives on the designed feedback. We explain the process behind modifying the generated feedback based on the perspectives faculty bring during the interviews.

In addition to recruiting perspectives on the generated feedback, feedback should be developed such that it can be supportive of individual faculty. Goertzen, Scherr, and Elby[34] conducted a case study of Alan, a teaching assistant (TA) for tutorials. The feedback Alan received from his fellow TA instructors in a professional development setting had a little impact on him. This was due to the lack of support the provided feedback allowed Alan to re-evaluate beliefs that govern his teaching practices. Alan's case study provided deeper insights into both his teaching practices and beliefs that governed these teaching practices. Thus, Goertzen, Scherr, and Elby[34] emphasized providing feedback not only on how Alan taught, but the beliefs that made him teach in a certain way. The following quote provides evidence to support these authors' stance on providing feedback to Alan, which is centered around his beliefs beyond just the practices.

> *"Thus, feedback given to Alan needs to respond not only to behavior like his tendency to assume students understand when they provide the correct conceptual answer but also to respond to his belief that instructors should give students the benefit of the doubt rather than assume they are incorrect."*

Similar to Alan, physics faculty also have beliefs associated with their classroom practices. Thus, to make an impact, feedback for physics faculty should also be centered around their beliefs, but not just the practices. We do not limit ourselves to just exploring beliefs, but "processes internal to faculty" including beliefs, that can govern their practices associated with "modifying courses." We refer to "feedback" for faculty which is rooted in research-based assessments developed by physics education research (PER) practitioners.

In Chapter 5, we provide two case studies of physics faculty. We do not intend to provide concrete suggestions on either features that can guide effective feedback development for a particular assessment or modifying existing feedback structure of a particular assessment. Instead, we call for an approach to designing feedback for physics faculty that leverages both their practices and processes internal to them.

We explore how faculty themselves view the process of course modifications by attending to the "processes internal to them," when modifying courses. We explore the types of external feedback faculty typically receive and how that feedback is influential to the "processes internal to them." This exploration would guide us to articulate the nuances the external feedback should entail, supporting the modifications faculty typically execute.

In the next chapter (Chapter 2), we provide a literature review related to the work presented in this dissertation. In Chapter 3, we provide a mechanism to develop assessment tasks that can assess students' abilities to blend physics concepts with scientific practices. We provide an approach to develop feedback for faculty in Chapter 4. In Chapter 5, we provide two case studies that provide us the evidence why considering experiences associated with modifying courses by each faculty member is important when providing external feedback to them. In the last chapter of this dissertation (Chapter 6), we provide a summary of the work presented in this dissertation along with future work.

Chapter 2

Background

In this chapter, we provide background on Assessments in PER (Sec. 2.1), Assessment Design and validation (Sec. 2.2), Problem-Solving and "Using Math" in Physics (Sec. 2.3), Feedback for Faculty in General (Sec. 2.4), Research-Based Assessment Feedback for Faculty (Sec. 2.5), Research-Based Material Dissemination (Sec. 2.6), and Research-Based Assessment Dissemination (Sec. 2.7).

2.1 Assessments in PER

Assessments can be broadly viewed as either formative or summative. Formative assessments are used on a daily basis to support student learning by giving students the feedback needed to reflect on their own learning and to adjust the subsequent instruction of the instructor. On the other hand, summative assessments are used to provide evidence of achievement to make decisions such as grading and retention[35;36]. The available and widely used standardized assessments (such as concept inventories) in PER typically are used for summative purposes[37]. As of now, there are almost 100 research-based assessments available for the physics education community as listed in the PhysPort website[38]. These assessments are identified by PhysPort as assessing content knowledge, problem-solving, scientific reasoning, lab skills, beliefs/attitudes, and interactive teaching.

These standardized assessments primarily measure students' conceptual knowledge (63 out of 95 assessments measure content knowledge) in numerous physics concepts[38]. Thus, these off-the-shelf assessments have a significant impact on education reform by providing a universal way of evaluating student understanding that leads teachers to assess and revise their teaching methods[39]. For example, these assessments have been used to evaluate teaching methods[30–32], learning outcomes of different student populations[40–42], and curriculum reforms[43;44].

The most common standardized assessments used at introductory level are Force Concept Inventory[29], Force and Motion Conceptual Evaluation[45], Brief Electricity and Magnetism Assessment[46], and Conceptual Survey of Electricity and Magnetism[47]. While these concept inventories are assets in eliciting students' conceptual understanding, they are not designed to elicit students' engagement in scientific practices[37]. However, calls to include scientific practices into K-12 level and college curricula brought assessment developers' attention to design tasks to assess students' abilities to engage in scientific practices and concepts. For example, Wolf *et al.*[48] developed and validated a practical exam to assess student abilities to engage in scientific practices in introductory physics laboratories. While this work provides a promising way to assess scientific practices in laboratory settings, it is unclear how this approach can be generalizable to typical large-scale introductory classrooms where paper-based assessments play a prominent role.

2.2 Assessment Design and Validation

Assessments give us vital information about student learning. The "information" refers to the types of inferences we make out of students' work, attributing a certain set of knowledge and skills to the student performance that align with the designer's goal for the assessment. The process of making inferences is referred to as "reasoning from evidence" that describes the process of drawing inferences accumulating a set of supporting evidence from students' work[36].

This process can also be portrayed as a triangle where the triad represents the three key

elements highlighted in designing assessments, the *assessment triangle*, a model of student cognition and learning in the domain, a set of beliefs about the kinds of observations that will provide evidence of student competencies, and an interpretation process for making sense of the evidence[36]. The assessment triangle conceptualizes the nature of assessment tasks, but an elaborative framework is needed to operationalize those conceptualizations.

Evidence-Centered Design is embedded in the logic of the assessment triangle. It provides a methodological and systematic approach to the assessment task design that helps elicit students' proficiencies attributed to the designer's intention. It has also been identified as a promising approach for developing assessment tasks that effectively measure concepts intertwined with scientific practices[35;36]. In particular, there are several works, in which researchers have adopted Evidence-Centered Design to design assessment tasks that assess scientific practices and concepts[3;16;17;19;21;25;26].

Assessment task design is accompanied by validating the designed tasks. There are several approaches to task validation in the research literature. One approach takes the form of content validity where the alignment between the task content with the subject matter framework is evaluated by experts in a particular domain[36]. Extending this approach to include empirical evidence to determine the extent to which designed tasks tap the intended cognitive processes is also emphasized in several works[36;49]. This argument-based approach to validity consists of two parts: interpretive and validity arguments[50;51]. First, the interpretation and use of assessment scores are proposed prior to disseminating the assessment tasks to students (interpretive-argument). Second, the plausibility of the interpretive-argument is validated via student think-aloud interviews (validity-argument)[35;36;52]. During this process of validation, students' unintended problem-solving approaches that tap unintended cognitive processes differing from the designer's intention can be documented. Thereby, the iterative modifications to task design can be made until the proposed interpretations and use of assessment scores are reasonable.

Evidence-Centered Design, in particular, leverages the argument-based validity approach to validate assessment tasks where the claims about student knowledge and skills are backed by evidence[53;54]. We acknowledge that there are numerous approaches for task validation in

the research literature other than the approaches we described in this section (e.g., criterion validity[55], classical test theory[56], and item response theory[56]). Such statistical approaches for task validation are not the focus at this stage of the study, but will be the focus in the future. Thus, we do not provide an extensive literature review on that.

2.3 Problem-Solving and "Using Math" in Physics

Mathematics is one of the cornerstones in physics problem-solving. However, use of mathematics in physics is found to be different from mathematics alone[57–61]. This nuance often causes problems due to the gap between student and instructor expectations of what it means to do math in physics.

Physicists believe use of math in physics occurs in a certain, prescribed way[57;58]. Thus, one way to evaluate students' use of math is to probe their work products produced during problem-solving with the prescribed models for using math in physics. For example, Redish[57] developed a model describing the bare bones of how to use math in physics. This model includes 1) Mapping the physical system into a mathematical model, 2) Processing the mathematical model to simplify it, 3) Interpreting the results obtained to explore what they tell about the physical system, and 4) Evaluating the result to validate its extent to accurately represent the physical system.

However, it is worth noting that students do not necessarily follow that procedure when solving problems[57]. Instead, they approach problems in ways different from Redish's prescribed procedure. These approaches are typically considered as ineffective[62]. These ineffective ways might arise due to the lack of a systematic strategy that guides students to apply their knowledge. Thus, lots of research work has targeted teaching students specific problem-solving strategies[62–64] and to evaluate students' engagement in problem-solving[64–70].

Another approach to evaluate students' use of mathematics is to explore how students use math on their own terms. The body of research work on this aspect leverages theoretical perspectives such as resources[71], framing[72], and epistemic games[73] to explore students' in-the-moment reasoning while solving physics problems[74]. Resources are the fragments of

knowledge being activated based on how students tacitly determine what kind of knowledge might be appropriate for the problem at hand (framing). This leads to a set of locally coherent activities (moves) students do during problem-solving (epistemic games).

One such study that leverages the aforementioned theoretical perspectives is the work by Tuminaro and Redish[75] where they observed six epistemic games introductory students play while solving physics problems. The tacit judgement students make to decide which game to play depends on their expectations for the problem at hand. These expectations determine which resources to bring into a particular problem context. Bing and Redish[76] leveraged resources and epistemological framing to capture how upper-level students use math. A recent study by Modir, Thompson, and Sayre[77] developed a theoretical framework that models upper-level student framing in math and physics adapting epistemological framing.

While theoretically well-grounded approaches are more robust than prescribed models to explore student use of math in physics, they pose challenges on instructors who are not familiar with these theoretical constructs. Attending to these challenges, the ACER framework[74;78] bridges the gap between prescribed models for student use of math with the resources framework and epistemic frames.

ACER stands for the Activation of the tool, Construction of the model, Execution of the mathematics, and Reflection of the results[78]. These components are pertaining to the activation of the mathematical tool, mapping between the physics and mathematics of a problem, working with the procedural aspects of the mathematical tools, and interpreting and checking the intermediate and the final steps of the solutions respectively. Each of these components consist of several subcodes in which students shift back and forth while solving physics problems. These subcodes are not categorized in any specific order, rather it describes what steps students might take while going through the problem. For example, in construction of the model component, students might be making assumptions or developing a representation that describes the physical system. The subcodes under the components rely on the nature of the assessment tasks.

In addition to exploring students' problem solving approaches, communicating to faculty the approaches students took to solve problems is also important. Faculty can then build

upon this information to modify their courses to better facilitate student learning.

2.4 Feedback for Faculty in General

Feedback in general is viewed as the information one receives about their performance in relation to an expected performance threshold, which helps guide one's learning[79]. Feedback can be either internal or external. Internal feedback is the feedback one generates on their behalf based on their perception about the performance. On the other hand, external feedback is received from external sources (such as an audience member) on one's performance as compared to external standards about that performance[2]. We now turn into related background on external feedback to faculty, which is the focus of our study.

External feedback for faculty has been identified as an important element which can be supportive of faculty professional development[80;81], though there is a lack of consensus around how to facilitate it[82]. The common and well-known approaches to providing feedback to faculty include student and peer evaluations on instructional practices.

Feedback for faculty in the form of student evaluations (such as mid or end semester course evaluation forms) can enhance teaching[83;84]. Student evaluations mostly focus on teacher-centered practices, such as the teacher preparedness or format of the classroom[84–86]. Through student evaluations, faculty often receive limited information about students' content knowledge, a crucial element that can contribute to designing feedback for faculty[87;88]. Thus, student evaluations provide minimal information about student learning which can inform subsequent content modifications by a faculty member to better facilitate student learning[89–91].

Faculty also often receive evaluations from their peers, after peer faculty observe their real-time teaching[80]. Peer evaluation has been identified as supportive of faculty professional development[81]. However, similar to student evaluations, peer evaluations also carry some limitations. One such limitation is that peers might not have self-confidence in their opinions when evaluating their fellow faculty members[92]. There are also concerns about the biased evaluation that can be made by a peer faculty member, and therefore to have a more balanced

perspective, evaluations from several peers are needed[93]. Faculty themselves resist peer evaluation due to the lack of its contribution to career advancement, when it is compared to the contribution from student evaluations[94].

Centra stated that,

> *"Emphasizing learning rather than teaching (that is, what the instructor does) has recently been promoted as the preferred paradigm."*

We situate our work in the same paradigm, where student learning is emphasized rather than evaluating teaching practices of faculty. Thus, we use students' responses to research-based assessments to capture student learning.

2.5 Research-Based Assessment Feedback for Faculty

Research-based assessments in physics have been used to characterize student learning under different learning environments[52;95]. Characterizing student learning through these assessments provides opportunities for faculty to implement instructional interventions that can better facilitate student learning[39].

Historically, research-based assessments in physics have been used to evaluate different teaching methods[30–32], performance of diverse student population[40–42], and efficacy of curriculum reforms[43;44]. These large-scale studies help researchers to identify learning environments that can facilitate better student learning. A potential "modification" made to a course by a faculty member informed by these studies has been considered as a *change* occurred to their instructional practices[96;97].

One of the key strategies identified as successful in creating instructional change is to encourage faculty to reflect on their instructional practices and improve[96;97]. An important feature of this strategy is the consideration of faculty as individuals with knowledge and experiences to improve themselves[98]. One approach to encourage and support faculty to reflect on their instructional practices has been identified as the "external feedback to faculty" [99;100], which can be helpful for instructional transformations made by faculty[101;102].

The form of external feedback that is widely available to an individual faculty member is simply the numerical scores rooted in research-based assessments (e.g., students' conceptual gains in the form of pre- and post-test scores[29;45–47], and attitudinal changes as pre- and post-test measures in Likert scales[103;104]). As we explained in Chapter. 1, this form of external feedback rooted in current research-based assessments limits opportunities for an *individual* faculty member to reflect and make informed decisions of their teaching[33], though this is not the case when results from a large scale study is available to a faculty member (e.g., Hake's study[30]).

2.6 Research-Based Material Dissemination

There is an extensive time, effort, and resources that PER researchers put into developing research-based materials in recent decades[105]. One major goal of the developers of these materials is to communicate up-to-date PER findings to faculty with the intention to improve student learning upon faculty's use of these materials in their classrooms. These goals of the developers were achieved to some extent, resulting in improved student learning[106]. However, there are areas that yet need improvement.

Henderson and Dancy introduce four categories that lay out the connection between the developers of research-based materials and intended users of these materials. These four categories are not discrete categories, rather they are situated in the adoption-invention continuum (see Fig. 2.1). These four categories include adoption, adaptation, reinvention, and invention[107].

Adoption Developers disseminate curricular materials to faculty and expect that they use them with fidelity.

Adaptation Developers disseminate curricular materials to faculty and expect that they make slight changes to the materials.

Reinvention Developers disseminate curricular materials to faculty and faculty make significant changes to the materials or create new ideas based on these materials.

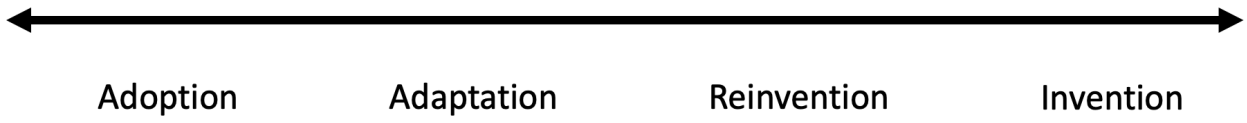

Figure 2.1: *The adoption-invention continuum. This figure is recreated from its original version presented on Henderson and Dancy*[1].

Invention Faculty themselves develop curricular materials using their ideas.

Typically, most of the STEM-related, research-based materials are designed with the mindset to disseminate them to faculty and expect that they use these materials with fidelity – adoption[108;109]. However, research shows that faculty make significant changes to these materials when they use them in their classrooms[108;110–112]. One of the reasons for this is the lack of involvement of faculty during the research-based material development process. Instead, these materials were considered as both "instructor-proof" and "context-proof" (materials independent from instructor and contextual effects such as the personality of faculty and departmental norms, respectively)[97;112].

Research shows that there are individual and contextual features that can hinder faculty using these materials in their classrooms with fidelity[106;107;107;112]. Dancy, Henderson, and Turpen stated that it is less likely that research-based materials can be implemented in a classroom with fidelity, but can undergo changes during faculty's implementation of them to adjust them to their unique departmental contexts, student's perspectives, and personality of the faculty[108].

Instead of considering faculty as partners during the process of research-based material design, there is an implicit assumption inherent behind designing and disseminating research-based materials to faculty. That is the consideration of faculty as individuals who can make informed decisions on implementing these materials into their classrooms with fidelity[108]. For example, consider Roger's innovation-decision process, which includes five stages from an intended user gaining knowledge about research-based materials to confirming the continued use of them in their classroom[113].

Knowledge An individual is exposed to the existence of the research-based materials and gains knowledge about how these materials function.

Persuasion An individual creates a favorable or unfavorable attitude towards research-based materials.

Decision An individual makes a decision on whether or not to adopt research-based materials.

Implementation An individual puts the research-based materials into practice.

Confirmation An individual seeks to reinforce the decision to use the research-based materials.

Developers of research-based materials are successful at making faculty gain knowledge of these materials, particularly through talks, workshops, and journal publications[108;114]. Showing evidence of improved student learning in classrooms when faculty use these materials, intended users are also successfully persuaded to implement these materials into their classrooms. However, $\frac{1}{3}$ of the faculty who used research-based materials discontinue to use them over time. One of the reasons for this discontinuation is that the faculty do not necessarily follow what the developers of these materials suggest as indicated in the "adoption" category. Instead, faculty follow practices laid out similar to "reinvention" or "invention" categories[114].

Research also suggests that it is time to move away from viewing faculty from a deficit point of view[115;116]. This strand of research highlights the similarity between viewing students and faculty from a deficit point of view. In a deficit point of view, the difficulties students have in a classroom were considered due to the deficiencies within individual students, but not due to the education system or society. Similarly, from a deficit point of view, faculty are considered as individuals who do not motivate to improve, rather resisting to change their teaching practices[117].

There is another strand of research, which advocates for considering faculty as resourceful individuals, and thus promoting asset based point of view[98]. Thus, this strand of research calls to include faculty during the process of developing research-based materials. That way, developers of research-based materials respect and value both faculty's own classroom practices (such as own beliefs about teaching) and contextual factors (such as departmen-

tal norms) that can be influential when adopting new materials into their classrooms. As Henderson and Dancy[1] stated:

> *"When disseminating educational innovations, the research community should focus on working with faculty as partners, either individually or in small groups, to improve instructional practices in individual situations. Under this framework, faculty would be recognized as a valuable part of this process with learning occurring on both sides. This is in contrast to current dissemination activities describing deficiencies with traditional instructional practices, providing polished ready-to use curricula, and having change agents promote only the curricula that they developed."*

2.7 Research-Based Assessment Dissemination

We now turn to the feedback for faculty rooted in research-based assessments. One such feedback faculty often receive is numerical scores, which assessment developers (similar to other research-based material developers) believe as helpful for faculty to modify their subsequent instruction to better facilitate student learning. However, faculty indicate need of better ways to interpret the scores that they receive after conducting research-based assessments, which can productively communicate to them the effectiveness of their teaching[33]. Madsen *et al.* stated:

> *"Faculty also want a deeper understanding of what the results mean, for example, a better understanding of what a specific numerical score tells them about their teaching."*

We take the example above to note that there is a mismatch between what assessment developers think what types of feedback work well for faculty and what faculty themselves believe works well for them. Bringing perspectives from Alan's case study that we introduced in Chapter 1 into the discussion again, we note that this mismatch arises due to the lack of

consideration of what course modifications mean for faculty, instead only focusing on what we believe them doing as course modifications, when providing feedback to faculty which is rooted in research-based assessments.

Informed by the background provided in this chapter, we first provide a methodology to design assessment tasks that can assess students' abilities to blend the scientific practice of "Using Math" with physics concepts, in the following chapter.

Chapter 3

Assessing Scientific Practices in Physics Paper-Based Assessments

In this chapter, we answer the following research questions:

1. How do we develop assessment tasks to assess students' use of mathematics along with physics concepts?
2. How can we validate students' work products in response to these tasks for their potential to elicit students' abilities to intertwine mathematics with concepts?
3. How much evidence of their abilities to intertwine mathematics with concepts do we get from looking at the written responses?

In Sec. 3.1, we provide theoretical background for our task design and validation process followed by research questions in Sec. 3.2. In Sec. 3.3, we explain our methodology for task design and the analysis of student responses to the designed tasks followed by data analysis exemplars. We finally provide some insights into our results suggesting potential implications for assessment design and validation in Sec. 3.4.

3.1 Theoretical Background

In this section, we articulate the theoretical approach to our task design process adapting Evidence-Centered Design and the Three-Dimensional Learning Assessment Protocol[7] along with our theoretical assumptions for task validation. We first articulate the general principles of Evidence-Centered Design as laid out by its developers and then how researchers adapt that to incorporate scientific practices. We then explain the utility of the Three-Dimensional Learning Assessment Protocol into our work. We also provide insights into our task validation approach within the Evidence-Centered Design.

3.1.1 Evidence-Centered Design

Employing educational assessments can be viewed as a process of reasoning from evidence, i.e. how we can use assessments to infer what students know and can do[36]. However, designing assessments to measure these constructs requires careful and thoughtful approaches. As Mislevy[20] mentions,

> *"Assessment design is often identified with the nuts and bolts of authoring tasks. However, it is more fruitful to view the process as first crafting an assessment argument, then embodying it in the machinery of tasks..."*

This way the distinction between testing and assessment is emphasized.

Drawing from previous work, beyond this point, we explain the basics behind Evidence-Centered Design (ECD)[20;22–24]. ECD suggests that we first gather substantial information of the domain of interest (such as physics). This substantial information includes, but is not limited to, concepts, student knowledge representations, and terminologies. Then the information gathered can be depicted into a design pattern.

Design pattern comprises several elements to ensure coherent nature between the claim about what students should know and be able to do (Student Model), expected evidence to meet the claim (Evidence Model), and the task to elicit the evidence (Task Model). The Student Model articulates the knowledge, skills, and abilities identified as important. The

evidence for these knowledge, skills, and abilities are required to justify the claim about what students should know and be able to do. The Evidence Model articulates the potential observations in the student work that constitute evidence for knowledge, skills, and abilities. The "Task Model" makes sure that the task features have the potential to elicit potential observations in students' work.

After laying out the basics of ECD, we now turn to work that utilizes ECD as a design approach to design assessment tasks that assess scientific practices articulated in the framework for K-12 science education[12]. The theoretical views below mostly capture the ideas in Harris *et al.*[3], and we suggest this reference for readers who are interested in the detailed assessment task design approach laid out here.

Our assessment task design approach, which is mostly reflected the approach by Harris *et al.*[3] is also built around the three models, i.e. student, evidence, and task models. The Student Model, claims about what students should know and be able to do takes the form of learning performances. Learning performances articulate assessable statements that measure student abilities to intertwine scientific practices with concepts. The knowledge, skills, and abilities required to meet the learning performances are also articulated in the Student Model. Evidence Model consists of evidence statements that provide evidence that students have the required knowledge, skills, and abilities. The Task Model makes sure that the assessment tasks have the potential to elicit the evidence statements.

As we develop tasks for the introductory-level physics students, it is worthwhile to explore the valued scientific practices and the ways those can be elicited in assessment tasks in introductory level. Thus, we next bring your attention to the Three-Dimensional Learning Assessment Protocol (3D-LAP)[7], a tool that can be used to design assessment tasks to elicit student abilities to engage in scientific practices. The 3D-LAP consists of criteria each for scientific practice, and to align with a scientific practice, all of the underlined criteria should be met. This criteria was developed with a team of disciplinary experts that consisted of researchers in the field of education-based research and more traditional faculty members. The 3D-LAP was successfully validated for its reliability to differentiate tasks that have the potential to elicit scientific practices and concepts with the tasks that do not have the

potential to do so[7].

To have a coherent task design, we couple the 3D-LAP with the ECD. In other words, the criteria in the 3D-LAP for tasks to elicit scientific practices can be used as task features in the Task Model in ECD which we explain with more details in the Sec. 3.3.

3.1.2 Assessment Task Validation

Assessment task design is followed by the validation of those tasks[3;21]. Adapting the 3D-LAP, a tool that has been validated for its reliability to differentiate assessment tasks with and without having potential to elicit scientific practices along with concepts, contributes to our tasks' content validity[7]. The assessment task validation also ensures the extent students demonstrate the evidence that the tasks intended them to be showcased. One way to evaluate such validity is to examine the processes students go through when they encounter these tasks and look for evidence to determine that the task functions as intended. In this way, the assessment tasks can be connected with the students' ideas[39]. In particular, the student solutions should be explored in light of evidentiary arguments to determine the extent to which assessment tasks have the potential to elicit appropriate predefined evidence. Think-Aloud[118;119] interviews have been suggested as a way of eliciting student problem-solving processes to the assessment tasks[35;36;52].

The task validation process requires us to allow descriptive, unexpected student evidentiary representations to take into consideration. In other words, the predefined evidence statements that give us the evidence that students have targeted knowledge, skills, and abilities can be modified based on student solutions to entirely capture their potential to elicit the learning performance. As part of these modifications, students' fine-grained evidentiary representations that pertain to the evidence statements can emerge. Thus, an analytic framework that closely captures the predefined evidence statements can be adapted to interpret student work products.

We expand our ESs, and thereby the "Evidence Model" by coupling with an analytic framework that closely captures the predefined ESs which is the ACER framework in our case.

We adopted the ACER framework because 1) the component in the framework well-aligned with our predefined ESs for assessment tasks, 2) it gives insight into learning theories[74] while remaining open for instructors who are not familiar with the theoretical constructs of the framework, and 3) its emphasis on organizing students' written work products (as compared to only video data). This approach modifies the "Evidence Model" by introducing student knowledge representations based on their own terms.

Table 3.1: *Stages included from the ECD process for the task in Fig. 3.1. Task features are the criteria in the 3D-LAP to design constructed response assessment tasks to elicit the learning performance associated with the scientific practice of "Using Math" intertwined with "force." The * represents the elements added as part of the task's validation process based on student responses.*

Learning Performance	Students will be able to use math to determine kinematic values from data about the motion presented and use that information to reach a conclusion about the nature of the motion.
Knowledge, Skills, and Abilities	KSA1: Identify kinematics principles as appropriate to determine the nature of the motion. KSA2: Identify relevant physics equations or generate mathematical equations to connect the variables in the physical system. KSA3*: Conduct appropriate mathematical manipulations. KSA4: Determine the nature of the motion.
Evidence Statements	ES1: Statements of the unpacking of appropriate physics concepts to solve the problem. ES2: Statements of the use of mathematical equations that represent the given physical system. ES3*: Statements correspond to mathematical manipulations. ES4: Statements interpreting the results from the mathematical manipulations.
Task Features	Question gives an event, observation, or phenomenon. Question asks students to perform a calculation or statistical test, generate a mathematical representation, or demonstrate a relationship between parameters. Question asks students to give a consequence or an interpretation (not a restatement) in words, diagrams, symbols, or graphs of their results in the context of the given event, observation, or phenomenon.

3.2 Research Questions

Our research questions articulated in Sec. 1 turned in to a form below after incorporating the theoretical perspectives we lay out in Sec. 3.1. Thus, in this work, we answer the research questions,

1. How do we develop assessment tasks to assess the extent to which students achieve learning performances that intertwine scientific practices and concepts?

2. How can we validate student work products in response to these tasks for their potential to elicit expected evidence to achieve the target learning performances?, and

3. How much evidence of their abilities to meet the learning performances do we get from looking at the students' written responses?

3.3 Methodology

As we move forward on this section, we explicate our assessment task design process, data collection, and data analysis to answer our research questions in Sec. 3.2. The presented methodology in this section does not reflect the exact process we followed during our research. We modified and optimized the process based on our research experience.

3.3.1 Assessment Task Design

Harris *et al.*[3] articulated their task design process adapting ECD along with multiple design stages to ensure coherent task design to intertwine concepts with practices. We build on that work to design assessment tasks in the context of undergraduate physics, specifically introductory mechanics. Table 3.1 summarizes the stages in the ECD process (described below) used to develop the task shown in Fig. 3.1.

We first need to identify what we value that students should know and be able to do in the domain of physics. We then construct an assessable statement that blends what students should know (concept) and be able to do with their knowledge (scientific practice)

Assume you are responsible to carry out an accident reconstruction case at your local police station. The car accident left a skid mark of length 40.3 m on the road. The driver claims he was driving under the speed limit.

In order to further clarify this case, you did an experiment at a crash site with similar accident conditions. The data shows an average skid mark of length 22.4 m when the brake was locked while the car was travelling at the speed of 15.2 m/s.

Describe how you can determine the speed of the car before the accident.

Your job is to determine whether or not the driver was speeding before the car accident. If the speed limit of the area that the accident occurred in is 18 m/s, is the driver at fault?

Figure 3.1: *Car accident reconstruction problem from the assessment.*

in the form of a Learning Performance (LP). We then determine the Knowledge, Skills, and Abilities (KSAs) to achieve that LP. Then we articulate the Evidence Statements (ESs), which specify what we need to see in a student's response to demonstrate that they have the KSAs we articulated previously. In the final stage, we define the task features needed to elicit the evidence articulated in the ESs.

As we stated in Sec. 3.1, the criteria in the 3D-LAP lays out the basis for the task features to elicit the ESs[7]. The protocol consists of a set of criteria for each scientific practice where all the specifications of the criteria should be satisfied in order for an assessment task to have the potential to elicit a scientific practice. For example, to elicit the scientific practice of using mathematics, we should develop the task such that it 1) gives an event, observation, or phenomenon 2) asks students to perform a calculation or statistical test, generate a mathematical representation, or demonstrate a relationship between parameters, and 3) asks student to give a consequence or an interpretation (not a restatement) in words, diagrams, symbols, or graphs of their results in the context of the given event, observation, or phenomenon. The phenomenon can be integrated with concepts around core ideas in physics (in our case, force) to take the form of task features to elicit a LP that addresses the scientific practice of "Using Math."

Similarly, each task of the assessment is accompanied by a logical argument that can be

built by following the aforementioned generalized procedure. We discussed and refined the assessment tasks with another researcher until each one met all the criteria in the 3D-LAP (task features) to elicit student abilities to engage in the scientific practice of "Using Math" blended with "force."

3.3.2 Data Collection

We conducted Think-Aloud interviews[118;119] with students to answer our second and third research questions. Think-Aloud protocols have been used with individuals with varying levels of expertise in a domain of interest to articulate the information these individuals attend to at a given time and how this information is organized during problem-solving[119]. Interviewers ask subjects to "think-aloud" and verbalize their thought processes while performing cognitively demanding tasks such as problem-solving. According to Ericsson and Simon[118], a subject's verbalization that occurs simultaneously during problem-solving does not alter their thought processes as long as the interviewer does not interrupt with probes.

The participants of our study were students in first or second semester introductory-level, calculus-based physics courses. The students voluntarily participated in this study, and they were remunerated with twenty dollars in gift cards for their participation. We scheduled individual interview sessions that facilitated a quiet environment for subjects to think-aloud simulating an exam environment[119]. We asked students to think-aloud while working on the assessment tasks. For each student, the think-aloud interview lasted about one hour. Similar to an exam, and in keeping with the think aloud protocol, the interviewer did not assist the students with the problems or answer questions about the problems. Like a normal exam, in our interviews, students moved back and forth between problems as they wished, and they determined when they were done with each particular problem.

When students paused for several seconds, the interviewer reminded them to keep thinking aloud. We took notes during the interview that can be followed-up when the interviewee finished the tasks. This led us to further clarify subjects' problem-solving processes and reasoning. The interviews were video and audio recorded and work products in the form of

written solutions were collected and scanned for further analysis.

We did not include interviews with audio issues, and interviews where students did not regularly think-out aloud even after being encouraged to do so several times by the interviewer. Overall, we had 7 distinct assessment tasks that addressed the scientific practice of "Using Math" among 8 interviews, thus giving us 56 total instances for the analysis. Out of the 56 instances, in 3 instances students did not respond to the assessment tasks. Thus, we transcribed remaining 53 student verbal responses both manually and using an AI transcription service[120]. We corrected some of the transcriptions obtained from AI transcription service for their clarity. The accompanying 53 written solutions were gathered for the analysis. All names used in this chapter are pseudonyms.

3.3.3 Data Analysis

In this section, we provide our data analysis approach that helped us answer our second and third research questions. We first provide insights into how we developed our codebook using the ACER framework to analyze student data, incorporating their own knowledge representations[121–123]. We then demonstrate how we code verbal and written responses.

3.3.4 Code Book

To develop the codebook (see Table 3.2), we started by selecting one assessment task and going through all students' responses to that task before looking at another task. We carried out the coding process looking for appropriate subcodes, merging them when they overlapped. We finalized our codebook when no additional subcodes were identified as needed to represent students' work products, i.e. the codebook was saturated. The codebook also captures errors students make while solving the problems by including an "X" in the code. Full codebook with examples from data can be found in Table A.1 and Table A.2 in Appendix A.

Table 3.2: *Portion of the codebook used to analyze data. Each subcode is assigned with a symbol to make the navigation in between subcodes efficiently. See Appendix A for the full codebook with definitions and examples.*

Component	Subcode	Description of the Subcode
Activation (A) ∼ES1	A1	Identify appropriate physics concepts.
	A2	Identify general physics equations to be applied.
	A3	Identify target parameters.
Construction (C) ∼ES2	C1	Apply the general equations to a particular situation.
	C2	Make assumptions.
	C3	Develop representations.
	C4	Develop mathematical relations based on the concepts used.
Execution (E) ∼ES3	E1	Manipulate symbols.
	E2	Perform an arithmetic calculation.
	E3	Execute math conceptually.
	E4	Substitute expressions.
	E5	Manipulate mathematical expressions.
Reflection (R) ∼ES4	R1	Make sense of the answer with the information given in the prompt.
	R2	Make sense of the answer found in an intermediate/final step.
	R3	Make sense of the result for use in a subsequent step.

3.3.5 Coding Verbal and Written Responses

We now turn to the goal of identifying if the tasks were capable of eliciting evidence to achieve the LP. If the assessment tasks have the potential to elicit the expected evidence, they should provide evidence for each component in the ACER framework (that is, activation, construction, execution, and reflection).

Once the codebook was finalized, we coded the students' transcribed verbal responses sentence by sentence. Once a student's verbal response to an assessment task was coded, we compiled the subcodes corresponding to that problem-solving into a list (the "verbal-codes"). Then, the written solutions were coded by assigning an appropriate subcode to each line of a student's solution. Once a student's written solution for an assessment task was completely coded, we compiled the set of subcodes corresponding to that problem solution into another list (the "written-codes"). The student's verbal and written codes were then

synthesized into a single coding pattern (the "combined-codes") which constitutes their overall problem-solving approach. The motivation to obtain combined-codes is to capture a student's complete problem-solving approach. This process was repeated for all 53 verbal and written student responses.

For each assessment task, we analyzed the combined-codes across all students in our data set to explore the task's potential to elicit the expected evidence. This gives us evidence whether or not the intended cognitive processes were tapped during students' problem-solving. If the task was able to elicit the expected evidence – at least one code from each ACER component; A, C, E, and R – from a majority of the students, we determined that the task was good enough to differentiate student abilities to meet the LP. We define majority in our context as $> 50\%$ of students.

If the assessment tasks elicited the expected evidence, we further explored the extent to which the students' written solutions (which are what typically get graded in coursework) accurately reflected their overall engagement with the problem. In order to determine if students' written solutions provided enough evidence to support the claim that they were or were not achieving the LP, we compared each students' written-codes with their combined-codes.

On the other hand, if the task was not eliciting the expected evidence, the component(s) of the ACER which is lacking was documented. This is for the future revisions of that task to deliberately elicit that component(s).

3.3.6 Example Coding

In this section, we provide an example coding for Catherine. The subcodes from Table 3.2 are provided within quotes in square brackets. Catherine started the accident reconstruction problem (see Fig. 3.1) by going through the problem statement and trying to make sense of it. Then, she referred to the equation sheet looking for information that she can relate to the problem. Catherine vocalized her initial thoughts on the problem as follows,

"...the car is going from a, hmm, faster speed down to a stop I can assume that

the, hmm, initial velocity is equal to the, hmm, speed you travel at [inaudible] final velocity is equal to zero just until he stopped and since I have the length of skid mark I can assume that the initial position was zero and then the final position is how long that skid mark was."

She also realized that she was not given any information related to time. Therefore, she chose to use the general equation $V^2 = V_0^2 + 2a(x - x_0)$ to solve for acceleration.

"Hmm, I do not have the time for any of those states. So hmm I... am so this is saying that if the brakes were locked so hmm that's kind of the maximum hmm decrease in acceleration hmm so I'm going to use hmm V squared equals V naught squared plus two a in parenthesis x minus x note hmm [A2]. That way I can solve for a [A3] ..."

Thereafter, she made an assumption that if she knows the acceleration at the crash site, the same acceleration can be used at the real accident. With this assumption in mind, she applied the general equation activated to the crash site to figure out the value for the acceleration. She manipulated the numerical value for the acceleration using the calculator.

"because assuming that the actual accident, the driver locked the brake then to then and I'm just using that same acceleration to see if the driver was at fault or not [C2]. So hmm having zero squared equals fifteen point two meters per second squared plus two a and then in parenthesis it is the twenty two point four meters [C1]. So then just solving for a, [inserting values in the calculator to find the numerical value for acceleration, a] [E2]."

She reflected on the negative value obtained as the acceleration to make sense that it was a reasonable answer as the driver was going from a faster speed down to a lower speed by stating,

"The acceleration is equal to negative five point two one six which once again is a reasonable answer since they are going from a faster speed down to a lower speed. [R2]"

Her goal for this problem was to see when she applied the same acceleration to the actual accident to see what skid mark length it would give and then to compare it with the skid mark given for the actual accident. Thus, it is not that she came up with the answer, but reflected on the answer to see how she can use that information in subsequent steps of the problem-solving.

> *"Hmm then I'm going to take that acceleration and plugging into the exact same problem [R3] hmm to see if hmm the skid mark length that I get is equal to the actual skid mark length. [A3]"*

Then she applied the general equation activated before to find the skid mark length given the speed limit 18 m/s and calculated the numerical answer using a calculator.

> *"Well, I know that x note is gonna be zero just as I'm going from no skid mark to the length I'm just solving for the x [A3], so it will be V is equal to zero again and then I'm going to use V eighteen per second squared two times the negative five point meters per second and then in parenthesis it's x since x minus zero is just x [A2][C1]. Hmm so solving for that is [Calculating the numerical value for x using the calculator] [E2]."*

Once she got the value for the skid mark length as thirty-one point zero five eight meters, she made a comparison with the given value of forty point three meters and determined the driver was at fault.

> *"The x is equal to thirty one point zero five eight meters which is shorter than the skid mark length of forty point three meters. So yes the driver was at fault [R1]."*

We note that Catherine got the answer right following the expected line of a reasoning. Her response pattern corresponds to A3, A2, C2, C1, E2, R2, R3, A3, A2, C1, E2, and R1. Our coding of Catherine's solution includes at least one code for each element in the ACER framework, indicating that the task met the minimum condition to elicit expected evidence

to make conclusion about students' abilities to meet the LP. In addition to Catherine, if the task elicited expected evidence for majority of the students ($> 50\%$ as we mentioned in Sec. 3.3.3), we concluded that the task can elicit students' abilities to meet the LP. Otherwise, it's required that we modify assessment tasks until they elicit the expected evidence to argue about students' proficiency to meet the LP.

On the other hand, Catherine's written solution is associated with the response pattern, A2, C1, E2, A2, C1, E2, R1. The corresponding Written solution mirrors Catherine's problem-solving approach except for the subcodes A3, C2, R2, and R3, eliciting at least single evidence for each component of the ACER framework.

We applied the same process to all assessment tasks by taking into account the students' responses to those tasks to evaluate the tasks' potential to elicit the expected evidence in achieving the LP. If tasks ensured their potential to elicit expected evidence, we further explored the extent to which students' written solution mirrored that potential accurately. In the next section, we attend to some of the interesting aspects of our analysis with more details.

3.3.7 Inter-Rater Reliability

After finalizing the codebook, 5 instances of transcribed verbal and written responses were independently coded by another researcher. These 5 instances included different assessment tasks among multiple students in our data set. After discussion, the coders came to a 100% agreement on 4 of the 5 instances and a 96% agreement on the remaining instance.

3.3.8 Limitations

One limitation of this work is associated with the assumption of the Think-Aloud protocol: verbalized information is the information acquired and heeded by the subject at a given time[119]. However, the human thought processes are rapid enough such that it is likely that subjects verbalize a portion of their thoughts leaving other non-verbal. Thus, only problem-solving processes and verbalized reasoning should be used to make inferences about subjects'

abilities. We also note that our data set includes a small population of students. Expanding the data set to include more (and more diverse) students and incorporating their reasoning in response to assessment tasks would be needed to strengthen their validity for something similar to a standardized assessment. However, we only mean to show this work as a proof of concept.

Another limitation of this work is that the fine-grained ESs are unique to these assessment tasks, and cannot be generalizable across different assessment tasks that address additional concepts and scientific practices. However, given the methodology, one needs to develop their own codebook based on the student evidentiary knowledge representations in their population of students. It is likely that the additional subcodes might appear during that process, but we argue that it cannot be significantly different between the similar problem types we presented in this chapter.

3.4 Results and Discussion

3.4.1 Assessment Tasks Elicited the Expected Evidence for Students' Abilities

As noted in Table 3.3, majority of the tasks (i.e. 6 out of 7) elicited students' reasoning that enabled us to capture their evidence pertaining to each component in the ACER framework [1] except for the Ferris wheel task. This enabled us to capture that when students got the answers right, they got their answers for the right reasons, i.e the expected cognitive processes were tapped.

Situating our work in the ECD approach articulated in Harris *et al.*[3] provides great insight into task design that assesses students' abilities to engage in scientific practices along with concepts. Our work strengthens the generalizability of ECD as a task design approach showcasing its potential to similarly extend into assessing student abilities to engage in

[1]Though this is the minimum condition required to ensure the tasks' potential to elicit the expected evidence, it is typical that many students in our data set used numerous subcodes within each component in the ACER framework while meaningfully engaging in problem-solving.

Table 3.3: *The number of students who responded to each assessment task, the tasks that elicited the expected evidence, and number of students whose written solutions mirrored the elicited expected evidence are provided. "N/A" in the last column refers to the task (Task #3) that does not have the potential to elicit the expected evidence (i.e. Ferris wheel task). Task #1 can be found in Fig. 3.1. See Fig. A.1, Fig. A.2, Fig. A.3, Fig. A.4, Fig. A.5, and Fig. A.6 for task #2, 3, 4, 5, 6, and 7, respectively in Appendix A.*

Task #	# Responses	# Evidence	# Matched
1	7	7	5
2	8	7	5
3	7	1	N/A
4	8	6	6
5	7	4	4
6	8	7	4
7	8	5	4
Total	53	37	28

scientific practices at introductory level physics courses.

We also note that coupling the 3D-LAP with ECD to facilitate task features is promising when it comes to assessing students' abilities to intertwine scientific practices with concepts. Our work further validates the 3D-LAP as an effective tool to elicit student abilities to engage in scientific practices with the support of students' data. Thus, for task developers who have limited time, we suggest the 3D-LAP as a tool to begin with task development. However, we first recommend doing a thorough analysis of the domain of interest to determine the valued concepts to be assessed. This process can be followed by the integration of those concepts with the criteria for scientific practices of interest in the 3D-LAP to develop assessment tasks that can elicit student abilities to intertwine concepts with the scientific practices.

We note that utilizing the ACER framework to analyze both written and verbal work products takes student in-the-moment reasoning into account. This work also expands the utility of the ACER framework to capture students' mathematical reasoning at introductory level.

Overall, we argue that a coherent, systematic approach to designing assessment tasks by coupling ECD with the 3D-LAP is productive when it comes to assessing students' abilities to intertwine scientific practices with concepts. Further, we argue that a framework that

articulates what it means to use math in physics, i.e. the ACER framework guides our task validation process by capturing students' in-the-moment reasoning.

3.4.2 Modifying the Task that Failed to Elicit 'Using Math'

As we explained above, 6 out 7 assessment tasks elicited the expected evidence which showcased students' abilities to achieve the LPs that address the scientific practice of "Using Math." In this subsection, we explain how we can modify the remaining task that failed to elicit the expected evidence (i.e. Ferris wheel task) into a form which potentially would elicit the evidence as intended.

Unlike the assessment tasks that include numerical quantities in their problem statement (6 tasks), the task that includes symbolic variables, the Ferris wheel problem (1 task) did not prompt students to elicit the expected evidence. While our intention was that the Ferris wheel problem has the potential to elicit the expected evidence, 6 out of 7 students who responded started with a conceptual analysis of the problem to determine the positions where a rider in a Ferris wheel feels heaviest and the lightest. However, the follow-up question, "Approximately how large would ω have to be for this to have a noticeable effect on your weight?" prompted them to elicit the expected evidence. For example, given below is how William figured out in which positions the rider feels the heaviest and the lightest in the Ferris wheel problem.

> *"So rotating counter-clockwise. Whenever it's moving up, the acceleration is kind of pulling it outwards so it's not really feeling like wait but when you're at the top hmm you're starting to go down the acceleration straight out so feels like you're moving up, so you're lightest at the top and going down. Heaviest at the bottom and going up just like an elevator."*

He started the problem doing a conceptual analysis of the problem as above, and then made an explanation to figure out the positions where the rider feels the heaviest and the lightest. Thereafter, answering the question about figuring out ω that might give a noticeable

effect on weight, he showcased appropriate evidence as he figured out a reasonable expression for ω.

Thus, we believe that the structure of the variables in the form of symbolic or numerical might affect the way students activate their knowledge, skills, and abilities at hand. Though we see that students well-interpreted and elicited the expected evidence for the assessment tasks that include numerical variables, we do not see the same when it comes to the assessment task with symbolic variables. While we did not specifically probe the question during the interview about why students approached the way decided in response to the task that includes symbolic variables, this gives us some initial clues about the ways they interpret the tasks with respect to the nature of the variables in the problem statement.

In particular, the utterances students made in response to the Ferris wheel task, *"There is no numbers so... It seems kind of broad", "[student is asking a question from the interviewer] So, with this question, how does it depend on diameter? Do you wanna leave those [symbolic variable of "D" for diameter] in our answer?"* provided us initial evidence that they paid attention to the nature of the variables in the assessment tasks. However, more work is needed to strengthen the argument behind the dependability of the variable types in the problem statement that prompts students to elicit the expected evidence encouraging mathematical reasoning.

One potential future work is to intentionally design tasks that include both symbolic and numerical variables and explore their problem-solving approaches with respect to the variable types. Such work can give great insight into the ways in which we can prompt students to engage in more conceptual analysis of the problem rather than mere mathematical manipulations. Bringing our attention back to the task validation process, we further need to revise this task until it has the potential to elicit expected evidence to capture students' abilities to intertwine math with physics concepts.

3.4.3 Written Solutions Mirrored the Elicited Expected Evidence for Students' Abilities

In order to address our third research question, we analyzed the extent to which the written solutions accurately represented the students' reasoning during problem-solving. For this analysis, we looked only at the six tasks that successfully elicited evidence of Using Math. From those six tasks, 36 of the combined-codes included all four components of the ACER framework. In 28 out of those 36 instances, the written-codes covered the same elements of the ACER framework as the combined-codes (see Table 3.3). In other words, though in these instances students elicited evidence for each component in the ACER framework, the reflection component (i.e. R1, R2, and R3, provided in Table 3.2) is not mirrored in the written solution.

Students who engaged in reflections verbally, but did not include it in their written work might be an important aspect to further look into. This is because students' reflections during physics problem-solving are crucial and cognitively demanding. What students wrote down as part of working through a problem might be the things they believe instructors are valuing in their work. Therefore, by strengthening the importance of reflecting on responses students obtain to make sense of them at an earlier stage such as an introductory level is crucial. The lack of evidence for students' reflections in the written work suggests that we can modify our task features in a way that those elements are more conspicuous. In particular, we can scaffold the task prompting students to demonstrate proficiencies associated with reflections at an earlier stage. As students progress through the curriculum, the scaffolding can be removed to promote their autonomy to engage in reflection.

For example, Kang *et al.*[124] show that high quality scaffolding can provide students opportunities to demonstrate their disciplinary proficiencies. Careful scaffolding to elicit student reasoning in assessments is also encouraged in the work by Cooper and Stowe[125]. However, the scaffolding should not guide any specific problem-solving patterns. Rather the assessment tasks should allow students to construct solutions on their own to preserve the authenticity of the scientific practices along with concepts[36]. For example, we can explicitly

guide students to utilize a self-constructed representation of a system that models a real-world phenomenon by including a question prompt similar to, "construct a representation that models the physical system as part of your solution."

In addition to developing assessments to elicit students' abilities, communicating these abilities to faculty is also important. Communicating this information to faculty can enable them to modify their courses to provide better learning opportunities for students. In the next chapter, we provide a methodology to design feedback for physics faculty that can guide course modifications, in parallel to designing a new research-based assessment.

Chapter 4

Designing Research-Based Assessment Feedback for Physics Faculty

In this chapter, we present our methodology to design feedback for faculty in parallel to designing a new research-based assessment. This feedback can support and encourage an individual faculty member to modify their course. In this chapter, we answer the following research questions.

1. How can we design feedback for faculty that facilitates course modifications?
2. What are the perceptions of the faculty towards the generated feedback?
3. How can we incorporate perceptions of the faculty to improve the generated feedback?

In the next section (Sec. 4.1), we provide the theoretical background that informs the design of feedback. In Sec. 4.2, we provide the methodology to develop feedback, obtain perspectives from faculty about the developed feedback through interviews, and incorporate those perspectives to update the feedback. We provide a discussion around this methodology in Sec. 4.3.

4.1 Theoretical Background

4.1.1 Including Feedback within a Research-Based Assessment

Development of a research-based assessment includes validating its measurement argument [50;51]. Measurement argument includes the scores an assessment can generate after students took the assessment, and interpretation to these scores that can reflect student learning. Validating a measurement argument ensures the interpretations made on scores are accurate measures of student learning.

There is a call to conceptualize assessments not just as measuring instruments, rather instruments that can create impact on individuals. Articulating a theory-of-action for an assessment is an approach to address such call [53;126–128]. As Bennett [129] stated,

> *"Theory of action considers the assessment system to not only be a measure, but also a "treatment"."*

Viewing a research-based assessment as a *treatment* suggests not just validating a measurement argument, but also a *theory-of-action*. A theory-of-action for an assessment subsumes its measurement argument within a broader assessment argument.

A theory-of-action explicitly focuses on the change a research-based assessment can make on both individuals (e.g., students, faculty, and administrators) and institutions (e.g., departments, and schools) (see Fig. B.10 in Appendix B for an example). Along with the measurement argument, theory-of-action includes elements such as "components of the assessments," "intended effects," and "action mechanisms" to cause the intended effects. In the following paragraph, we explain these elements using an example.

The components of an assessment can include questions intended to elicit students' abilities to engage in logical reasoning. One of the intended effects of administering these questions would be to improve students' abilities to make logical reasoning. To achieve this intended effect, one of the action mechanisms that can be enacted by a faculty member would be to evaluate students' responses to the provided questions. This evaluation can provide information to the faculty member about students' abilities to engage in logical reasoning.

If students need more opportunities to practice logical reasoning, the faculty member can create a classroom activity that can provide opportunities for students to engage in logical reasoning.

Instead faculty themselves evaluating students' responses to the provided questions, researcher-generated feedback within a research-based assessment can provide recommendations to faculty, which can facilitate their subsequent course modifications. Similar to valuing "logical reasoning," this researcher-generated feedback can be developed based on the evidence presented in students' work for any valued aspect of student learning the assessment targets. Thus, faculty members' subsequent course modifications guided by the feedback can enhance student learning.

4.1.2 Arguing from Evidence to Reason about Student Performance

There are calls to transform present day physics classrooms into authentic science learning environments. In these learning environments, students are supposed to make use of their knowledge to create new knowledge. This is to bring together students' knowledge and its use into situations that are novel to students – students' knowledge-in-use[3]. One approach to doing this is bringing core physics concepts (core ideas), concepts that are common across science (crosscutting concepts), and practices that generalize actions of scientists (scientific practices) together into physics curriculum, instruction, and assessment[12;36]. Assessing students' knowledge-in-use is crucial to understand what opportunities can be given to them to enhance their learning[35].

Evidence-centered Design (ECD)[20] has been identified as a framework that can be leveraged to assess students' knowledge-in-use[12;36]. ECD orients around an assessable knowledge-in-use statement, known as a **learning performance (LP)**[3], which informs the **knowledge, skills, and abilities (KSAs)** being targeted by the assessment. The students' KSAs are explored through the lens of **evidence statements (ESs)**, which articulate the observable features of the student data. To elicit ESs from the student work, the task features that

can characterize students' knowledge-in-use can be used[130].

The Three-Dimensional Learning Assessment Protocol includes a distinct set of criteria for core ideas, crosscutting concepts, and scientific practices[7]. To elicit students' abilities to engage in these three crucial aspects, an assessment task should collectively meet all of the criteria. A developed task that collectively meets all of the criteria can elicit students' knowledge-in-use. Thus, these criteria can be used as task features to develop assessment tasks that can potentially elicit ESs that inform knowledge-in-use from student work.

The development of the feedback for faculty is guided by the students' demonstrated KSAs as observed through the lens of ESs. The extent to which students demonstrated the expected KSAs can inform the opportunities faculty members can provide students to better facilitate learning as needed[28].

4.1.3 Promoting Learner-Centered Approach when Designing External Feedback for Faculty

External feedback was historically considered as a unidirectional transmission of information from an expert to a novice learner[79]. Under this notion, more emphasis was placed on the ways in which an expert person can improve their practices associated with providing feedback to a novice learner. Focusing on the expert person, instead of the novice learner, limits the uptake and use of the received feedback by the novice learner. If a person takes up and uses the external feedback received, it has been shown to improve their learning[131;132]. Thus, to increase the likelihood of taking up and incorporating external feedback by the novice learner into their ongoing learning, the feedback design should focus on the learner[79;133–135].

One such approach to leverage external feedback where the learner is centered to the ongoing learning process is provided in the model developed by Butler and Winne[2] (see Fig. 4.1). This model explains the processes a learner goes through as learning progresses and how external feedback can intersect with that ongoing learning. We now turn to explaining these processes explicitly along with the impact the external feedback can create on these

processes.

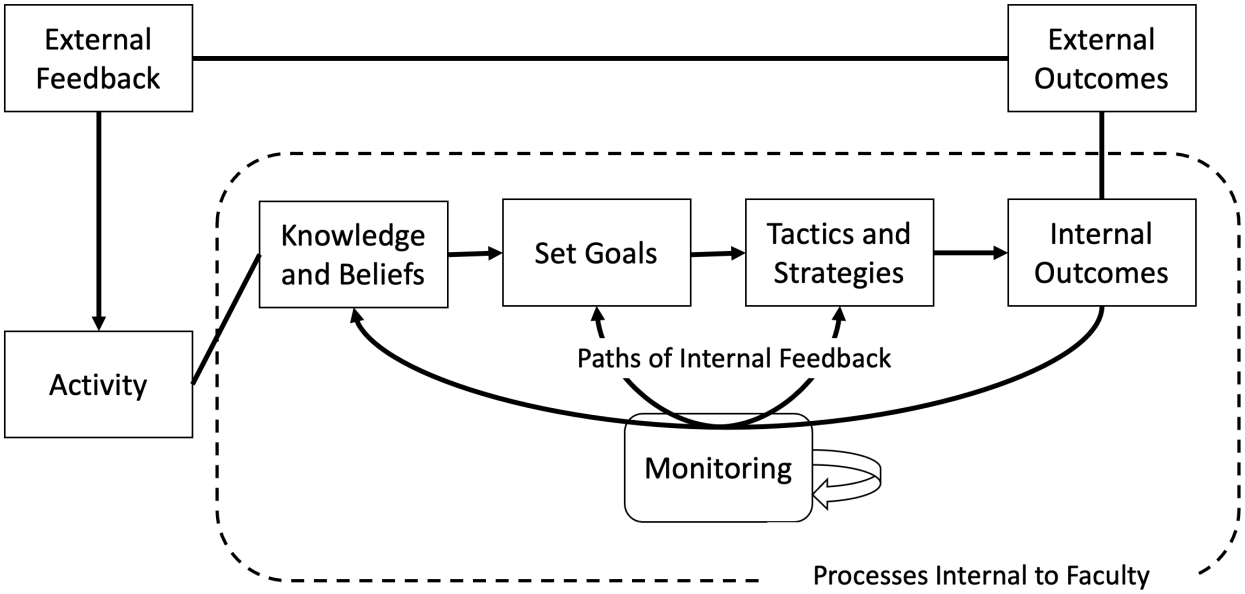

Figure 4.1: *A model for self-regulated learning in the context of external feedback. This figure is recreated from its original version presented on Butler and Winne[2]. The "Task" and "Cognitive System" from the original version were changed to "Activity" and "Processes Internal to Faculty" respectively, to align with the context of our study. In addition to align with our context, "Task" was replaced with "Activity" to reduce the potential confusion between "Task" and "Assessment Task." The "Products" and "Performance" from the original version were changed to "Internal Outcomes" and "External Outcomes" respectively, for effective communication.*

When a learner engages in an activity, they draw on their knowledge and beliefs, set goals for themselves, draw on tactics and strategies to achieve goals, and produce outcomes as they engage in the activity. These outcomes could be either internal or external. An example for an internal outcome could be emotions one feels once they engage in the activity as they either achieved or partially-achieved the goals they set. On the other hand, external outcomes are the outcomes that are visible to outsiders. An example for an external outcome could be a talk a learner gives.

Once the outcomes are produced, a learner starts comparing outcomes produced with the goal that was initially set. This self-monitoring process leads to generating internal feedback. Internal feedback can cause a learner to reinterpret the task, set new goals, draw on different tactics and strategies, and produce new outcomes as they make progress in the activity.

In addition to the internal feedback a learner generates for themselves, external feedback can be received on the external outcomes a learner produces. External feedback is usually received in relation to the standards set externally. The external feedback that can support a learner should[133]:

i. Clarify the expected performance,
ii. Communicate the current state of the performance, and
iii. Provide opportunities to close the gap between the current and the expected performance.

Similar to students having learning opportunities in the classroom, faculty also have learning opportunities in the classroom[136]. These learning opportunities are enacted when faculty design courses, deliver instruction, and reflect and modify courses to better support student learning. Promoting a learner-centered approach to designing external feedback for faculty, we use the model described above along with the features listed above to similarly characterize the learning process faculty undergo[98].

The development of the feedback for faculty is conducted in parallel to developing a new standardized assessment under development, which is the Thermal and Statistical Physics Assessment (TaSPA). We have deliberately chosen the TaSPA as the research-based assessment that can provide evidence for student learning. This is because the development of the TaSPA is informed by ECD, and we are not aware of any other available research-based assessment in PER that attends to student learning from an evidence-based reasoning approach. In the next section, we first provide some background to the TaSPA, following methods pursued to develop the feedback for faculty.

4.2 Feedback Development Methodology

4.2.1 Context

Figure 4.2 shows how TaSPA can be operationalized in a classroom. When available to use, the TaSPA will allow faculty to choose the LPs they value and prefer to assess in their

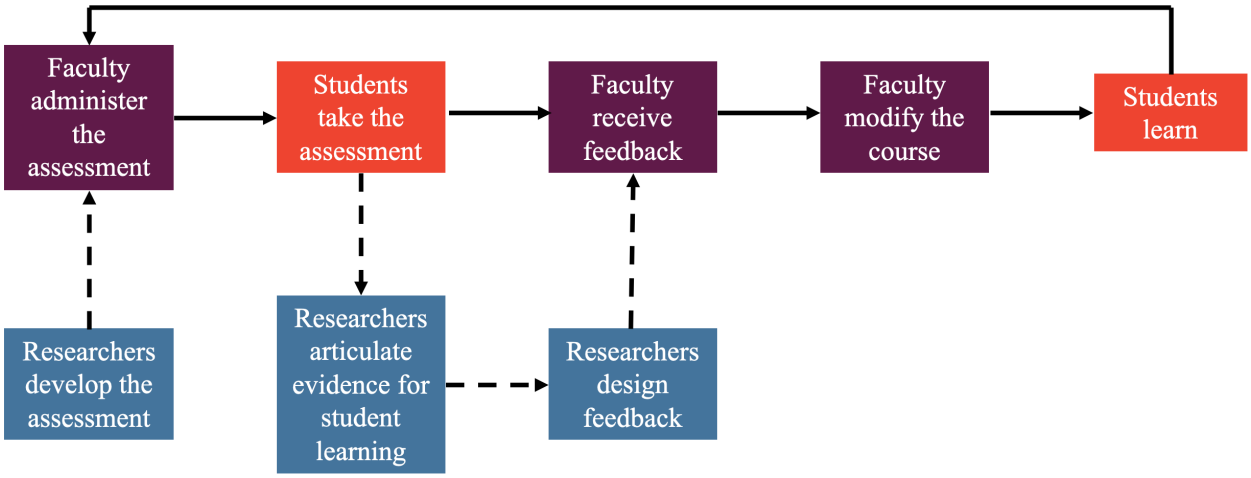

Figure 4.2: *Operationalizing the TaSPA in a physics classroom. "Purple" represents the actions intended to be taken by faculty. "Orange" represents actions associated with students. "Blue" represents the researcher's involvement during the creation of the TaSPA and its corresponding feedback. Dashed line further separates actions intended from researchers with faculty and students.*

classroom. Faculty would then administer the assessment in their classroom, which includes assessment tasks aligning with the chosen LPs. These assessment tasks take a coupled, multiple-response (CMR) format to facilitate online test administration with streamlined evaluation of student work[137;138].

The development of a CMR task begins from developing, piloting, and analyzing student responses to free response (FR) version of the task[138]. The students' common response patterns identified via analysis of student responses to the FR task inform the answer options available for students to select in the corresponding CMR assessment task.

Typically, a CMR task is structured as a cluster of questions, i.e. multiple-choice question (one answer is allowed for students to select) followed up by multiple-response questions (multiple answers are allowed for students to select). The students are asked to select an appropriate answer from the multiple-choice question along with the justification for that selection from the multiple-response questions.

Based on students' selections to the CMR version of the assessment task, faculty would receive feedback on their students' performance. This feedback communicates to faculty about the extent to which students achieve a LP along with suggestions that can guide

course modifications. These suggestions include opportunities that can be given to students to stimulate learning, as needed. The uptake and use of the feedback by the faculty would help students better achieve valued LPs.

4.2.2 Steps Associated with the Feedback Development Process

In this subsection, we answer the first research question that we laid out in the beginning of Chapter 4: *How can we design feedback for faculty that facilitates course modifications?* Designing feedback for faculty involves several steps, which can be found in Fig. 4.3. In the following paragraphs, we explain each step in detail.

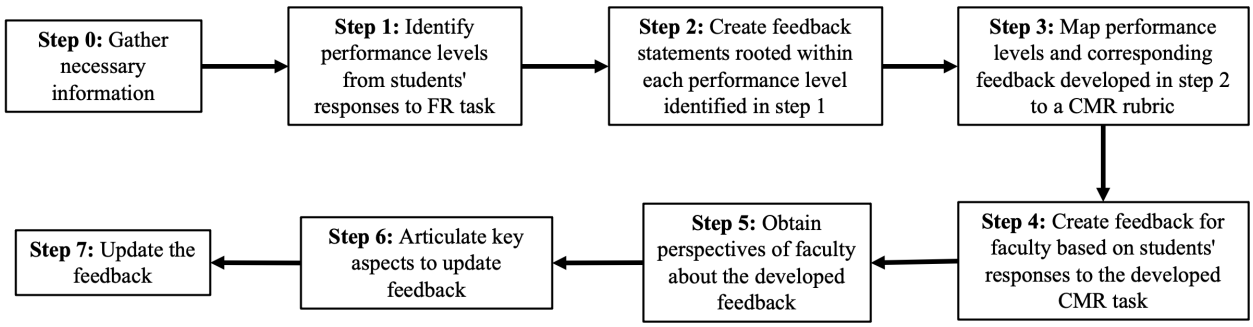

Figure 4.3: *Steps included in the development of the feedback for faculty.*

Step 0: Gather necessary information

Since the design of the feedback for faculty is rooted in the TaSPA, we first need to gather some background information from the TaSPA development process. This includes adopting a designed LP, and its corresponding KSAs and ESs from the assessment task development stage of the TaSPA[27;28]. Table 4.1 provides an example LP, and its corresponding KSAs and ESs that we used to demonstrate the methods for designing feedback. Though we present all three KSAs and ESs required to demonstrate students' engagement with the provided LP in Table 4.1, we only use **KSA2 and ES2** to demonstrate the feedback development methodology. We note that the methods presented can be mirrored to other KSAs and ESs as well.

Table 4.1: *Information necessary to develop feedback for faculty, which was adopted from the assessment development stage of the TaSPA. The definitions of Learning Performance (LP), Knowledge, Skills, and Abilities (KSAs), and Evidence Statements (ESs) were adopted from Ref.*[3].

Component	Definition	Example
LP	Assessable knowledge-in-use statement.	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.
KSAs	Proficiency needed to demonstrate the LP.	KSA1: Unpack relations that connect change in internal energy to heat and work. KSA2: Generate explanation about the change in internal energy of the system using relations that include heat and work. KSA3: Construct a statement about the change in internal energy of the system.
ESs	Observable features of student proficiency.	ES1: Relations that connect change in internal energy to heat and work. ES2: Generated explanation about the change in internal energy of the system using relations that include heat and work. ES3: Statement about the change in internal energy of the system.

In addition to adopting the LP, and its corresponding KSAs and ESs provided in Table 4.1, we used the student data collected during the TaSPA development stage. We used student data in response to the FR task in Fig. 4.4, developed to address the LP, and its corresponding KSAs and ESs in Table 4.1. We also used student data in response to the CMR task provided in Fig. 4.5 that addresses the same LP, KSAs, and ESs in Table 4.1, which was developed based on student data in response to the corresponding FR task as we explained in Sec. 4.2.1. The development of a CMR task from its corresponding FR task can be found elsewhere[27]. We note that our goal is to provide a methodology to design feedback for faculty which is rooted in the student responses to the CMR task, which is the task format the TaSPA entails. However, we cannot ignore the student responses to the FR task as this directly informs the development of the CMR task.

The student written work in response to FR task in Fig. 4.4 were collected in a final exam setting ($N = 72$). The student data in response to CMR version of that task in Fig. 4.5 were

collected online as a pre-class assignment ($N = 43$). These data were collected at a public research university in the US. We note that the goal of this work is to provide a mechanism to design feedback for faculty, rather than characterizing student responses, for example, gaining qualitative insight into students' difficulties.

Step 1: Identify performance levels from students' responses to FR task

We analyzed the students' responses to FR task in Fig. 4.4 with the lens of ES2 provided in Table 4.1. The goal was to explore the nuance of students' KSA2 with respect to ES2, as they are presented on students' responses. We identified three characteristic problem-solving patterns students enacted when presented with the assessment task provided in Fig. 4.4.

When this task is provided to students, to achieve the required proficiency for ES2, we ideally expect them to incorporate both heat and work as energy forms to generate explanations about the changes to internal energy of the system. These explanations should be ideally made when the piston moves upwards and downwards.

For example, when students used *both* heat and work to generate explanation about changes to internal energy of the system when the piston moves upwards and downwards, we identified that as the "performance-met" condition with respect to ES2 (see Fig. 4.6). When students used *either* heat or work to generate explanation about changes to internal energy of the system (but not both) when the piston moves upwards or downwards, we characterized that as the "performance-partially met" condition (see Fig. 4.6). On the other

> Your lab partner sets up today's apparatus: a gas-filled metal cylinder with a movable piston on top and a burner underneath. Following the lab instructions, you turn the burner on and observe the piston slowly move upwards. After a moment, your partner slowly pushes the piston down and then holds it at its initial position.
>
> What has happened to the total internal energy of the system since the beginning of the experiment? Justify your answer using appropriate physics principles and be explicit about any assumptions you made.

Figure 4.4: *The FR version of the internal energy task to elicit student work that aligns with the LP and corresponding ESs provided in Table 4.1.*

How did you reason about the expansion? *(select all that apply)*

- ☐ The expansion was isothermal
- ☐ The expansion was isobaric
- ☐ The internal energy increased during the expansion
- ☐ The internal energy decreased during the expansion
- ☐ The internal energy remained the same during the expansion
- ☐ Heat entered the system
- ☐ Heat left the system
- ☐ Temperature increased during the expansion
- ☐ Temperature decreased during the expansion
- ☐ Energy flowed into the gas due to work
- ☐ Energy flowed out of the gas due to work
- ☐ No work was done during the expansion
- ☐ Other:

Figure 4.5: *A portion of the corresponding CMR version of the internal energy task in Fig. 4.4 to elicit student work that aligns with the LP and corresponding ESs provided in Table 4.1. This portion of the CMR task only informs ES2.*

hand, when students *did not use either* heat or work to generate explanation about changes to internal energy of the system, we identified that as "performance-not met" condition. We only use performance-met and -partially met conditions for the rest of this chapter to demonstrate the methodology.

In the future, we intend to conduct student interviews with FR tasks to incorporate perspectives that students can bring when solving those tasks. We acknowledge that this can lead to identifying performance levels which could not have been captured through only written solutions.

Step 2: Create feedback statements rooted within each performance level identified in step 1

Our next goal was to bridge the student performance with respect to ES2 with the feedback for faculty. This way we can communicate to faculty the ways in which student learning associated with KSA2 can be better facilitated, as needed. We communicate this information to faculty, aligning with the three features that can strengthen learning of faculty members

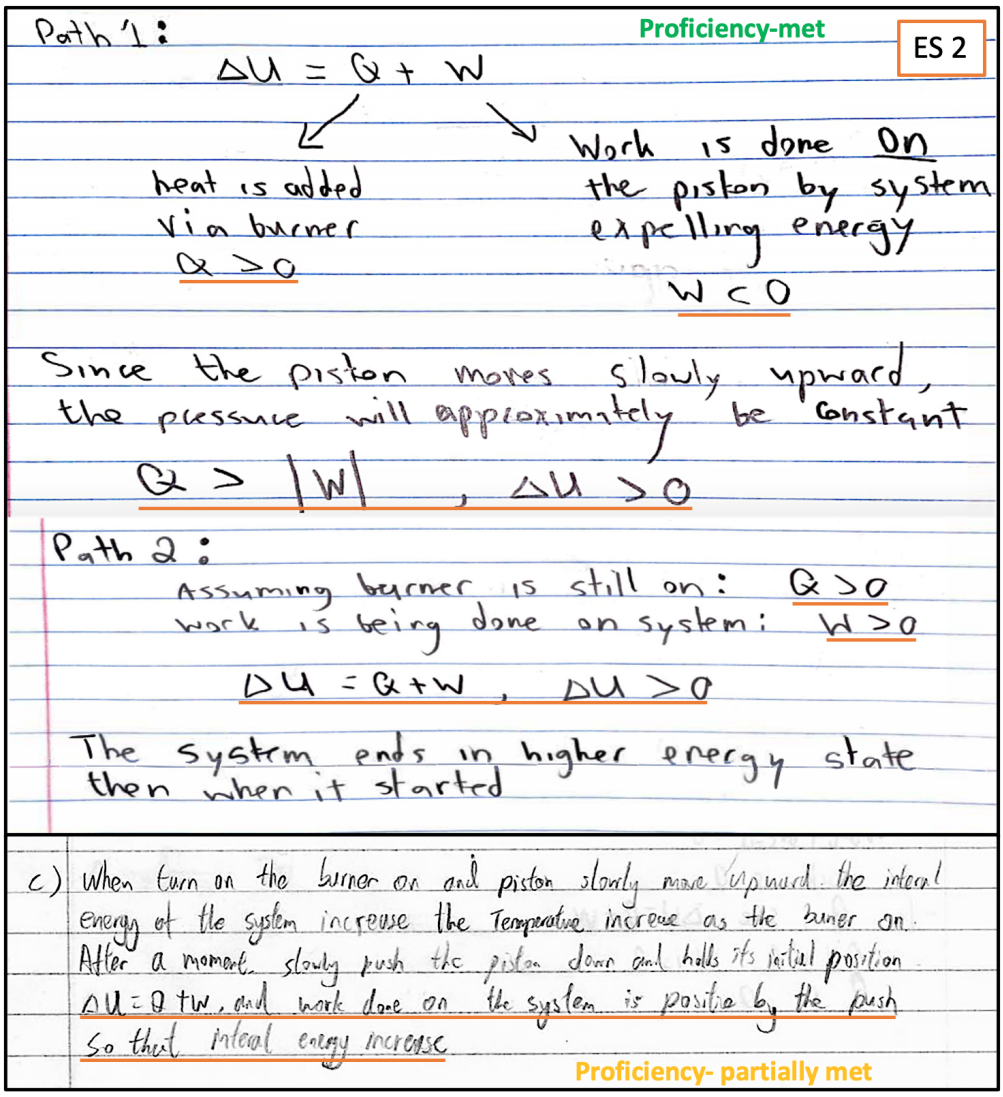

Figure 4.6: *The student's solution on the top corresponds to the performance-met condition, while the student's solution on the bottom corresponds to the performance-partially met condition. Both of these solutions were coded with respect to ES2.*

as we laid out in Sec. 4.1.

We first provide information to faculty about the expected performance. For example, when aligning feedback with respect to ES2, we provide the expected performance as "Students construct arguments about the changes to internal energy of a system. These arguments are composed of coherent reasoning that takes into account the contributions from *both* heat and work as forms of energy flow into and out of that system."

Secondly, we provide information about the current state of the performance in com-

parison to the expected performance. For example, if students in a class achieved the performance-met condition, the current state of the performance that communicates to faculty would be "Students met the expected performance."

Thirdly, we provide faculty recommendations that include suggestions to close the gap between the expected and current performance, as needed. Given the expected and current state of the performance above, where there is no gap, faculty would be given "No course modifications are suggested to address the expected performance."

On the other hand, if students in a class achieved the performance-partially met condition with respect to ES2, we communicate to faculty about the current state of the performance as "Students constructed an argument about the changes to internal energy of a system during the full duration of the considered process by taking into account the contributions from *either* heat or work as forms of energy flow into and out of that system, but not *both*." The next step is to provide suggestions to faculty to close the performance gap to help students achieve the performance-met condition with respect to ES2, in comparison to performance-partially met condition.

Thus, we communicate to faculty that "Students can be given more opportunities to generate coherent explanations about how both heat and work as forms of energy can concurrently contribute to the changes in internal energy of a system. Embedding these opportunities in real-world scenarios that include systems undergoing multiple processes could be helpful for students. Such processes could include isobaric, isochoric, adiabatic or isothermal expansions or compressions." We create these suggestions to faculty and then collectively discuss among other researchers for its accuracy and clarity. We acknowledge that if student interviews led to identifying additional performance levels during step 1 in the future, the feedback statements should also be updated aligning to those identified performance levels. A set of feedback aligned with a set of LPs can be found in Table B.1 and beyond in Appendix B.

Step 3: Map performance levels and corresponding feedback developed in step 2 to a CMR rubric

Since the assessment tasks in the TaSPA take the format of CMR, we need to explore how feedback can be designed for tasks in such format. As we explained in Sec. 4.2.1, CMR task is built from its corresponding FR task, and the answer options available for students in the CMR tasks are informed by students' responses to the corresponding FR task. Thus, we mapped the performance level identified in step 1 to a CMR rubric. Within the CMR rubric, students' appropriate answer selections from the available set of options that attribute to a certain performance level are also noted (see Fig. B.11 and beyond in Appendix B for the CMR rubric). This way, based on students' selections, the associated feedback statements will be available for faculty.

For example, if majority of the students selected at least the following selections from the CMR task provided in Fig. 4.5, faculty would receive the feedback associated when students met the *performance-met* condition, which we explained in Sec. 4.2.2: "The internal energy increased during the expansion," "Heat entered the system," and "Energy flowed out of the gas due to work."

On the other hand, if majority of the students selected at least the following selections from the CMR task provided in Fig. 4.5, faculty would receive the feedback associated when students met the *performance-partially met* condition, which we explained in Sec. 4.2.2: "The internal energy increased during the expansion" along with "Heat entered the system," but *not* "Energy flowed out of the gas due to work" or "The internal energy increased during the expansion" along with "Energy flowed out of the gas due to work," but *not* "Heat entered the system."

We created the corresponding CMR task for the FR task given in Fig. 4.4, prior developing the corresponding CMR rubric by following the methods in Ref.[27]. After creating the CMR rubric, we iteratively modified the developed CMR task to better align with the corresponding CMR rubric.

Step 4: Create feedback for faculty based on students' responses to the developed CMR task

We present an example feedback that can be provided to faculty, which was generated by using the developed CMR rubric to code the students' responses received after piloting the CMR task. When coding students' responses based on the CMR rubric, the students' selections that are aligned with performance levels were used.

Figure 4.7 shows an example of feedback, based on the students' responses to the piloted CMR task as they were captured through the CMR rubric. We note that we used the overall classroom trends to determine the feedback that would appear for faculty. For example, within each ES, we explore the percentage distribution of the student population among performance levels identified. We explore the extent to which this distribution skewed towards "performance-met," "performance-partially met," or "performance-not met" conditions. If the majority ($70\% - 100\%$) of the students leaned towards any of these performance levels, the associated feedback with that respective performance level would appear for faculty.

Expected Performance	Evidence of Expected Performance			Students' Performance based on Overall Class Trends	Recommendations for Course Modifications to Address the Expected Performance
	No	Some	Yes		
ES1	…	…	…	…	…
ES2	95%	5%	0%	Students did not take into account the contributions of either heat or work as forms of energy flow into and out of that system to construct an argument about the changes to internal energy of a system during the full duration of the considered process.	• Students can be given more opportunities to generate coherent explanations about changes in internal energy of a system when the considered process involves concurrent contributions from both heat and work. • Situating these opportunities in real-world scenarios that include systems undergoing multiple processes (e.g., expansion and compression of gases) could be helpful for this population of students.
ES3	…	…	…	…	…

Figure 4.7: *Example feedback for faculty, aligning with the students' responses to the piloted CMR version of the internal energy task provided in Fig. 4.5. Only a portion of this task is provided in Fig. 4.5. The feedback associated with only ES2 is provided. "No," "Some," and "Yes" correspond to the "performance-not met," "performance-partially met," and "performance-met" conditions. Overall class trend is determined by the "No" state as 95% is within the majority of $70\% - 100\%$. Thus, recommendations aligned with the "performance-not met" condition appears for faculty.*

Step 5: Obtain perspectives of faculty about the developed feedback

Because this form of feedback is new and outside the norm, it is important to understand how faculty would respond to the feedback provided to them. For that, we conducted one-on-one, semi-structured interviews with faculty who teach upper-division/intermediate thermal physics. We refer to "thermal physics" as the combination of classical thermodynamics and statistical mechanics. To recruit faculty for the interviews, we first sent out emails to department chairs in diverse sets of institutions within the US, including minority-serving institutions (MSIs) in addition to the typically larger predominately white research institutions. We requested that they forward an email to faculty who teach upper-division/intermediate thermal physics in their departments. The participants who responded indicating interest in participating in interviews were then filtered to incorporate perspectives based on several factors identified beforehand as potentially influential for faculty's assessment and instructional practices.

These factors include faculty's institution type (e.g., research-focused, teaching-focused, MSIs), teaching experience (e.g., first time teaching thermal, multiple years of experience in teaching thermal), academic rank or administrative role (e.g., full professor, lecturer, chair), and number of students typically enrolled in the course (e.g., instructors who explicitly mentioned in the response email that their thermal course includes a small number of students). We used a demographic survey at the end of the interview to collect faculty's demographic information (e.g., race/ethnicity and gender), and participants were asked to respond only if they preferred to. We conducted the interviews ($N = 10$) through Zoom, and interviews were video and audio recorded. We also utilized Zoom's built-in transcription option. The transcripts were later revised to correct for anything Zoom got incorrect or missed. Table 4.2 provides more information about the interview participants.

Our interview protocol (see Appendix B.1 for the full interview protocol) captures three broader scopes which we explain next. These three scopes include questions that prompt faculty to elicit the classroom practices associated with modifying the thermal physics course they teach, views on a set of LPs provided and the likelihood of assessing them in their class-

Table 4.2: *Demographic information of the interview participants. All the names provided are pseudonyms. Faculty reported, an introduction to thermal physics by Schroeder[4], thermal physics by Kittel and Kroemer[5], and thermal physics by Baierlein[6] as the textbooks that they use in their classrooms. The number of students presented here are the average number of students in classrooms, as faculty reported. The acronyms, PWIs and HSIs correspond to Predominantly-White Institutions and Hispanic-Serving Institutions, respectively. The race/ethnicity and gender were self-identified. In the survey form, race/ethnicity was provided as "Caucasian/White" and "Hispanic/Latinx."*

Pseudonym	Textbook	# Students	Institution	Race/Ethnicity	Gender
Dr. William	Schroeder	5-22	PWI	White	Man
Dr. Andreas	Schroeder	10-15	PWI	White	Man
Dr. Michael	Kittel and Kroemer	2-6	PWI	White	Man
Dr. John	Baierlein	4-15	HSI	White	Man
Dr. Ginny	Schroeder	10-20	HSI	White	Woman
Dr. Demetri	Schroeder	7-17	HSI	White	Man
Dr. Justin	Schroeder	15-23	HSI	White	Man
Dr. Basilio	Schroeder	6-8	PWI	White, Hispanic	Man
Dr. Arthur	Schroeder	25-30	PWI	White	Man
Dr. Thomas	Schroeder	30-35	PWI	White	Man

rooms,[1] and perspectives on feedback aligned with two LPs. One of the feedback reports was aligned with the LP we used to illustrate our methodology in Sec. 4.2: *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.* The other feedback report was aligned with a different LP: *Use a representation of a physical system to determine the number of microstates for a given macrostate to predict the system's macroscopic property of entropy.* Both of these LPs were included in the set of LPs provided to faculty, when we asked their views on them.

One of our major goals for conducting faculty interviews was to explore how faculty consider modifying a course based on the current state of the student performance. For this, we provided different combinations of the percentage of the student population who met each performance level within each ES for the above mentioned two LPs (see Fig. 4.7 for example). Thus, we used these combinations as artifacts that allowed faculty to reflect on and determine how they might make modifications to their courses.

[1]We selected these LPs for the faculty interviews as we already developed assessment tasks and feedback aligned to these LPs.

We provided distributions when 50% of students met, 50% of students partially met, or 50% of students did not meet the proficiency within the ES. Other combinations included instances where percentages were almost equally divided in between three performance levels, or either skewed significantly (e.g., 85% and 73%) to performance-met or -not met levels. We chose these specific percentage distributions in part because we as a research team had varied perspectives on developing feedback for distributions such as bimodal distributions. Thus, we need to recruit perspectives from faculty on how they might interpret these percentages.

During interviews, we provided these percentages aligned with each LP to faculty and asked them to reflect and provide interpretations on them, in-the-moment. We had a set of questions at the end of each feedback report asking faculty about the provided feedback. The goals of these questions were to explore faculty's perspectives on the perceived utility of the provided feedback in the classrooms, i.e. usefulness and challenges of the feedback, along with ways in which it can be improved.

Step 6: Articulate key aspects to update feedback

In this subsection, we answer our second research question that we laid out in Sec. 1: *What are the perceptions of the faculty towards the generated feedback?* We used thematic analysis to characterize the interview data, along with an inductive qualitative approach[139]. We provide a detailed description of the analysis below.

Familiarize with the data: We have iteratively watched each video and audio recorded interview data of 10 physics faculty (about 30 min each), when they reflected on and provided perspectives on the provided researcher-generated feedback aligned with the two LPs. As we were watching data, we recorded aspects that faculty brought up during the interview that were aligned with the second research question. A narrative was generated for each interview that summarizes perspectives of faculty in relation to the second research question.

Generate initial codes: We started coding the data, which is the process of identifying and labeling features of the data – codes – that relate to the second research question. We assigned a code to a block of text on the transcript. While coding, both video data and

the transcript modified for clarity were simultaneously used. To provide an example for the process of coding, when going through the utterance,

> *"So one thing that might help along with the recommendations and all this is looking at what the questions are like if I and just, you know, if you gave here were the questions that assess this learning goal. . . ,"*

we assigned a *code*,

> *"Faculty need information about the assessment task in which the learning goal being assessed."*

When coding, each of the relevant portions of the transcript was color-coded and assigned a descriptive label (like the *code* given above) with the same color-coding to identify them easily.

We kept reading the data until the next potential block of text that relates to the second research question was identified. Then, we examined whether a same code can be applied, or a new code is needed to capture that block of text. As the coding progressed on, the descriptive labels attached to the codes were modified to entirely capture the meanings of the new block of text with similar features. This process was repeated for 10 interviews.

Search for Themes: A theme well-captures the pattern of the data in relation to the research questions. Identifying themes involves clustering codes into similar units which reflect similar characteristics. When clustering, only codes that appear in more than one interview were used.

Review Potential Themes: In this phase, the themes need to be reviewed with respect to both the coded data and entire data set. First, the four themes generated were reviewed with respect to the coded data. After each theme was reviewed in relation to the coded data, it was reviewed in relation to the entire data set. This makes sure themes well-capture the crucial aspect of the data with respect to the second research question. We provide each theme, the codes constitute each theme, and faculty who represent each code in Table 4.3 and Table 4.4. In the following paragraphs, we explain the final set of themes with examples

from the data.

Theme 1: Faculty discuss making modifications to a course based on the extent to which students met (or not met) the expected performance.

This theme was generated based on six codes (given below), each of which corresponds to instances where faculty reflected on how they would make course modifications based on the current state of the student performance provided to them. The current states of the student performance are characterized though "yes," "some," and "no" (see Fig. 4.7). These are representative of the "performance-met," performance-partially met," and "performance-not met" conditions. The number of students for each of the step as a percentage were also provided to the faculty. Note that in the version presented to faculty, we marked the percentage of zero simply with a "-".

The following paragraph describes the six codes:

1. Faculty identify 50% "no" is something significant that still needs to be addressed even when the rest of the 50% achieved "some."
2. Faculty identify 50% "no" is something that still needs to be addressed even when 50% achieved "yes."
3. Faculty identify 73% "no" is something that still needs to be addressed even when 27% achieved "yes."
4. Faculty identify 50% "some" is something that still needs to be addressed even when the rest of the 50% achieved "yes."
5. Faculty identify evenly distributed percentages among "no", "some", and "yes" as something that still needs to be addressed.
6. Faculty identify 85% "yes" and 15% "no" as a successful performance.

Faculty discussed making modifications to a course based on the extent to which students met (or not met) the expected performance. For example, from the provided percentage distributions, if the performance data skewed significantly towards students not meeting

Table 4.3: *Themes, codes, and faculty representing codes.*

Themes	Codes	Faculty representing codes
Faculty discuss making modifications to a course based on the extent to which students met (or not met) the expected performance.	Faculty identify 50% "no" is something significant that still needs to be addressed even when the rest of the 50% achieved "some."	Dr. William, Dr. Andreas, Dr. Michael, Dr. John, Dr. Ginny, Dr. Demetri, Dr. Justin, Dr. Arthur, Dr. Thomas.
	Faculty identify 50% "no" is something that still needs to be addressed even when 50% achieved "yes."	Dr. William, Dr. Andreas, Dr. Michael, Dr. Ginny, Dr. Demetri, Dr. Justin, Dr. Basilio, Dr. Arthur, Dr. Thomas.
	Faculty identify 73% "no" is something that still needs to be addressed even when 27% achieved "yes."	Dr. William, Dr. Andreas, Dr. Ginny, Dr. Demetri, Dr. Justin, Dr. Basilio, Dr. Arthur.
	Faculty identify 50% "some" is something that still needs to be addressed even when the rest of the 50% achieved "yes."	Dr. William, Dr. Andreas, Dr. Michael, Dr. Ginny, Dr. Demetri, Dr. Justin, Dr. Basilio, Dr. Arthur, Dr. Thomas.
	Faculty identify evenly distributed percentages among "no", "some", and "yes" as something that still needs to be addressed.	Dr. William, Dr. Andreas, Dr. Michael, Dr. Demetri, Dr. Justin, Dr. Basilio, Dr. Arthur.
	Faculty identify 85% "yes" and 15% "no" as a successful performance.	Dr. William, Dr. Andreas, Dr. Michael, Dr. Demetri, Dr. Justin, Dr. Basilio, Dr. Arthur.
Faculty react positively to the feedback.	Faculty react positively towards the feedback report in general.	Dr. William, Dr. Michael, Dr. Ginny, Dr. Demetri, Dr. Basilio, Dr. Arthur, Dr. Thomas.
	Faculty positively react to the recommendation of giving students more opportunities to practice when skills are lacking.	Dr. William, Dr. Andreas, Dr. Michael, Dr. John, Dr. Ginny, Dr. Demetri, Dr. Justin, Dr. Basilio, Dr. Arthur, Dr. Thomas.
	Faculty tend to like real-world scenarios, which have the complexity to it, can help students make sense things better.	Dr. Andreas, Dr. Michael, Dr. Justin.

Table 4.4: *Themes, codes, and faculty representing codes (continued).*

Themes	Codes	Faculty representing codes
The grain-size of the feedback seems to facilitate flexible course modifications.	Faculty's decision to make course modifications depends on how vital the LP is for the course.	Dr. William, Dr. Andreas.
	Faculty tend to pay attention to the *LP* that needs attention based on the student performances.	Dr. William, Dr. Thomas.
	Faculty tend to pay attention to the *expected performance* that needs attention based on the student performances.	Dr. Andreas, Dr. Demetri, Dr. Arthur, Dr. Thomas.
	Faculty prefer customizing recommendations when adopting.	Dr. William, Dr. Andreas.
Faculty identify limitations that the feedback overlooks.	Faculty have concerns about including real-world scenarios in recommendations.	Dr. William, Dr. Demetri, Dr. Arthur.
	Faculty prefer specific recommendations over general recommendations (e.g., example real-world scenarios, what constitutes a logical flow of an argument, example problem, or multiple recommendations that one can keep trying).	Dr. William, Dr. Ginny, Dr. Arthur, Dr. Thomas.
	Faculty suggest a recommendation about explicitly communicating to students about the valued learning goals and its constituent expected performances.	Dr. William, Dr. Thomas.
	Faculty need information about the assessment task in which the learning goal is being assessed.	Dr. Michael, Dr. Justin, Dr. Arthur.
	Faculty would like to have specific feedback that captures student proficiency levels with respect to number of students.	Dr. Michael, Dr. Demetri, Dr. Basilio, Dr. Arthur.
	It's not clear to faculty what the levels of proficiencies communicate to them how students go wrong.	Dr. John, Dr. Ginny, Dr. Demetri, Dr. Justin, Dr. Thomas, Dr. William.
	Faculty do not recognize when there's no "some" category.	Dr. William, Dr. Thomas.

the proficiency required for the expected performance, faculty discussed making significant modifications to a course (see codes #1, 2, 3, and 4 above). According to Dr. William,

> *"Well, the fact that nobody was able to construct the argument tells me that more work needs to be done..."*

This is also the case where distribution is equally distributed among performance levels (see code #5 above). According to Dr. Arthur,

> *"Okay, and so, yes this is clearly something also very important that was all over the place, so it has to be addressed. It is too spread."*

Though faculty discussed making significant course modifications in these instances, they prioritized the expected performance or LP that required the most attention based on students' performance. According to Dr. Demetri,

> *"I don't know, it really depends on the rest of the performance indicators, if that makes sense, you know if the rest of the performance indicators are really bad and this one is, you know, 50 [%] "some", 50 [%] "yes", I'm probably not going to mess with that piece of the course, because there are other areas that are far more in need of change."*

Therefore, after cross examining the expected performance or LP which needs the most attention, faculty might prioritize which modifications to make.

Faculty also determined the instances where no significant course modifications were required (see code #6 above). For example, faculty identified student performance as satisfactory when the percentage distribution of students' performances skewed towards students achieving the full proficiency required for the expected performance. According to Dr. Andreas,

> *"Yeah, it might be a bit optimistic. That will be a pretty good outcome. Maybe, I would expect in my class, more to be like 70, 15, 15, but if I saw this [85% "yes" and 15% "no"], I would be quite happy."*

However, a few physics faculty who were experienced with typically low student-enrollment classes were willing to follow up with the students who did not demonstrate the full proficiency required for an expected performance. This is because they did not like them to be left behind. According to Dr. Michael,

> *"...I could see that as a way that most would do it. I hate to leave 15% behind but, so, I may work with them individually, personally, but yeah."*

This is because with the majority of students achieving the full proficiency required for an expected performance, faculty believed that the student who lagged behind could be problems at an individual scale, rather than the problems associated with the delivery of instruction.

Theme 2: Faculty react positively to the feedback.

This theme consisted of three codes each of which corresponded to instances where faculty showcased their preference towards the feedback, recommendations, and nuance of the provided recommendations. These nuances include providing students learning opportunities that are contextualized in real-world scenarios. The following paragraph describes these three codes:

1. Faculty react positively towards the feedback report in general.
2. Faculty positively react to the recommendation of giving students more opportunities to practice when skills are lacking.
3. Faculty tend to like real-world scenarios, which have the complexity to it, can help students make sense things better.

During the interview, we specifically asked faculty about their perspectives on the feedback provided to them. In general, faculty showcased preference towards the feedback. According to Dr. Ginny,

"But in general, this is more of a general format. 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."

Similarly, Dr. William states that,

"Everything about it I think is the right way to approach assessment for a particular goal and so I think this is exactly what an interested teacher would want to improve their classroom. So, I think, go for it, this is great and I would be very interested in using this kind of thing in my classroom."

Faculty also responded positively towards the researcher-generated recommendations, which would potentially facilitate course modifications. According to Dr. Andreas,

"So in this case, I would want both [recommendations], I would first want to give more opportunities, like the first point says..."

According to, Dr. William,

"Um, I agree completely with the first one. I think they need to be given more opportunities. Usually, repetition is a good way to internalize a concept."

Faculty also positively reacted towards contextualizing student learning in real-world scenarios. They believe that this contextualization can help students make sense of things better. According to Dr. Michael,

"Okay. I like the real-world scenarios...If it's using a real world scenario that they relate to, they understand and grasp, that could help some of the 50% that did not get it to get it there after, so okay, alright."

Similarly, according to Dr. Andreas,

"Because when you have a real-world scenario that has complexity to it, the brain is forced to make sense out of things and put it together. Right, and that can enhance your previous understanding, I think."

Theme 3: The grain-size of the feedback seems to facilitate flexible course modifications.

This theme was built from four codes which highlights the feasibility of the feedback in allowing faculty to flexibly make course modifications. These four codes are:

1. Faculty's decision to make course modifications depends on how vital the LP is for the course.
2. Faculty tend to pay attention to the *LP* that needs attention based on the student performances.
3. Faculty tend to pay attention to the *expected performance* that needs attention based on the student performances.
4. Faculty prefer customizing recommendations when adopting.

As we mentioned before, faculty have the autonomy to choose which LP to assess in their classroom with the TaSPA. Thus, the feedback they will receive will be aligned with the chosen LP. However, faculty prioritized the LP in which a modification was required based on how vital a LP was for the course. According to Dr. William,

"But if that goal [LP] is considered vital to the course, which in my mind it sounds like it might be and should be, then more practice needs to be given to making those arguments, developing those arguments in this particular context."

In addition to that, faculty prioritized the LP that needs the attention based on the students' performance. According to Dr. William,

"I would say that, based upon the feedback that you've gotten from students' performances, this learning goal [LP] needs a lot more attention than the first learning goal [LP] does. Simply because you have more students that seem to be not demonstrating what needs to be demonstrated..."

Further, faculty also paid attention to the expected performance in which the immediate modification to the course was needed when compared to the other expected performances within a LP. According to Dr. Demetri,

> *"I don't know, it really depends on the rest of the performance indicators, if that makes sense, you know if the rest of the performance indicators are really bad and this one is, you know, 50 [%] "some", 50 [%] "yes", I'm probably not going to mess with that piece of the course, because there are other areas that are far more in need of change."*

Moving further, faculty also expressed a desire to customize the recommendations provided when planning on implementing them to their course modifications. According to Dr. Andreas,

> *"So in this case, I would want both [recommendations], I would first want to give more opportunities, like the first point says, and when I get the feeling that has changed the outcome that is in column one, then I will do the second recommendation."*

Theme 4: Faculty identify limitations that the feedback overlooks.

Though faculty liked the feedback, recommendations provided by the researchers, and suggestion of contextualizing student learning in real-world scenarios, they also brought up concerns related to these aspects. For example, they highlighted areas where the feedback could be adjusted to better align with their needs. A few faculty had contrasting opinions about contextualizing student learning in real-world scenarios as it might convolute student learning. All of these concerns and suggestions are grouped into the theme, faculty identify limitations that the feedback overlooks, which is discussed in detail below.

This theme contained seven codes addressing concerns and suggestions faculty brought up when they reflected on the feedback provided. These seven codes are:

1. Faculty have concerns about including real-world scenarios in recommendations.

2. Faculty prefer specific recommendations over general recommendations (e.g., example real-world scenarios, what constitutes a logical flow of an argument, example problem, or multiple recommendations that one can keep trying).
3. Faculty suggest a recommendation about explicitly communicating to students about the valued learning goals and its constituent expected performances.
4. Faculty need information about the assessment task in which the learning goal is being assessed.
5. Faculty would like to have specific feedback that captures student proficiency levels with respect to number of students.
6. It's not clear to faculty what the levels of proficiencies communicate to them how students go wrong.
7. Faculty do not recognize when there's no "some" category.

While some faculty preferred the recommendation of contextualizing student learning in real-world scenarios, others did not prefer to do so. The primary reason for this was the complexity real-world scenarios added when students are trying to make sense of basic concepts. According to Dr. Demetri,

> *"I often find real world scenarios muddy things, when students are trying to understand these really basic concepts."*

Faculty preferred specific recommendations over general recommendations. Faculty preferred to have examples of real-world scenarios in which they can build on. Also, when we provided recommendations rooted in the extent to which students can construct arguments, faculty preferred recommendations attributed specifically to the constructing argument aspect, rather than tailoring that back into concepts. The other suggestions faculty provided on making feedback specific was to provide them an example problem to build on along with recommendations, and a set of recommendations which they can customize. An excerpt to illustrate the specificity faculty asked for is by Dr. Michael, where he states:

> *"...But, it is speaking in generalities. If there is any way to say, hey, here is a real-world scenario, maybe give me two or three or four real-world scenarios,*

and saying, yeah, you could do this concept, this example, these, whatever. That would be beneficial. Give me something concrete that I could use. That would be the biggest benefit and then I would be not trying to think, because if you were to say, come up with a real-world scenario, I would be going...what, you know, and I would be struggling to find one. So, if you go a step further, and find a few examples of those, that would be a big help."

Some faculty also suggested a recommendation to include in the feedback, which is to suggest faculty explicitly communicate students the valued LPs and its constituent expected performances. In that way, faculty believed, rather than only giving students more opportunities, students would realize what faculty expect them to be proficient at. According to Dr. William,

"Not only just practice, but maybe like I said, emphasis that this is a learning goal [LP] and this is how we break it down and this is what will be expected of you. So, students understand that this is how I would approach it."

Also, the feedback presented to faculty did not include the assessment task in which the student learning was being assessed. The absence of the assessment task in the feedback was something faculty pointed out during the interview. According to Dr. Thomas,

"I would want to know also what the items were that students were working on, to know whether it is like difficult, whether it looks difficult slash tricky or like standard."

Faculty preferred to have feedback which was narrowed down to the information relating to students. Faculty who were familiar with teaching in small-enrollment classes liked to receive feedback that contained information related to individual students. The other reason why faculty wanted information relating to the number of students who took the test was to infer the significance of the students' outcomes to modify the course. According to Dr. Basilio,

"Okay, so, again, I'm operating under the framework of small classes. So, it's very easy to do global solutions for problems that sometimes deal with individuals. If you give me information on who were the 50% and who were the other 50%, I would probably follow up with the students..."

Faculty brought up a concern about the lack of information provided by the feedback about how students went wrong in their solutions. The faculty who brought up this concern would like to see how their students were struggling to achieve the expected performance, and the nuance of those struggles, to help navigate them to the right path. According to Dr. John,

"...okay, "some", okay, but what is the evidence? I need to see what are they saying, what are they making, how are they making mistakes, how are they showing confusion or understanding..."

Additionally, faculty had a hard time recognizing when there was no "some" category presented in the feedback. It seemed that the lack of recognition of this might reduce faculty's ability to reflect and make meaningful inference about the students' outcomes and its significance on making course modifications. According to Dr. Thomas,

"So, in your feedback, if there's no way that a student could get some, I would like to see that indicated on the form...But, if one of those up a column is not a possibility, in N/A there or something would help me interpret it more accurately, so that I wouldn't worry that the "no" means that my students are in serious trouble with, you know, in distinction from a "some" answer."

Step 7: Update the feedback

In this subsection, we answer our last research question that we laid out in Sec. 1: *How can we incorporate perceptions of the faculty to improve the generated feedback?*

Based on the four themes emerged from the analysis, we claim that the feedback provided to faculty include elements that they identified as important with the utility of that in the

classroom. At the same time, we claim that there is room for improvement to better align the feedback with faculty members' needs. Thus, in the following paragraph, we explain what we keep including in the feedback and the modification that can be made to it, incorporating faculty's perspectives. We note that the order the explanations would appear is based on the first, second, and third themes, where faculty's perspectives contributed to what works well for them in the provided feedback. Finally, we provide an explanation on the fourth theme, where faculty identified limitation that the feedback overlooks.

We first expand the first theme presented in Sec. 4.2.2. Depending on the extent to which students achieved an expected performance, which is the current performance, faculty determined course modifications as required. If the percentages of student population skewed more towards performance-not met level, faculty identified that as an instance where significant changes to a course are required, which is also aligned with our interpretation of when to make a course modification.

On the other hand, if the percentages of student population skewed more towards performance-met level, faculty identified that as an instance where no significant changes to a course is required, which is also aligned with our interpretation of when to make course modification. However, one exception here was the need raised by faculty who have experience teaching in small student-enrollment classrooms. They were interested in seeing student-level information to follow-up with the student population who lagged behind, though many students achieved the expected performance already. However, our goal was not to evaluate student-level information, rather course-level information. Thus, to address this need, we will provide the total number of students who represent the percentage (e.g., 72% ($N = 24$)). Addressing this need also addresses a need faculty brought up under the third theme in Sec. 4.2.2. This was also relating to requesting information about the number of students occupied for each performance level.

Within the second theme which we provided in Sec. 4.2.2, we explained how faculty reacted positively to the feedback provided to them. They liked the feedback in general, recommendations provided to them suggesting course modifications to better facilitate student learning, and contextualization of student learning in the real-world contexts. However, one

exception to this was the concern brought by faculty about situating student learning in real-world contexts as we explained under the forth theme in Sec. 4.2.2. Some faculty believed that the inclusion of real-world contexts can potentially make student learning convolute, when students learn basic concepts.

In the third theme which we explained in Sec. 4.2.2, we articulated how the grain-size of the feedback seems to facilitate flexible course modifications for faculty. This flexibility was advocated through the grain-size the feedback was designed with. For example, providing them with feedback aligned to each LP, each expected performance within that LP, and multiple recommendations for each expected performance allowed faculty to customize the feedback to adopt based on their priorities, when provided. Thus, we continue to keep the grain-size the feedback represents, allowing faculty to customize the feedback to modifying a course based on their priorities.

In the fourth theme articulated in Sec. 4.2.2, we described the limitations faculty identified. Faculty identified real-world scenarios as adding complexity to student learning as students try to understand basic concepts. We, however, note that the inclusion of real-world scenarios in the recommendations provided to faculty was a deliberate choice that is linked to the knowledge-in-use perspective. Student's knowledge-in-use is characterized through their ability to apply knowledge into novel situations such as real-world scenarios. This is in part addressing the goal of the TaSPA via feedback which ultimately helps promote classrooms that provide students with opportunities to nurture knowledge-in-use. This goal behind contextualizing student learning in real-world contexts will be provided to faculty as a primer for the TaSPA. We note that depending on the goals of a particular assessment, the nuance of the recommendations can vary, and thus faculty might react differently.

Faculty preferred to have specific recommendations as opposed to general recommendations. The specificity faculty referred involves having example real-world scenarios and problems, and the assessment context in which the student learning was evaluated. However, we note that providing minimal information about such contextual information was a deliberate choice we made. We believe that providing these information to faculty would encourage "teach to the test" as opposed to promoting authentic learning for students. If

the contextual information are familiar to students when engaged with the assessment, it does not support the goal of the TaSPA, which is to evaluate the extent to which students can apply their knowledge to novel situations.

Faculty also brought up the idea of being specific about what constitutes an argument as opposed to blending argumentation with physics concepts. However, we note that blending argumentation with physics concepts was intentional. This is to encourage scientific argumentation in the presence of physics concepts, rather than in isolation. Thus, this suggestion left us with making no modifications to the feedback. Another angle to the specificity faculty brought up was about including a set of recommendations in which faculty can customize. This will be something that we can add to the process of designing feedback for faculty.

Faculty suggested including a recommendation to faculty about faculty explicitly communicating to students about the valued LPs and its constituent expected performances. That way students are also aware of what has been valued about learning by the faculty. Though we consider this as a great suggestion, we believe this provides evidence of how feedback enabled faculty to reflect on their practices and think about the ways student learning can be facilitated, rather than a modification that needs to made to the provided feedback. Thus, we do not make any changes to the feedback.

Faculty identified the need for information relating to the assessment task. When feedback is provided to faculty, we made a deliberate choice for not providing the assessment task in which the student learning was assessed. This is because providing assessment tasks could let faculty "teach to the test," rather than providing students opportunities to acquire the knowledge, skills, and abilities required to achieve the expected performance. In addition to that, if faculty used the same assessment task as an example during the instruction, the context the learning being assessed will no longer be a novel context to students. Thus, it is less likely for faculty (or researchers) to make meaningful inferences about students' abilities to use their knowledge into novel situations, when presented.

Faculty showcased a need of information on individual-student scale. Not providing information on the scale of individual-student in the feedback was intentional. This is because the purpose of the TaSPA is to assess the course (and thus, of the instruction and curriculum),

but not the students.

Faculty brought up the concern of the provided feedback having a lack of information about how students go wrong. Faculty stated that knowing how students went wrong can enable them to help students learn better. This is due to the lack of clarify associated with the "yes," "some," and "no" categories provided to faculty in communicating students' performance levels. We believe that this concern can be addressed via framing these three categories in an asset based point of view. For example, along with providing faculty "Students did not meet the expected performance" when it is related to "no" category, we can also provide the common approaches students followed that resulted in receiving "no" category.

It seemed like the feedback failed to communicate when there is no performance level associated with the expected performance. Agreeing with one of the faculty's suggestion on placing "N/A" when there is no performance level attributed with an expected performance, we will update our feedback for faculty accordingly.

4.2.3 Inter-Rater Reliability

The processes during each step associated with the design of feedback for faculty in Fig. 4.3 were discussed with other researchers, who have experience in developing knowledge-in-use assessments for upper-division along with feedback. We carried out discussions until we came to a consensus.

4.3 Discussion

One of the primary goals of this work was to articulate a methodology to provide actionable feedback for faculty which can guide course modifications, facilitating better learning opportunities for students. We conducted this study in parallel to developing a new research-based assessment under development, i.e. TaSPA. The assessment tasks in the TaSPA take the CMR format to promote online test administration with streamlined evaluation of student work. Thus our feedback development methodology provided in Sec. 4.2 was contextualized

in the students' responses to tasks in the CMR format. We embedded features that can strengthen faculty's learning into the feedback design process, envisioning them having the autonomy to make modifications to their courses.

We note that the feedback development methodology introduced in this work has broader implications to the physics education research community, especially the methodology behind including explicit feedback for faculty in the research-based assessment design. We acknowledge that providing a methodology for the assessment design was not the scope of this work. However, we note that developing feedback for faculty is embedded within the early stage of the process of developing a research-based assessment, and thus, if valued, should be paid attention to at an early stage, rather than later.

Identifying LPs, KSAs, and corresponding ESs which provide observable features of student work to make conclusions about the extent to which students achieve LPs are crucial steps of the process of developing assessments. From the feedback development perspective, ESs lay out the foundation to characterize student performance levels, which amalgamate information relating to the extent to which students achieve a LP. Thus, the utility of such a characterization bridges students' outcomes with the suggestions that can be provided to faculty via researcher-generated feedback, that would potentially facilitate students to better achieve a LP.

We demonstrated the process of designing feedback for faculty which is rooted in the CMR assessment task format. Though that is the case during the feedback generation process, we used the nuance of student work in response to both FR and CMR assessment tasks. The creation of the CMR rubric was rooted in the performance levels identified in students' responses to the FR task. Such a methodology was promoted due to the CMR task being built from its corresponding FR task.

We emphasize that providing the methodology for designing feedback as rooted in both FR and CMR assessment tasks has implications for physics education research practitioners. For example, if one wants to design feedback rooted in FR assessment tasks, which is the task structure broadly used in physics classrooms, the methodology provided in this work can be adopted to serve that purpose. On the other hand, if one wants to move a step

further and design feedback rooted in CMR assessment tasks, the format in which has the potential for online administration with streamlined evaluation of student work, that is also viable with the methodology presented in this work.

With the inclusion of feedback for faculty within the TaSPA, we elevate its potential to explicitly cause changes to instructional practices of faculty members. For example, as we explained in Sec. 4.1.1, feedback (i.e. a component of the assessment) provides actionable information (i.e. action mechanisms) for faculty, if uptaken, can improve student learning (i.e. an intended effect). Thus, we encourage assessment developers to include explicit feedback for faculty within their assessment design, which would explicitly attend to changes in the individuals the assessments target.

We note that the choice of using LPs was deliberate to serve the purpose of the TaSPA, which is to transform upper-division classrooms to promote students' application of knowledge, rather knowledge in isolation. However, based on the designer's goal for the assessment, assessment developers can adopt what they really value into the design of LPs, KSAs, and the corresponding ESs, and yet follow the same process to design feedback. We remind the readers that this whole process of designing feedback is rooted in ECD, which provides the theoretical basis for developing LPs, KSAs, and ESs. Given the broader utility of ECD to characterize student learning, the same approach we used for the feedback development can still be used by other assessment developers, with what they value in student work differing to what we value in the context of the TaSPA.

We remind that the other prominent goal of this work was to incorporate faculty's perspectives into the feedback design process. This is an important step to ensure how faculty would react to the provided feedback, especially given that the form of feedback we developed is novel. Though we conducted semi-structured interviews to obtain perspectives from the faculty, we note that others can use approaches differing to what we used. One such example is, informal discussions where researchers and faculty can get together to discuss the developed feedback. Providing faculty with sample feedback would allow them to reflect and provide their perspectives on the sample feedback efficiently. We note that enabling in-the-moment reflections on the provided feedback during the interviews, rather providing

them prior to the interviews would closely simulate faculty's reflections when the feedback is received in real-time.

Analyzing and interpreting interview data can reveal faculty's perspectives on the feedback provided to them, especially the aspects that better support course modifications as well as the aspects that need improvements. For example, we used thematic analysis as a data analysis method with inductive approach to explore the nuance of faculty's perspectives. The themes articulated can inform what faculty valued as important in the provided feedback and what can be modified to better support their needs.

Thus, we emphasize the utility of the faculty interviews in improving the quality of the feedback. Conducting faculty interviews by providing the developed feedback for faculty to reflect on, analyzing these interviews, and incorporating the results to inform and update the developed feedback are important steps. Though LPs, KSAs, and ESs one uses could be different than the ones we used in the context of the TaSPA, the methodology provided in this work can be incorporated similarly. In the following paragraphs, we explain how the methodology presented in this work can be adopted by others to design feedback in parallel to developing their respective assessments.

Step 0: *Gather necessary information*

Developing feedback for faculty is conducted in parallel to developing a FR assessment task. Thus, there are several components that are necessary to adopt from the task development stage. These include a learning goal, set of skills expected from students to demonstrate the achievement of that learning goal, and the observable evidence to claim that students have these skills. The students' work in response to the developed task is also needed to gather evidence. Collecting such evidence can be done in numerous ways. For example, it is possible to provide the assessment tasks to students in an exam setting and then the students' written work can be gathered. It is also possible to ask students to go through the assessment task and reflect their thoughts out loud while going through the task[118]. This can be conducted in a video and audio recorded interview setting and the corresponding

written solutions can also be gathered.

Step 1: *Identify performance levels from students' responses to FR task*

Analyzing students' work with respect to the evidence identified as required in step 0 can accommodate students' common problem solving patterns in response to the assessment task. These patterns demonstrate the richness of students' knowledge and provide evidence for their abilities to achieve a learning goal. Each unique pattern can be assigned with a performance-level such that the variety of approaches students took when responding to an assessment task can be captured through these performance levels.

Step 2: *Create feedback statements rooted within each performance level identified in step 1*

The feedback for faculty is rooted within each performance level. This bridges information relating to the extent to which students achieve a required skill, with opportunities faculty can provide students to achieve the required skill (if needed) via feedback. Such feedback can be aligned with the three features articulated in Sec. 4.1.3, which can support faculty's learning process as we unpack next. We first provide the required skill that student should have. When students' work is characterized through evidence required to showcase that skill, it provides information relating to the extent to which students achieved that skill. Thus, secondly, faculty are provided with this current state of the student achievement. If there is any gap between the current and expected skill, faculty are thirdly provided with recommendations that facilitate modifications to their instruction to better facilitate students to achieve the required skill. These recommendations can be created by individuals who are expert in the subject matter the assessment task contains.

Step 3: *Map performance levels and corresponding feedback developed in step 2 to a CMR rubric.*

The performance levels identified during step 1 and feedback aligned with these performance levels can be mapped to a corresponding CMR rubric. The students' selections appropriate to achieve each performance level are also listed in the CMR rubric. Faculty will receive feedback based on the selections students made in response to the CMR task.

Step 4: *Create feedback for faculty based on students'responses to the developed CMR task*

At this step, it is important to explore whether the feedback can be generated after students responded to the CMR task. This process enables to explore whether CMR rubric can be operationalized, where each response pattern of the students can be characterized into a performance level. If the CMR rubric functions as intended, then the associated feedback can also be generated based on the selections students made on the CMR task.

Step 5: *Obtain perspectives of faculty about the developed feedback*

We identify obtaining perspectives from faculty about the developed feedback as an important step. This ensures that the researcher-generated information are interpreted accurately by faculty. Additionally, this step ensures the utility of recommendations to be incorporated into faculty's instructional practices. These perspectives can be obtained in numerous ways. Semi-structured interviews, or informal discussions can be a few examples to obtain perspectives of faculty. These interviews or informal discussion can include questions to elicit thoughts and views of faculty when they are provided with sample feedback.

Step 6: *Articulate key aspects to update feedback*

Analysis of faculty interviews can reveal the faculty's interpretation of the researcher-generated feedback. One way to analyze faculty interviews is to conduct a thematic analysis

with the inductive qualitative approach. The themes generated out of this analysis can inform the modifications that can be made to the feedback by incorporating needs and preferences of faculty.

Step 7: *Update the feedback*

The themes generated in step 6 communicate the perspectives faculty brought up during the interview. These perspectives can capture both what works well in the feedback and the improvements needed to align with the needs of faculty. Thus, these themes can inform whether or not the feedback needs to be updated.

When designing feedback for faculty, it is also important to consider their unique approaches to course modifications. In the next chapter, we provide two case studies of physics faculty to explore the design of feedback that can support their course modifications.

Chapter 5

Designing Feedback for Physics Faculty Supporting their Course Modifications: Two Case Studies

In this chapter, we provide two case studies that provide insights into similarities and differences of practices and processes internal to two faculty members – Dr. William and Dr. Andreas – take in the context of modifying physics courses. These two case studies answer the following research questions.

1. How do faculty discuss modifying their course?
2. What are the types of external feedback faculty receive, and how do they incorporate that feedback into modifying a course?
3. How might we incorporate experiences of faculty associated with course modifications into designing researcher-generated feedback?

In Sec. 5.1, we provide information about the model of self-regulated learning in the context of feedback, which is the theoretical framework used in our study. In Sec. 5.2, we provide the data collection and selection, and data coding approaches pursued in our study, following two case study analysis in Sec. 5.3. A synthesis of these two case studies can be

found in Sec. 5.4. In Sec. 5.5, we provide the features of the external feedback that can be supportive of Dr. William and Dr. Andreas.

5.1 Theoretical Framework

We use a model of self-regulated learning in the context of external feedback as the lens to characterize the "processes internal to an individual" and "role of external feedback" when an individual engages in an activity[2;79;133]. Figure 5.1 provides the elements that constitute this model. We deliberately choose this model due to its explicit intersection of processes internal to an individual with external feedback. This explicit intersection enables us to identify features of the external feedback that can support the internal processes of an individual when they engage in an activity. The other reason why we chose this model is due to the dynamic nature the internal processes can take when engaging in an activity. Instead of considering internal processes as static, this model advocates for the iterative adjustments an individual makes during the activity. Thus, this iterative nature leads to an individual's learning by bringing, utilizing, and restructuring their existing views and practices.

In the following paragraphs, we explain each element that constitutes the model depicted in Fig. 5.1 by pulling out information from Refs.[2;79;133]. In parallel, we explain how each of these elements can be used in the context of a faculty member modifying a course. Thus, a "learner" in our context is a "faculty member" who undergoes a learning process while "modifying courses." The "external feedback" is the one that faculty themselves interpret or consider as external to them. Each of these explanations and associated examples is summarized in Table 5.1 and Table 5.2.

When a learner engages in an activity, they draw on their knowledge and beliefs to have a personal interpretation of the activity. An example for this can be the knowledge and beliefs a faculty member draws on when they start the activity, which is "modifying a course." The faculty member can believe that the physics courses can and should teach skills such as communication in addition to the physics content. Then the faculty member can draw on knowledge relating to the ways in which their course can be structured to facilitate students

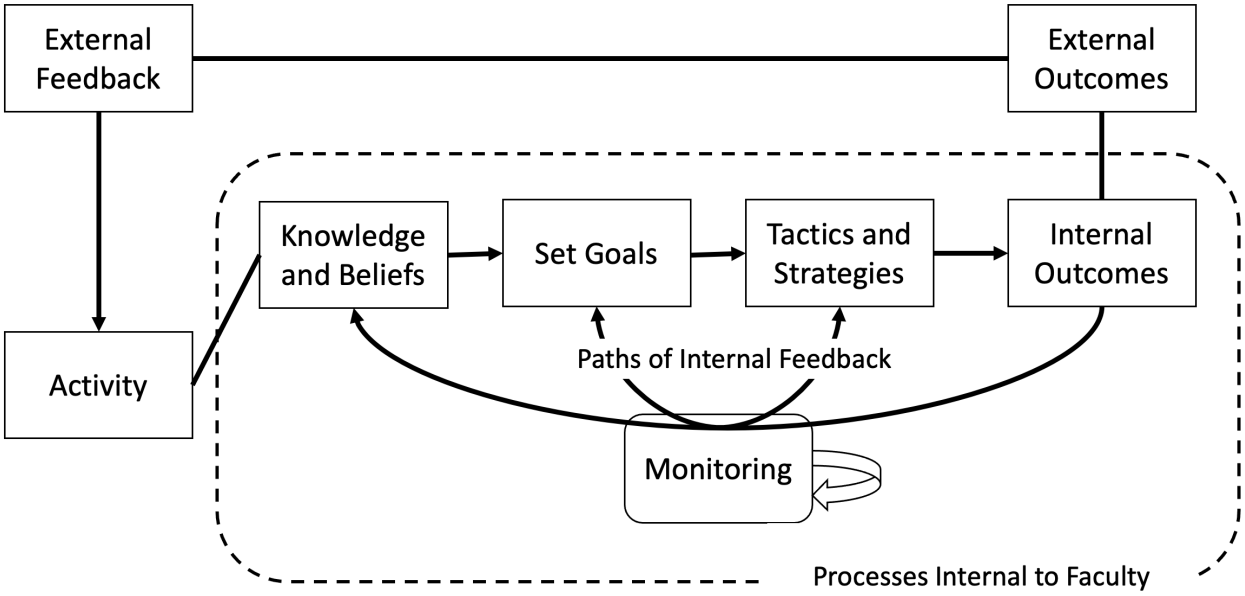

Figure 5.1: *A model for self-regulated learning in the context of external feedback. This figure is recreated from its original version presented on Butler and Winne*[2]*. The "Task" and "Cognitive System" from the original version were changed to "Activity" and "Processes Internal to Faculty" respectively, to align with the context of our study. In addition to align with our context, "Task" was replaced with "Activity" to reduce the potential confusion between "Task" and "Assessment Task." The "Products" and "Performance" from the original version were changed to "Internal Outcomes" and "External Outcomes" respectively, for effective communication.*

to practice communicating ideas to the general public.

After drawing knowledge and beliefs, the learner set goals for themselves. An example for this can be the phase where the faculty member sets up explicit goals by building on the knowledge and beliefs about improving students' communication skills. This goal setting phase explicitly attends to the specifics about the nature of the learning, the faculty member expects their students to gain through practicing communication skills. For example, the goal can be "course should be modified such that students should be able to communicate to the general public about the concept of "entropy.""

To achieve the goals, the learner draws on tactics and strategies in the context of the activity. For example, to help students learn how to communicate to the general public about the concept of "entropy," the faculty member can allocate some of their lecture time for students' presentations. Students can be encouraged to create their presentations with a

Table 5.1: *Descriptive information of each element in Fig. 5.1 and the examples for each element in the context of physics faculty modifying their courses.*

Model Elements	**Definitions**	**Examples**
Activity	The activity a learner engages in.	Modifying a physics course by a faculty member.
Knowledge and Beliefs	The knowledge and beliefs the learner draws on to interpret the activity.	The faculty member believes that the physics courses can and should teach skills like communication in addition to the physics content, and they draw on knowledge related to the ways in which the course can be structured to help students gain such skills.
Set goals	The goals the learner has for the activity.	Course should be modified such that students should be able to communicate to the general public about the concept of "entropy."
Tactics and strategies	The tactics and strategies the learner uses to achieve the goals.	The faculty member plan on allocating some of their lecture time for students' presentations.

non-scientific audience in mind.

As the learner uses the tactics and strategies, they produce two types of outcomes: the internal and external outcomes. Internal outcomes are the outcomes internal to the learner. These outcomes include, but are not limited to the changes in the cognitive or emotional states of the learner. An example for this can be the excitement the faculty member had when seeing their students learn communication skills better, based upon the students' presentations they implemented. External outcomes are the outcomes the learner produced that are visible to others. These include both tangible outcomes (such as lecture notes) and behavioral outcomes (such as lecturing). An example for the tangible outcomes can be the updated syllabus of the faculty member by allocating time for students' presentations. An example for the behavioral outcomes can be that the faculty ask students to create and perform the presentations.

As the learner progresses through the activity and produces outcomes, they start monitoring their progress by comparing outcomes to the goals. This generates the internal feedback.

Table 5.2: *Descriptive information of each element in Fig. 5.1 and the examples for each element in the context of physics faculty modifying their courses (continued).*

Model Elements	**Definitions**	**Examples**
Internal Outcomes	The outcomes that are internal to the learner (changes to cognitive or emotional states of the learner).	Excitement the faculty member had when seeing their students learn communication skills better, based upon the students' presentations they implemented.
External Outcomes	The outcomes that are tangible (e.g., lecture notes) and behavioral (e.g., lecturing).	Updated syllabus (tangible) of the faculty member by allocating time for students' presentations and the faculty member asking students to create and perform the presentations (behavioral).
Monitoring	The process in which the learner compares the outcomes with respect to the goals.	Checking whether students' skills associated with communicating the concept of "entropy" to the general public are improving.
Paths of internal feedback	The monitoring process leads to generating paths of internal feedback (i.e. internal feedback can guide the learner to reinterpret the task and/or draw on different knowledge and beliefs and/or set different goals and/or use different tactics and strategies, and/or produce new outcomes).	The faculty member recognizes that they need to include a mechanism to give feedback on students' communication (such as the recognition to include a discussion session after students' presentations – re-evaluate tactics and strategies) to further clarify how to better communicate the concept of "entropy" to the general public.
External feedback	Feedback received externally on the performance in comparison to external standards.	The faculty member can receive feedback from the department that the students' presentations take a huge amount of time, preventing them from covering the intended course material.

This internal feedback, if needed, can cause the learner to reinterpret the activity, draw on different knowledge and beliefs, use different tactics and strategies, set new goals, and/or produce new outcomes as they make progress through the activity. An example for this can be the case where the faculty member recognizes that they need to include a mechanism to give feedback on students' communication (such as the recognition to include a discussion session after students' presentations) to further clarify how to better communicate a concept

to the general public.

On the other hand, the external outcomes the learner produced, that are visible to others, can receive external feedback. This external feedback can also cause the learner to reinterpret the activity, draw on different knowledge and beliefs, use different tactics and strategies, set new goals, and/or produce new outcomes. For example, the faculty member can receive feedback from the department that the students' presentations take a huge amount of time, preventing them from covering the intended course material. This feedback can cause the faculty member to re-evaluate the modifications made to their course.

5.2 Methodology

5.2.1 Data Collection and Selection

We conducted one-on-one, semi-structured interviews with faculty who teach intermediate/upper-division thermal physics courses. The interviews were conducted through Zoom, and video and audio recorded. We used in-built transcription option of Zoom and corrected the transcript later for clarity. Each interview lasted about an hour. The interview protocol consists of three broader aspects (see Sec. B.1 for the full interview protocol). Faculty were asked about their practices associated with their course modifications, perspectives on researcher-generated learning goals, and perspectives on researcher-generated feedback aligned with two learning goals. For this study, we used information gathered from the first aspect targeted during the interviews, which informs the nuance of faculty's approach to modifying their courses.

To elicit the processes internal to faculty when modifying courses and role of external feedback, our interview protocol included several open-ended questions. For example, these questions include "how, if at all, do you go about course modifications to your own course?", "can you remember a time when you got feedback from external sources such as...," and "overall, what does effective feedback look like to you." To elicit in-depth nuance of faculty members' experiences within each of these broader questions, our interview protocol

included a set of follow-up questions. These follow-up questions included, but were not limited to "what motivates you to do those modifications?", "how often do you do course modifications," "what determines that you receive external feedback?", and "what features of that feedback were most helpful for you?". The interviewer asked questions improvised during the interview for either clarifications purposes or to further elicit faculty's responses to questions.

In this study, we provide two case studies – Dr. William and Dr. Andreas – to explore the role of feedback that guide faculty's ongoing learning processes when modifying courses. The demographic information of these two case studies are provided in Table 5.3. Case studies provide rich and in-depth perspectives into individual and situational characteristics which allows us to explore similar and different roles the external feedback can take when facilitating faculty members' course modifications. Analyzing these case studies lay out a preliminary approach that can guide extending this work to a large scale project in developing external feedback for physics faculty[28;140].

We chose to focus on these two cases because these two faculty members went into significant depth about their course modification process, making it easier to apply the model in Fig. 5.1. We also note that the questions in our interview protocol were informed by this model. For example, including questions that specifically asked about motivation of faculty to do course modifications, whether or not faculty measured the success of course modifications, and whether or not faculty receive external feedback, was to elicit crucial aspects of the model of self-regulated learning and feedback provided in Fig. 5.1.

Table 5.3: *Demographic information of Dr. William and Dr. Andreas. All the names provided are pseudonyms. They reported, an introduction to thermal physics by Schroeder[4] as the textbooks that they use in their classrooms. The number of students presented here are the average number of students in classrooms, as they reported. The acronym, PWIs corresponds to Predominantly-White Institutions. The race/ethnicity and gender were self-identified. In the survey form, race/ethnicity was provided as "Caucasian/White."*

Pseudonym	Textbook	# Students	Institution	Race/Ethnicity	Gender
Dr. William	Schroeder	5-22	PWI	White	Man
Dr. Andreas	Schroeder	10-15	PWI	White	Man

5.2.2 Data Coding

We used the method for coding suggested in Braun and Clarke[139] as the data analysis approach to characterize interview data. When coding, we used the model of self-regulated learning in the context of feedback (see Fig. 5.1) as the lens to characterize the learning process of faculty associated with modifying courses. Table 5.1 and Table 5.2 helped clarify what specifically we looked for in data when operationalizing the terminologies in Fig. 5.1 into our context. For example, the "Activity" corresponds to "modifying the thermal physics course by faculty" in our context. While analyzing interview data, these interpretations were modified to accommodate characteristics that appear in data. We provide detailed description of these interpretations when we bring up case studies in Sec. 5.3.

The two case analyses provided in this chapter captured about the first 30 minutes of the interviews. We watched the recordings several times to familiarize ourselves with the data. While watching, we corrected the transcript for its clarity, as Zoom in-built transcription sometimes does not capture pronunciations or jargon accurately. We started coding the data by identifying and labeling features of the data, i.e. codes. This coding process was informed by the first two *research questions* listed in the beginning of this chapter:

1. How do faculty discuss modifying their course?
2. What are the types of external feedback faculty receive, and how do they incorporate that feedback into modifying a course?

The coding process was also informed by the *definitions* provided in Table 5.1 and Table 5.2. This coding process has been conducted in the order of the responses appeared in the transcript, following questions by the interviewer.

A code was assigned to a block of text on the transcript. Each code was labeled aligning to one of the elements of the model provided in Fig. 5.1 and its definition on Table 5.1 and Table 5.2. For example, we labeled the following block of text as "set goals," which refers to an instance where a faculty member discussed their goal of creating a classroom which is active.

"I try to make sure that I have a classroom that is fairly active. I don't wanna simply be talking to them."

We highlighted each code with a unique color to differentiate codes from each other. The rightmost column in Table 5.1 and Table 5.2, which operationalize the terminologies in Fig. 5.1 into our context, were simultaneously modified to incorporate ideas and views presented in the data.

After codes were assigned, while watching the recordings, we read the transcript until the next block of text that is related to the research questions provided above and definitions provided in Table 5.1 and Table 5.2, was found. We then determined whether a same code can be applied or a new code was needed to capture that block of text. If a new code was assigned, the rightmost column in Table 5.1 and Table 5.2 were iteratively modified to incorporate ideas and views presented in the data. If the same code was assigned, the rightmost column in Table 5.1 and Table 5.2 were adjusted to capture codes with similar features. This process was repeated for both Dr. William and Dr. Andreas's interview data.

We compared and contrasted codes associated with the knowledge and beliefs, goals, tactics and strategies, and outcomes of Dr. William and Dr. Andreas in the context of their discussions related to course modifications. This led us to explore the similarities and differences of the nuance of their experiences associated with modifying courses. We compared and contrasted the monitoring processes of Dr. William and Dr. Andreas and how those processes led them to re-evaluate their knowledge and beliefs, goals, tactics and strategies, and outcomes. We made a note of the timescales of Dr. William and Dr. Andreas's learning processes during such re-evaluations.

We also compared and contrasted the codes associated with the external feedback Dr. William and Dr. Andreas received and how they incorporated that feedback into course modifications. This led us to explore the similarities and differences of the impact the external feedback can make on the course modifications made by Dr. William and Dr. Andreas. We made a note of the re-evaluation of the knowledge and beliefs, goals, tactics and strategies, and outcomes caused by the external feedback Dr. William and Dr. Andreas

received. We also made a note of the timescales for such re-evaluations for both Dr. William and Dr. Andreas.

5.3 Case Study Analysis

In this section, we provide case study analysis of Dr. William and Dr. Andreas. For both of these case studies, we extracted the illustrative quotes with assigned codes, and created narratives that unpacked both faculty members' experiences when modifying courses. The elements of the model in Fig. 5.1 as they appear in each case study are noted in parenthesis within quotes.

5.3.1 Dr. William's Case Study

When we asked Dr. William about the modifications he made to his thermal course, he started describing the readings he did on the best teaching methods for physics. This indicated Dr. William's willingness to improve his teaching by seeking knowledge on the best teaching methods for physics, while believing in them (knowledge and beliefs).

> *"Well, I read a little bit about the best, obviously the best teaching methods for physics (knowledge and beliefs)."*

Dr. William wanted his classroom to be an active learning environment for his students, rather than simply talking to his students (set goals).

> *"I try to make sure that I have a classroom that is fairly active. I don't wanna simply be talking to them (set goals)."*

To facilitate active learning environments for students, Dr. William used several teaching methods in his classroom. He shifted away from having a pure lecture-based course to one that includes but is not limited to mini lectures, readings, and in-class assignments (tactics and strategies), and implemented these strategies in his classroom (external outcomes). Dr.

William's continued willingness to improve his teaching led him to expand his knowledge about other teaching methods that can be used in his classroom. He planned on letting students work on a module and take quizzes until they achieved a certain score on that, which he referred to as the mastery method (tactics and strategies). He had been trying to use this new teaching method in his classroom (external outcomes).

> *"So, over the past decade or so I changed it from a completely lecture-based course to one that is more of a course (external outcomes) where there are mini lectures, and there are readings, and there are in class assignments (tactics and strategies) and all of those modifications come about over time as students give me feedback on course evaluations. And as I find something out there in the literature that says, oh why don't you try this, and so I go around. Okay, I'll try that. Most recently I've been trying to use a course which is based upon the achievement of a certain level (external outcomes). So, the students will do a module and then take quizzes until they achieve a certain score on that. So, the mastery method, basically (tactics and strategies)"*

However, in every modification he planned on implementing, Dr. William himself had his own uncertainties about whether that modification was going to work (internal outcomes).

> *"And every modification I make, you know, you do that, as you bite your lip and hope that everything is going to work. And cross your fingers and whatever other superstitions you may have to make something from nothing is what you're doing when you teach a class when you create the method that you're going to teach it. Lecturing is easy, teaching is more difficult (internal outcomes)"*

Dr. William attended to students' feedback to improve his teaching methods. He reflected on the end of the semester students' evaluations (external feedback) to evaluate the mastery method he implemented. Reflecting on the students' evaluations led him to go back and inspect why things didn't work out as expected (re-evaluate tactics and strategies).

"Well, if I get a lot of students and by a lot, you know, if I only have five students a lot could be three. But, if I get enough students telling me that something just didn't work for them (external feedback), I'm going to go back and look at what's happening and see why it didn't work (re-evaluate tactics and strategies)."

Upon receiving students' evaluations, Dr. William had his own beliefs about why the mastery method didn't work out as intended. He assumed that students were not prepared for that method yet. In particular, he believed his students, being juniors in college, were less prepared for the mastery method, while the method would have been worked if students were seniors (re-evaluate beliefs).

"And it doesn't always work because I don't think the students are prepared for it being juniors in college, and perhaps it would work if they were seniors (re-evaluate beliefs)."

Students' evaluations led Dr. William to modify his course. For example, in response to students' evaluations, he planned on (re-evaluate tactics and strategies) and implemented an approach where students read papers related to thermal physics and wrote a summary of their readings (external outcomes). However, Dr. William's continued reflection on students' evaluations made him aware that the implemented approach on reading and summarizing papers wasn't working for his students. Instead, students suggested the approach would have been better if they were able to talk, rather than writing a summary (external feedback). This suggestion led Dr. William to modify his course to create a half hour session every third Friday for his students to simply gather and discuss the papers they have read (re-evaluate tactics and strategies), and he implemented this new approach in his classroom (external outcomes).

"One of the things I try to do is I try to introduce (external outcomes) some primary literature so they have to read one or two papers in thermal physics or related to thermal physics (re-evaluate tactics and strategies), and in the past, that wasn't working, just reading it and having them write up a summary of it

or something (external feedback). So, what I instituted (external outcomes) based upon their feedback on them said it wasn't working, it might be better if they were able to talk about it rather than write about it (external feedback). So, we created this (external outcomes), you know, every third Friday we have a half hour session where we simply talk about the articles that they've read. And it's sort of like a literature review in person (re-evaluate tactics and strategies)."

Dr. William eventually found out that students preferred the discussion sessions he implemented (external feedback).

"And I think it works better now. They get more excited about what they're reading, if they get to share it with their fellow students as opposed to just sharing it with me, and that came directly from the feedback that I received from course evaluations (external feedback)."

Though Dr. William relied upon the students' feedback on course modifications, he also acknowledged his authority to make changes to the course on his terms, rather than on students' terms.

"And if they say, this isn't working and other students say yeah, this isn't working, then I'm going to make a change. And so I'll make that on the fly if necessary. Of course I always couch that in the fact that I have a lot more experience in teaching than they do. At least I hope I do, and so I'll make the changes on my terms, rather than their terms, but like last year last fall, it wasn't working this module method, this mastery method. And so we altered it so that it could work with them in an online format."

In addition to students' evaluations, Dr. William believed that the student learning should be improved based upon the modifications he made to his course (beliefs). Dr. William wanted to evaluate whether students understood the concepts and/or used the correct methods to solve the problems (set goals). For this, he used certain questions marked

on quizzes (tactics and strategies) to track students' scores over time (external outcomes). Dr. William reflected on the students' scores in response to these selected questions to evaluate whether or not there was an improvement of student learning (monitoring). Though he saw an improvement due to the course modifications, while providing him internal feedback that the goal was achieved to some extent (paths of internal feedback), it was mostly not statistically significant (internal outcomes). He believed that this insignificance was in part due to the few students who typically enrolled in his course (re-evaluate beliefs).

> *"Yes, I have certain problems, marked on quizzes. I don't give large exams. I mean this course, our assessment questions. So are the students understanding these concepts and/or are the students using the correct methods to solve these problems (set goals). And so what I've done is I've marked over the years I've marked a certain number of questions throughout these quizzes, and I look at those questions (tactics and strategies), and I try to evaluate whether there's an improvement, based upon the modifications I've made or not (monitoring). Most of the time, the improvement might be small. Might be small and so it's really not statistically significant (internal outcomes). In terms because I only have small numbers of students right. If I had 30 students in each class, I might be able to see something (re-evaluate beliefs). But last semester again, when I made the changes there was definitely a significant improvement, looking at these assessment questions (paths of internal feedback). And so I think that it all depends on the situation basically, but I do have certain questions earmarked for assessment."*

> *"I am tracking the scores on those particular questions over the past 22 years I've been teaching the class (external outcomes)."*

The suggestion to include an assessment approach to evaluate student learning was something Dr. William received from external reviewers. The external review on the physics program was something required by his department. However, this external review was received on the physics program as a whole, rather than on his thermal course, in particular.

"Well, as I said it, the external feedback had very little to do with individual courses and had more to do with the program as a whole. It did mention that we needed to assess more, and so that's when we really started to implement these chosen questions that would be representative of our assessment questions. So, we evaluate both student development in problem solving and analysis as well as conceptual understanding. So if there was any input, it came as sort of a backdoor – improve your assessment – and then we reviewed, what the assessment and assessment process, and came up with this method that we currently use."

Dr. William found external review as an useful component though he identified limitations that the external review included. For example, he identified the limited information external review provided on the content in a course or mode of delivery (e.g., *"...Very little of the external reviews had to do with content in a course or mode of delivery."*).

"Well, I think the assessment. I mean, anytime someone comments oh you need more assessment, that's going to be helpful right because that's going to allow you to adapt your course to make it better, at least I hope so, or to perhaps to make it more complete. You know, in the sense that it's not just about content, it's also about method. And so the assessment approach from the peer reviewers [evaluators] and the external [reviewers] is great. Again, they don't talk about content, because usually external reviewers and peer evaluators aren't physicists and so they don't really understand the content."

Dr. William also received feedback from his peers (external feedback). Similar to external review, he found peer feedback was also having limitations. He also received less information on the thermal physics content as the peers who gave him feedback were not physicists.

"In terms of peer evaluation, every five years, now that I have tenure, I'm reviewed and have to have a peer evaluator, actually two of them sit in the classroom, and they can choose which classes they sit in. Most of the time it's not upper level physics, because my peers aren't physicists, and so, they come into

the lower level courses simply because they'll understand more. So, I've had very little input externally, whether it's peer evaluation or external to the college on upper level courses (external feedback)."

Dr. William acknowledged student feedback as most effective for him over the external review or peer feedback.

"The student feedback tends to be the most effective for me. It's not the external or peer feedback. Simply because the students have lived through the whole semester of the material, and they're going to be able to make an informed review of what happened."

Additionally, he made modifications to a course annually based upon students' evaluations, as opposed to making on-going modifications as the class progresses. The exception to this was the time where class went completely online due to the pandemic, thus leaving him to modify the course on the go.

"So I would say it [course modification] is an annual thing. But, I'm not averse to making changes mid semester, if something's not working, like last year, we went completely online, you know, and the methods simply did not work online so I had to switch on the go."

Such infrequent, on-going modifications that Dr. William made, also based on the students' feedback (external feedback), which left him to convert the mastery method to online format (re-evaluate tactics and strategies).

"I mean, in a class of 5 to 10 students, there's always going to be someone who's very willing to speak up. And if they say, this isn't working and other students say yeah, this isn't working, then I'm going to make a change. And so I'll make that on the fly if necessary (external feedback)... But like last year, last fall, it wasn't working this module method, this mastery method. And so we altered it so that it could work with them in an online format (re-evaluate tactics and strategies)."

5.3.2 Summary of Dr. William's Case Study

Each element of Fig. 5.1 in the context of Dr. William's course modifications can be found in Table 5.4 and Table 5.5. Dr. William had his own knowledge and beliefs about why he did the course modifications the way he did. These knowledge and beliefs led him to set certain goals and plan accordingly to achieve those goals. Meanwhile, Dr. William generated outcomes and started comparing these outcomes with the goals to monitor his progress. This monitoring process led him to decide whether he made progress or any changes needed to be made to yet achieve his goals. Rather than a linear process, Dr. William's learning occurred as an iterative endeavour. Thus, the processes internal to Dr. William iteratively updated based on his knowledge and beliefs, goals, tactics and strategies, internal outcomes, and external outcomes.

In addition to his own judgement about why and how he should do a course modification, students' end of the semester evaluations provided him a supportive hand for his learning process. He acknowledged that the students' feedback on the external outcomes he produced based upon a course modification was the most effective for his learning process. This was due to the course specific feedback – either focusing on course content or activities Dr. William implemented – that he received from his students.

In contrast, Dr. William found feedback received from the external reviewers and peer evaluators carry limited information about the course specific information. External reviewers focused on the physics program as a whole, rather than his thermal course, while peer evaluators provided feedback on the course, but they are not physicists and did not provide content specific feedback to Dr. William.

Based on students' end of the semester evaluations, Dr. William preferred to modify his thermal physics course annually, rather than on-the-fly. An exception to this was the modifications he made during the time of the pandemic. However, such on-the-fly modification was also implemented based upon students' feedback. However, due to the low number of students who typically enroll in Dr. William's course, he acknowledged the limited ability for him to meaningfully monitor the progress of his modification on improving student learning.

Table 5.4: *Descriptive information of each element in Fig. 5.1 and the examples for each element in the context of Dr. William modifies his course.*

Model Elements	**Definitions**	**Examples**
Activity	The activity a learner engages in.	- Modifying the thermal physics course by Dr. William.
Knowledge and Beliefs	The knowledge and beliefs the learner draws on to interpret the activity.	- Believe in best teaching methods for physics, while seeking knowledge about them through readings. - Believe that students being juniors are less prepared for the mastery method, and the method would have been worked if they were seniors. - Believe that the course modifications should help improve student learning. - Believe that the improved student learning based upon the modification he made was insignificant due to the small number of students in his classroom.
Set goals	The goals the learner has for the activity.	- Create active learning environments for students as opposed to completely lecture-based environments. - Evaluate whether students understand the concepts and/or use correct methods to solve problems, upon modifying the course.
Tactics and strategies	The tactics and strategies the learner uses to achieve the goals.	- Introducing mini lectures. - Introducing in-class assignments. - Introducing modules. - Introducing quizzes. - Introducing a session of reviewing and summarizing literature related to thermal physics. - Introducing a session of reviewing and discussing literature related to thermal physics. - Marking specific questions on the quizzes to track students' progress.

Table 5.5: *Descriptive information of each element in Fig. 5.1 and the examples for each element in the context of Dr. William modifies his course (continued).*

Model Elements	**Definitions**	**Examples**
Internal Outcomes	The outcomes that are internal to the learner (changes to cognitive or emotional states of the learner).	- Noticed an improvement by looking at the marked questions on quizzes. - Uncertainty about whether or not a modification would work successfully.
External Outcomes	The outcomes that are tangible (e.g., lecture notes) and behavioral (e.g., lecturing).	- Implemented mini lectures. - Implemented in-class assignments. - Implemented modules. - Implemented quizzes. - Session of reviewing and summarizing literature related to thermal physics. - Session of reviewing and discussing literature related to thermal physics.
Monitoring	The process in which the learner compares the outcomes with respect to the goals.	- Reflecting on the students' scores to the questions marked on the quizzes.
Paths of internal feedback	The monitoring process leads to generating paths of internal feedback (i.e. internal feedback can guide the learner to reinterpret the task and/or draw on different knowledge and beliefs and/or set different goals and/or use different tactics and strategies, and/or produce new outcomes).	- Course modification is successful to some extent, and thus the goal is achieved.
External feedback	Feedback received externally on the performance in comparison to external standards.	- End of the semester students' evaluations. - External reviewers (review is not on the individual course, but physics program as a whole). - Peer evaluators.

5.3.3 Dr. Andreas's Case Study

When we asked Dr. Andreas about the course modifications he did in his classroom, he started discussing his beliefs about the modern teaching methods that can create interactive learning environments for his students (beliefs).

> *"So, are you [interviewer] familiar with the methods of modern teaching that are more interactive? (beliefs)"*

Moving beyond just believing the modern teaching methods, he started planning and eventually implementing such methods in his classroom. It seemed like he already implemented these modern teaching methods in his lower-level classes, and wanted to preserve some of them into his upper-level statistical mechanics course as well (set goals).

> *"So from the beginning I try to preserve some of that [methods of modern teaching that are more interactive] from the lower level classes into this class (set goals)."*

To create interactive learning environments for his students, he designed and implemented a game in his typical enrollment classroom (tactics and strategies). This game was played at the beginning of one of the three lectures (external outcomes). Dr. Andreas called this game as "15 questions" if he had 15 students in his classroom, and "13 questions" if he had 13 students in his classroom. These questions allowed students to develop conversations, rather than creating a typical question-answering environment for them. The difficulty of the questions gradually increased from question one to the last question. These questions typically covered the material students had been exposed to in the previous week. The participation in these games was voluntary, but students earned bonus points if they participated.

> *"In particular, I do, it depends a little bit on the enrollment, so if the enrollment is very high, I have to modify that. But for the typical enrollment that I just mentioned, I play at the beginning of one of the three lectures, each week, a game that I call 15 questions if I have 15 students, and 13 questions if I have 13 students (external outcomes). So, it's basically engaging the material that*

the students have learned in the past week. And it's structured in such a way that participation is voluntary but students earn bonus points. And it tends to go more into conversation than actual question answering. And the difficulty of the question goes up from question one which is very easy to 13 which would be then, a very advanced question that even the good students would struggle with (tactics and strategies). That is a modification that I do in addition to just regular textbook related teaching."

Apart from the modifications Dr. Andreas made his course to provide students interactive learning environments, he also made changes to his curriculum as the semester progressed on (activity). He believed the last chapter of Schroeder, which was the textbook used in his classroom, included the real statistical mechanics (beliefs). Thus, he set up his plan to make it to the last chapter of the Schroeder (set goals).

"My motivation is that I really wanna make it into the last chapter (set goals) which is the real Stat Mech, the one that has the long applications about, say, for example, Fermi gas, or something like that. I don't want to run out of time for that (beliefs)."

Dr. Andreas's physics department only offered a freshman year course that bridged the upper-level statistical mechanics class, but it did not offer an intermediate-level course. This made him realize that the students needed sufficient background that could help them learn the material in the upper-level statistical mechanics course. For example, some prior background on the use of partial derivatives.

"I modify the first two chapters to build a stepping stone for our students to learn about that material of the intermediate class, which usually introduces the use of partial derivatives and things like that."

"Yeah. So, as I mentioned earlier, I do not follow chapter one of Schroeder religiously. I have built in something that is from an old textbook from the 1950s that

introduces the use of partial derivatives because we don't have that intermediate class that introduces that."

However, adding this additional material into Dr. Andreas's curriculum, which could have been covered in an intermediate course, led him to reduce some of the other content from the textbook (tactics and strategies). This way he could still reach his goal, which was to reach the last chapter of Schroeder.

"And so, since I added at the beginning the chapter one material that's not there, I have to leave something out later to make it to the end (tactics and strategies)."

"If you are familiar with Schroeder, chapter 4 is about engines which we cover a little bit more in depth already in the freshman level class. And I do a reminder about that but I don't do all the different chapters, some chapters that are in there. Chapter five is more like a solid state application chapter, and I go through some of it, but the later parts of that I leave out (tactics and strategies)."

As the semester went on, Dr. Andreas evaluated the progress he had been making on reaching the end of the chapter material in Schroeder. For this, he has used the games he implemented in his classroom (tactics and strategies). When students played the game, and thus their responses (external outcomes) helped him make sure whether he could reach his goals as expected (monitoring). He did this review throughout the semester about three to four times.

"Yeah, but really it's based on these games that I mentioned earlier, the 13 questions (tactics and strategies). It will usually reveal to me whether the average of the class is understanding the material or is falling behind."

"...so that is what I found out in these 13 questions. At the end of it, I realized there were 13 students in class, but really only six participated in these 13 questions in a way that was satisfactory. And the others gave answers that were either wrong or that were very very incomplete (external outcomes)."

"I probably review it [the students' responses produced after they played the game], maybe three times, four times throughout the term to compare, whether I'm still on target or not (monitoring)."

Dr. Andreas's reflection on the students' answers led him to re-evaluate the material that he could cover during the class, while accomplishing his goals to reach the end of the chapter material in Schroeder (paths of internal feedback). If he felt students fell behind, he tried to build a stepping-stone for them. This way students can be easily navigated to the material that will be covered afterwards. This is because the course contents were usually built on each other. To still accomplish Dr. Andreas's goals towards reaching the end of the chapter course material, while helping his students to catch up where they fell behind, he further removed some of the material from chapter four and five in Schroeder (re-evaluate tactics and strategies).

"And based on that, I get the sense, but it's really not a hard measure. It's more an impression that I get as a teacher. Okay, you know, these students know from their grades from other classes they should be able to do this, and yet they do not. That is a sign to me that I'm going too fast (paths of internal feedback)."

"And if they fall behind and then I try to build a stepping stone for them. That makes it easier to comprehend the material that is coming next. Because everything builds on each other right. So, when I see that then I fall behind and during the other reviews I can see that. Okay, I have to leave something more out of chapter four or five to make it still to the end (re-evaluate tactics and strategies)."

Though Dr. Andreas had to re-evaluate his tactics and strategies to accomplish his goals towards reaching the end of the chapter materials, the students' reactions to the game he implemented seemed engaging and working (monitoring), thus leaving him not making any changes to that approach (paths of internal feedback). For example, he himself witnessed how students better interact with the material (internal outcomes).

"Yeah, so it's usually the best students in the class who are giving very very elaborate answers sometimes in these games, and that is very rewarding to see right. So, they basically slip into the role of being a teacher themselves to the other students when they give these answers. And I try to encourage that and I really enjoy it when that happens (internal outcomes)."

Additionally, students themselves showed their interest towards the games that had been implemented by Dr. Andreas (external feedback).

"So I do have, usually a good relation to most of the students in the class who attend lectures. And whenever we meet somewhere in the hallways or so when I had these games that try to focus the students on the interesting parts of the material, they usually want to talk about it. So, it's you know, that they're chatting there or see two or three of them standing in the hallway after lecture. And I often join them, and find out okay, they really found this engaging and interesting (external feedback)."

Dr. Andreas also received student evaluations, but he emphasized it had been received after the course was completed. Thus, they did not inform Dr. Andreas's ongoing learning process during the semester.

"Yeah, so we have the end of term evaluations, but that alone, of course, after the course has been graded and finished, right."

Dr. Andreas also mentioned a peer evaluation system his department had. However, that only happened when he had to prepare material for promotion purposes.

"We do also have a peer evaluation system. But that is only in the years when we actually have to write up our packet for either the next promotion or for the renewal or whatever it may be, and so that I get less often."

His department also had an event called "teaching chat," where faculty and graduate students came together to talk about the current issues in teaching. However, Dr. Andreas

believed that only the colleagues who were interested in teaching reached out to that event. Thus, he believed that the feedback received from them could be biased. Instead, he preferred to obtain feedback from colleagues with different perspectives, including the ones who did not participate in the teaching chat. Thus, Dr. Andreas did not reach out to the teaching chat and his learning process during the process of course modifications was not informed by that event.

> *"But I do have something that we are doing in the department that's called a teaching chat, where the professors, and graduate students come together to talk about current issues in teaching. And these kinds of reports that somebody gives about their class can be part of that. So, in that sense I have. It's a little bit of biased feedback because it's really only the colleagues who are interested in teaching who come to that, right. I might want to also hear what the other colleagues think about it and I do not reach them with that event."*

Dr. Andreas also mentioned that he preferred to receive feedback in the form of discussions, rather than someone sending him a bulleted list of information.

> *"So, personally I like to be in the form of a discussion not just that somebody sends me a bullet list of points that they notice but that we can then sit together, maybe drink a cup of coffee or something and just chat about it. That gives it more depth than just having the objective feedback right. That way you get an impression for what they really think is important on that list and why it is important."*

> *"So, if you get just a written feedback, it's usually not clear you have to guess. If you then take the time to sit together and discuss it, that's usually when it becomes clear to me why they really thought it was important."*

5.3.4 Summary of Dr. Andreas's Case Study

Each element of Fig. 5.1 in the context of Dr. Andreas's course modifications can be found in Table 5.6 and Table 5.7. There were beliefs that led Dr. Andreas to operationalize the course modifications that he conducted. These knowledge and beliefs led Dr. Andreas to set goals, use tactics and strategies to achieve goals, and produce both internal and external outcomes. Once these outcomes were produced, Dr. Andreas started monitoring the extent to which he achieved the goals he set. This monitoring process led him to re-evaluate the tactics and strategies such that the goals he set up to cover the end of the chapter material can be yet achieved. Dr. Andreas did this monitoring process several times during the semester using the games he implemented to ensure that he could achieve his goals. He evaluated whether or not he could achieve the goals by evaluating the extent to which his students successfully engaged in the games he implemented.

In addition to his own judgement about his progress, he received feedback from his students about the games he implemented. This led him to evaluate whether his goal of creating an interactive learning environment was successfully achieved. He also typically received end semester students' evaluations. However, Dr. Andreas usually does his course modifications during the semester, on-the-fly. Thus, end semester students' evaluations did not inform his course modifications.

Additionally, his department conducted an event called teaching chat, which was not contributed for his course modifications either. He did not attend that event. He identified that the feedback he might get from the teaching chat was more of a biased feedback, as faculty who have showed interest in teaching only reached out to the teaching chat. On the other hand, the focus of this feedback was not on his thermal course, rather it was a discussion happened in general between faculty members. Dr. Andreas mentioned that sometimes the focus of that discussion was to go through other faculty member's tenure packet. He also used to get peer evaluations from his peers on the tenure materials prepared for promotion purposes.

Thus, Dr. Andreas's learning process as he engaged in the course modifications was

Table 5.6: *Descriptive information of each element in Fig. 5.1 and the examples for each element in the context of Dr. Andreas modifies his course.*

Model Elements	**Definitions**	**Examples**
Activity	The activity a learner engages in.	- Modifying the thermal physics course by Dr. Andreas.
Knowledge and Beliefs	The knowledge and beliefs the learner draws on to interpret the activity.	- Believe that modern teaching methods can create interactive learning environments for students. - Believe that the last chapter of Schroeder includes the real Stat Mech.
Set goals	The goals the learner has for the activity.	- Create interactive learning environments for students. - Reach the end of the chapter material in Schroeder.
Tactics and strategies	The tactics and strategies the learner uses to achieve the goals.	- Introducing games which include questions for students. - Reducing contents in some chapters in Schroeder to reach the last chapter. - Using implemented games as a way to evaluate students' progress through the course material.

mostly informed by his own judgement and feedback received from students.

5.4 Synthesis of the Two Case Studies

In this section, we synthesize our two case studies of Dr. William and Dr. Andreas, and answer our first two research questions: *"How do faculty discuss modifying their course?"*, and *"What are the types of external feedback faculty receive, and how do they incorporate that feedback into modifying a course?"*.

We bring up the similarities and differences of these two faculty members when it comes to their practices and own internal processes. Dr. William and Dr. Andreas's learning processes were primarily rooted in feedback received from students. However, their learning processes

Table 5.7: *Descriptive information of each element in Fig. 5.1 and the examples for each element in the context of Dr. Andreas modifies his course (continued).*

Model Elements	**Definitions**	**Examples**
Internal Outcomes	The outcomes that are internal to the learner (changes to cognitive or emotional states of the learner).	- Enjoyment seeing students' active engagement during the implemented games.
External Outcomes	The outcomes that are tangible (e.g., lecture notes) and behavioral (e.g., lecturing).	- Implemented games in the classroom. - Students' responses to the questions provided during games.
Monitoring	The process in which the learner compares the outcomes with respect to the goals.	- Evaluating the extent to which students satisfactorily responded to the questions. - Noticing students' reactions to the implemented games.
Paths of internal feedback	The monitoring process leads to generating paths of internal feedback (i.e. internal feedback can guide the learner to reinterpret the task and/or draw on different knowledge and beliefs and/or set different goals and/or use different tactics and strategies, and/or produce new outcomes).	- Removal of some content from chapter 4 and 5 in Schroeder to reach the last chapter in Schroeder. - Approving that the games seemed to work due to students' increased engagement with them.
External feedback	Feedback received externally on the performance in comparison to external standards.	- Students' explicit statements that they found that the implemented games are engaging and interesting. - Peer evaluation and "teaching chat", both of which are not received on the external outcomes.

associated with course modifications spanned across different time scales. For example, Dr. William's learning process typically occurred annually, except during the pandemic where he had to modify the course on-the-fly. On the other hand, Dr. Andreas preferred to do the course modifications regularly as opposed to doing them annually.

While both Dr. William and Dr. Andreas created active learning environments for their

students, they both made sure that their students were getting the required understanding of the course material as the semester progressed. Though these two faculty members had different approaches to evaluating students' progress, they both used some form of assessment tasks to inform this evaluation process. However, unlike the case with Dr. William, Dr. Andreas had strong preference towards the types of material he wanted to cover during the course. Additionally, unlike the case with Dr. Andreas, Dr. William had a low-enrollment class, which limited him in meaningfully reflecting on the assessment scores as it was hard for him to evaluate whether the assessment outcomes were significant with small number of students in his class.

We acknowledge that Dr. Andreas's course modifications were mostly influenced by student learning of the material, where Dr. William's course modifications were influenced by the ways in which his teaching can better facilitate effective learning environments for his students. Perhaps the limited opportunity for Dr. William to reflect on assessment scores led him to focus on his teaching instead of student learning.

Both Dr. William and Dr. Andreas also received feedback either from their peers or external reviewers. However, the focus of this feedback was not on the faculty members' respective courses. Instead, this feedback was focused on either the physics program as a whole or materials faculty members prepared for the promotional purposes. Thus, both Dr. William and Dr. Andreas identified the feedback received from their peers or external reviewers as less influential for their course modifications.

5.5 Features of the External-Feedback that can Support both Dr. William and Dr. Andreas

In this section, we answer our third research question: *"How might we incorporate experiences of faculty associated with course modifications into designing researcher-generated feedback?"*

Based on the similarities and differences discussed in Sec. 5.4, feedback which is aligned with the faculty members' local content coverage would be helpful for both of these faculty.

Since the time span of learning of these two faculty are different, feedback which is targeted both during the semester and at the end of semester would work well for both faculty. Additionally, if the feedback can be tailored to both low- and average-enrollment classes, that would also be helpful for both faculty.

These identified factors from case studies such as content coverage of a course, time frame for course modifications, and typical enrollment can be included in the development process of a survey, which can be distributed to recruit perspectives from a large group of faculty members.

Since in this study we focus on feedback rooted in assessments, to facilitate features of the feedback that can support course modifications of faculty, assessment development can include several steps. For example, a survey conducted nation-wide can provide information about the content variability within and across similar courses. The results from this survey can be used to develop the learning goals that are aligned with the tasks of an assessment.

Since the time span for course modifications are different, feedback can be aligned with learning goals that can be assessed on the scale of formative and summative. One other way to do this is to provide opportunities for faculty to customize the learning goals that they can adopt and assess in their classrooms. Aligning with the learning goals that faculty decided to assess in their classrooms, the associated feedback can be generated.

The lack of course specific feedback from external reviewers, peer evaluators, and peers provided limited opportunities for both Dr. William and Dr. Andreas to modify their courses. This strengthens our previous suggestion on providing feedback which informs faculty members' course specifics information, in particular, their local content coverage.

In the next chapter, we provide a summary of the work presented in this dissertation along with future work.

Chapter 6

Conclusion and Future Work

In responding to the need for assessments to evaluate the extent course transformations are effective in addressing scientific practices, in Chapter 3, we demonstrated a principled task design approach that can be utilized to design tasks that assess student abilities to intertwine physics concepts ("force" in our case) with the scientific practice "Using Math." As part of this process, we adopted ECD, and coupled that with the 3D-LAP to design assessment tasks (see Table 3.1).

We then used the ACER framework as a lens to look into students' responses to articulate the developed tasks' potential to elicit students' abilities to reasoning through mathematics when presented in Think-Aloud interview settings. This validation process takes into account both students' verbal responses and written solutions to holistically capture students' approaches to the presented assessment tasks. We updated the pre-defined ESs to accommodate student own knowledge representations emerging from the student data. The explicit validation process that includes written solutions expands our understanding about how these tasks can be modified for them to be utilized in paper-based summative assessment settings at large-scale college classrooms. Particularly in those settings, students' written solutions are the sole source from which to infer the extent to which their learning progresses.

Additionally, we explored the extent to which the written solutions accurately provide evidence of student reasoning. In addition to tasks' potential to elicit the expected evidence

most of the time, students' written solutions to our designed assessment mirrored student reasoning most of the time.

Therefore, we argue that utilizing and coupling both ECD and 3D-LAP is a productive approach to assess students' abilities to intertwine scientific practices with concepts. We also argue that a framework that articulates what it means to use math in physics, i.e. the ACER framework guides our task validation process by capturing students' in-the-moment reasoning. We note that the written solutions are reasonable artifacts from which to infer students' abilities to intertwine scientific practices with concepts.

This work has important implications for research-based assessments in PER. In particular, the approach to assessment task development adopting ECD and coupling with the 3D-LAP is promising at the introductory-level. Articulating an assessment argument ECD advocates which consists of the targeted performance, required knowledge, skills, and abilities to achieve the targeted performance, and evidence that supports students have the required knowledge, skills, and abilities is crucial prior developing assessment tasks. The task features to elicit the determined evidence are informed by the the 3D-LAP.

We validate the developed assessment tasks incorporating deeper insights into students' in-the-moment reasoning utilizing students' responses to these assessment tasks in Think-Aloud interview settings. Adopting an analytic framework – ACER – helps us define what it means to do math in physics, which is the target scientific practice for our study. Additionally, using the ACER framework and its perspectives on students' use of math minimizes the biases when analyzing data, in particular our own biases of what it means to do math in physics.

In validating assessment tasks, we also placed emphasis on the students' written work, which is the sole source of information available for instructors from which to infer students' knowledge, skills, and abilities. This addition provides us insights into the ways in which we can modify assessment tasks such that the students' engagement in a task can be meaningfully elicited and inferred from their written work. Though for this work we only use students' abilities to intertwine "math" with "force," other assessment designers can use what they value in students' work and still follow the process articulated in this work.

In addition to the scientific practice of "Using Math," we plan to expand this work to incorporate other scientific practices into our task design process. This future work will inform us about the extent to which our task design and validation process is consistent across different scientific practices. In the future, we plan to pilot the developed assessment to a student population with multiple backgrounds to explore how these assessment tasks promote equity.

This work also informs our on-going work of developing a new standardized assessment for upper-division thermal physics – The Thermal and Statistical Physics Assessment (TaSPA). In particular, this work informs the assessment tasks and associated feedback for faculty in-development based on students' responses to these tasks[28].

In chapter 4, we introduced a methodology to design feedback for faculty in response to the need for an approach to better communicate to faculty about students' outcomes rooted in response to research-based assessments. This methodology is intended to facilitate faculty implementing explicit course modifications. To incorporate perspectives of faculty members in to the researcher-generated feedback, we conducted semi-structured interviews with faculty. Analyzing these interviews provided us insights into the ways in which the generated feedback can be modified to better align with faculty's needs and preferences. In the following paragraphs, we explain the limitations of our study along with future work.

The methodology provided in this paper was completely applied to the example LP we used in Sec. 4.2. Thus, in the future, we plan on applying the same methodology to several other LPs. This is to evaluate the extent to which the methodology introduced in this work can be consistently expanded to other LPs. We collected students' written work in response to the FR task. We note that students' thought processes can be better elicited if we accompany video and audio recorded data while students go through the task. Thus, in the future, we plan on conducting think-aloud interviews to better capture students' reasoning. Additionally, we did not incorporate the diversity of the student population when we collected student data. We acknowledge that having a diverse student population when we collected data is an important aspect. This is to evaluate whether or not students are disadvantaged to represent their knowledge due to the structure of the assessment task or

the setting the assessment is conducted. Thus, in the future, we will try to capture diversity during student data collection.

We implemented a number of strategies to capture diversity when recruiting interview participants; however, we ended up with a group of predominantly white and male participants. We acknowledge that having more interview participants who self-identified as woman or people of color would better promote the diversity of our sample. In the future, we plan to keep trying to promote diversity in recruiting participants to obtain more diverse perspectives.

In Chapter 5, we conducted two case studies – Dr. William and Dr. Andreas – to explore the practices and processes internal to faculty when modifying courses and influence the external feedback make on those processes. We identified the similarities and differences both Dr. William and Dr. Andreas's learning processes entail and be influenced by the external feedback, during the process of course modifications.

Based on these identified similarities and differences, for feedback to be effective, it should include features that support the learning associated with course modifications of both faculty. Thus, we also provide insights into the ways in which a research-based assessment, in which the feedback is rooted can be designed such that feedback can inform the features that support each faculty member's learning.

We note that two case studies can only provide limited insights to the features the feedback should contain, if we intended to use such features for a large scale study. In the future, we analyze more interviews to expand our understanding of diverse set of internal processes and the influence external feedback can create on these internal processes. We also note that these case studies are rooted in the course modifications made by both Dr. William and Dr. Andreas on his upper-division/intermediate thermal physics course. It might be the case where different features for feedback are required if we focused on a different course. Choosing thermal physics was a deliberate choice we made as our large scale study informs developing a research-based assessment for upper-division thermal physics[27;28].

Additionally, our interview sample is not a representative sample of every faculty member. For example, it might be the case where the features of the feedback can be different if a

diverse set of faculty were included in these case studies. In the future, we plan on including faculty members which can increase the diversity of our case studies.

Bibliography

[1] Charles Henderson and Melissa H Dancy. Physics faculty and educational researchers: Divergent expectations as barriers to the diffusion of innovations. *American Journal of Physics*, 76(1):79–91, 2008.

[2] Deborah L Butler and Philip H Winne. Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research*, 65(3):245–281, 1995.

[3] Christopher J Harris, Joseph S Krajcik, James W Pellegrino, and Angela Haydel DeBarger. Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, 38(2):53–67, 2019.

[4] D. V. Schroeder. *An Introduction to Thermal Physics*. Addison Wesley, 1999. ISBN 9780201380279. URL https://books.google.com/books?id=1gosQgAACAAJ.

[5] Charles Kittel and Herbert Kroemer. *Thermal Physics*. W. H. Freeman & Co., 1980. ISBN 9780716710882. URL https://books.google.com/books?id=X4sfAQAAMAAJ.

[6] Ralph Baierlein. *Thermal Physics*. Cambridge University Press, 1999. doi: 10.1017/CBO9780511840227.

[7] James T Laverty, Sonia M Underwood, Rebecca L Matz, Lynmarie A Posey, Justin H Carmel, Marcos D Caballero, Cori L Fata-Hartley, Diane Ebert-May, Sarah E Jardeleza, and Melanie M Cooper. Characterizing college science assessments: the three-dimensional learning assessment protocol. *PloS one*, 11(9), 2016.

[8] Melanie M Cooper, Marcos D Caballero, Diane Ebert-May, Cori L Fata-Hartley, Sarah E Jardeleza, Joseph S Krajcik, James T Laverty, Rebecca L Matz, Lynmarie A Posey, and Sonia M Underwood. Challenge faculty to transform stem learning. *Science*, 350(6258):281–282, 2015.

[9] James McDonald. The next generation science standards: Impact on college science teaching. *Journal of College Science Teaching*, 45(1):13, 2015.

[10] Susan Singer and Karl A Smith. Discipline-based education research: Understanding and improving learning in undergraduate science and engineering. *Journal of Engineering Education*, 102(4):468–471, 2013.

[11] Joseph Kozminski, HJ Lewandowski, Nancy Beverly, Steve Lindaas, Duane Deardorff, Ann Reagan, Richard Dietz, Randy Tagg, M EblenZayas, J Williams, et al. Aapt recommendations for the undergraduate physics laboratory curriculum. *AAPT: College Park, MD, USA*, 2014.

[12] National Research Council et al. *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press, 2012.

[13] NGSS Lead States. *Next Generation Science Standards: For States, By States*. The National Academies Press.

[14] Rebecca L Matz, Cori L Fata-Hartley, Lynmarie A Posey, James T Laverty, Sonia M Underwood, Justin H Carmel, Deborah G Herrington, Ryan L Stowe, Marcos D Caballero, Diane Ebert-May, et al. Evaluating the extent of a large-scale transformation in gateway science courses. *Science advances*, 4(10):eaau0554, 2018.

[15] Melanie M Cooper. It is time to say what we mean, 2016.

[16] James Pellegrino. Session l: Measuring what matters: Challenges and opportunities in assessing science proficiency. URL https://research.acer.edu.au/research_conference/RC2015/17august/15.

[17] Kevin W McElhaney, Sania Zaidi, Brian D Gane, Nonye Alozie, and Christopher J Harris. Designing ngss-aligned assessment tasks and rubrics to support classroom-based formative assessment. In *NARST Annual International Conference, Atlanta, GA*, 2018.

[18] Jonathan Osborne. Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education*, 25(2):177–196, 2014.

[19] Christopher J Harris, Joseph S Krajcik, James W Pellegrino, and Kevin W McElhaney. Constructing assessment tasks that blend disciplinary core ideas, crosscutting concepts, and science practices for classroom formative applications. *Menlo Park, CA: SRI International*, 2016.

[20] Robert Mislevy and Michelle Riconscente. Evidence-centered assessment design: Layers, structures, and terminology (padi technical report 9), 2005.

[21] Norda S Stephenson, Erin M Duffy, Elizabeth L Day, Kira Padilla, Deborah G Herrington, Melanie M Cooper, and Justin H Carmel. Development and validation of scientific practices assessment tasks for the general chemistry laboratory. *Journal of Chemical Education*, 97(4):884–893, 2020.

[22] Robert J Mislevy and Geneva D Haertel. Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25(4):6–20, 2006.

[23] Robert J Mislevy, Linda S Steinberg, and Russell G Almond. Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, 1(1):3–62, 2003.

[24] Robert J. Mislevy, Linda S. Steinberg, and Russell G. Almond. *On the Roles of Task Model Variables in Assessment Design*. URL https://eric.ed.gov/?id=ED431804.

[25] Angela Haydel DeBarger, Christopher J Harris, Cynthia D'Angelo, J Krajcik, C Dahsah, J Lee, and Y Beauvineau. Constructing assessment items that blend core ideas and science practices. In *Learning and becoming in practice: The International Conference of the Learning Sciences (ICLS)*, volume 3, 2014.

[26] KW McElhaney, BD Gane, CJ Harris, JW Pellegrino, LV DiBello, and JS Krajcik. Using learning performances to design three-dimensional assessments of science profi-

ciency. In *annual meeting of the National Association for Research in Science Teaching, Baltimore, MD*, 2016.

[27] Katherine D. Rainey, Amali Priyanka Jambuge, James T. Laverty, and Bethany R. Wilcox. Developing coupled, multiple-response assessment items addressing scientific practices. In *Physics Education Research Conference Proceedings*, PER Conference, Virtual Conference, July 22-23 2020.

[28] Amali Priyanka Jambuge, Katherine D. Rainey, Bethany R. Wilcox, and James T. Laverty. Assessment feedback: A tool to promote scientific practices in upper-division. In *Physics Education Research Conference Proceedings*, PER Conference, Virtual Conference, July 22-23 2020.

[29] David Hestenes, Malcolm Wells, and Gregg Swackhamer. Force concept inventory. *The physics teacher*, 30(3):141–158, 1992.

[30] Richard R Hake. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics*, 66(1):64–74, 1998.

[31] Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23):8410–8415, 2014.

[32] Joshua Von Korff, Benjamin Archibeque, K Alison Gomez, Tyrel Heckendorf, Sarah B McKagan, Eleanor C Sayre, Edward W Schenk, Chase Shepherd, and Lane Sorell. Secondary analysis of teaching methods in introductory physics: A 50 k-student study. *American Journal of physics*, 84(12):969–974, 2016.

[33] Adrian Madsen, Sarah B McKagan, Mathew Sandy Martinuk, Alexander Bell, and Eleanor C Sayre. Research-based assessment affordances and constraints: Perceptions of physics faculty. *Physical Review Physics Education Research*, 12(1):010115, 2016.

[34] Renee Michelle Goertzen, Rachel E Scherr, and Andrew Elby. Respecting tutorial instructors' beliefs and experiences: A case study of a physics teaching assistant. *Physical Review Special Topics-Physics Education Research*, 6(2):020125, 2010.

[35] National Research Council et al. *Developing assessments for the next generation science standards*. National Academies Press, 2014.

[36] James W Pellegrino, Naomi Chudowsky, and Robert Glaser. *Knowing what students know: The science and design of educational assessment.* ERIC, 2001.

[37] James T Laverty and Marcos D Caballero. Analysis of the most common concept inventories in physics: What are we assessing? *Physical Review Physics Education Research*, 14(1):010123, 2018.

[38] Physport: Supporting physics teaching with research-based resources. https://www.physport.org/, 2020.

[39] Adrian Madsen, Sarah B McKagan, and Eleanor C Sayre. Resource letter rbai-1: research-based assessment instruments in physics and astronomy. *American Journal of Physics*, 85(4):245–264, 2017.

[40] Adrian Madsen, Sarah B McKagan, and Eleanor C Sayre. Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Physical Review Special Topics-Physics Education Research*, 9(2): 020121, 2013.

[41] Mercedes Lorenzo, Catherine H Crouch, and Eric Mazur. Reducing the gender gap in the physics classroom. *American Journal of Physics*, 74(2):118–122, 2006.

[42] Eric Brewe, Vashti Sawtelle, Laird H Kramer, George E O'Brien, Idaykis Rodriguez, and Priscilla Pamelá. Toward equity through participation in modeling instruction in introductory university physics. *Physical Review Special Topics-Physics Education Research*, 6(1):010106, 2010.

[43] Marcos D Caballero, Edwin F Greco, Eric R Murray, Keith R Bujak, M Jackson Marr, Richard Catrambone, Matthew A Kohlmyer, and Michael F Schatz. Comparing large lecture mechanics curricula using the force concept inventory: A five thousand student study. *American Journal of Physics*, 80(7):638–644, 2012.

[44] Matthew A Kohlmyer, Marcos D Caballero, Richard Catrambone, Ruth W Chabay, Lin Ding, Mark P Haugan, M Jackson Marr, Bruce A Sherwood, and Michael F Schatz. Tale of two curricula: The performance of 2000 students in introductory electromagnetism. *Physical Review Special Topics-Physics Education Research*, 5(2): 020105, 2009.

[45] Ronald K Thornton and David R Sokoloff. Assessing student learning of newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *american Journal of Physics*, 66(4):338–352, 1998.

[46] Lin Ding, Ruth Chabay, Bruce Sherwood, and Robert Beichner. Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical review special Topics-Physics education research*, 2(1):010105, 2006.

[47] David P Maloney, Thomas L O'Kuma, Curtis J Hieggelke, and Alan Van Heuvelen. Surveying students' conceptual knowledge of electricity and magnetism. *American Journal of Physics*, 69(S1):S12–S23, 2001.

[48] Steven F. Wolf, Mark W. Sprague, Feng Li, Annalisa Smith-Joyner, and Joi P. Walker. Introductory physics laboratory practical exam development: Investigation design, explanation, and argument. In *Physics Education Research Conference Proceedings*, PER Conference, Provo, UT, July 24-25 2019.

[49] Samuel Messick. Foundations of validity: Meaning and consequences in psychological assessment. *ETS Research Report Series*, 1993(2):i–18, 1993.

[50] Michael T Kane. Validation. *Educational measurement*, 4(2):17–64, 2006.

[51] Michael T Kane. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73, 2013.

[52] Wendy K Adams and Carl E Wieman. Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, 33(9):1289–1312, 2011.

[53] Randy E Bennett and Matthias von Davier. *Advancing human assessment: The methodological, psychological and policy contributions of ETS*. Springer Nature, 2017.

[54] Brian D Gane, Kevin W McElhaney, Sania Z Zaidi, JW Pellegrino, BD Gane, KW McElhaney, SZ Zaidi, and JW Pellegrino. Analysis of student and item performance on three-dimensional constructed response assessment tasks. In *NARST Annual International Conference, Atlanta, GA*, 2018.

[55] Paula V Engelhardt. An introduction to classical test theory as applied to conceptual multiple-choice tests. *Getting started in PER*, 2(1):1–40, 2009.

[56] Lin Ding and Robert Beichner. Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics-Physics Education Research*, 5(2):020103, 2009.

[57] Edward F Redish. Problem solving and the use of math in physics courses. *arXiv preprint physics/0608268*, 2006.

[58] Olaf Uhden, Ricardo Karam, Mauricio Pietrocola, and Gesche Pospiech. Modelling mathematical reasoning in physics education. *Science & Education*, 21(4):485–506, 2012.

[59] Rafael López-Gay, J Martínez Sáez, and J Martínez Torregrosa. Obstacles to mathematization in physics: The case of the differential. *Science & Education*, 24(5-6): 591–613, 2015.

[60] Claus Michelsen. Functions: a modelling tool in mathematics and science. *ZDM*, 38 (3):269–280, 2006.

[61] Ricardo Karam. Framing the structural role of mathematics in physics lectures: A case study on electromagnetism. *Physical Review Special Topics-Physics Education Research*, 10(1):010119, 2014.

[62] Fredrick Reif, Jill H Larkin, and George C Brackett. Teaching general learning and problem-solving skills. *American Journal of Physics*, 44(3):212–217, 1976.

[63] David S Wright and Clayton D Williams. A wise strategy for introductory physics. *The Physics Teacher*, 24(4):211–216, 1986.

[64] Patricia Heller, Ronald Keith, and Scott Anderson. Teaching problem solving through cooperative grouping. part 1: Group versus individual problem solving. *American journal of physics*, 60(7):627–636, 1992.

[65] Sahana Murthy. Peer-assessment of homework using rubrics. In *Aip conference proceedings*, volume 951, pages 156–159. American Institute of Physics, 2007.

[66] Jennifer Lynn Docktor. Development and validation of a physics problem-solving assessment rubric. URL http://conservancy.umn.edu/handle/11299/56637. Accepted: 2010-01-21T16:14:59Z.

[67] Edit Yerushalmi, Elisheva Cohen, Andrew Mason, and Chandralekha Singh. What do students do when asked to diagnose their mistakes? does it help them? i. an atypical quiz context. *Physical Review Special Topics-Physics Education Research*, 8(2):020109, 2012.

[68] Thomas Michael Foster. *The development of students' problem-solving skill from instruction emphasizing qualitative problem-solving*. 2000.

[69] Andrew Mason and Chandralekha Singh. Do advanced physics students learn from their mistakes without explicit intervention? *American Journal of Physics*, 78(7): 760–767, 2010.

[70] Jennifer L Docktor, Jay Dornfeld, Evan Frodermann, Kenneth Heller, Leonardo Hsu, Koblar Alan Jackson, Andrew Mason, Qing X Ryan, and Jie Yang. Assessing student written problem solutions: A problem-solving rubric with application to introductory physics. *Physical review physics education research*, 12(1):010130, 2016.

[71] David Hammer. Student resources for learning introductory physics. *American Journal of Physics*, 68(S1):S52–S59, 2000.

[72] David Hammer, Andrew Elby, Rachel E Scherr, and Edward F Redish. Resources, framing, and transfer. *Transfer of learning from a modern multidisciplinary perspective*, 89, 2005.

[73] Allan Collins and William Ferguson. Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educational psychologist*, 28(1):25–42, 1993.

[74] Marcos D Caballero, Bethany R Wilcox, Leanne Doughty, and Steven J Pollock. Unpacking students' use of mathematics in upper-division physics: where do we go from here? *European Journal of Physics*, 36(6):065004, 2015.

[75] Jonathan Tuminaro and Edward F Redish. Elements of a cognitive model of physics problem solving: Epistemic games. *Physical Review Special Topics-Physics Education Research*, 3(2):020101, 2007.

[76] Thomas J Bing and Edward F Redish. Analyzing problem solving using math in physics: Epistemological framing via warrants. *Physical Review Special Topics-Physics Education Research*, 5(2):020108, 2009.

[77] Bahar Modir, John D Thompson, and Eleanor C Sayre. Students' epistemological framing in quantum mechanics problem solving. *Physical Review Physics Education Research*, 13(2):020108, 2017.

[78] Bethany R Wilcox, Marcos D Caballero, Daniel A Rehn, and Steven J Pollock. Analytic framework for students' use of mathematics in upper-division physics. *Physical Review Special Topics-Physics Education Research*, 9(2):020119, 2013.

[79] Elizabeth K Molloy and David Boud. Feedback models for learning, teaching and performance. In *Handbook of research on educational communications and technology*, pages 413–424. Springer, 2014.

[80] Daniel J Bernstein. Peer review and evaluation of the intellectual work of teaching. *Change: The Magazine of Higher Learning*, 40(2):48–51, 2008.

[81] Therese Huston and Carol L Weaver. Peer coaching: Professional development for experienced faculty. *Innovative Higher Education*, 33(1):5–20, 2008.

[82] Trav D Johnson and Katherine E Ryan. A comprehensive approach to the evaluation. *New directions for teaching and learning*, 83:09–123, 2000.

[83] JU Overall and Herbert W Marsh. Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of educational psychology*, 71(6):856, 1979.

[84] Herbert W Marsh and Lawrence Roche. The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American educational research journal*, 30(1):217–251, 1993.

[85] Harry G Murray. Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of educational psychology*, 75(1):138, 1983.

[86] William E Cashin. Students do rate different academic fields differently. *New directions for teaching and learning*, 1990(43):113–121, 1990.

[87] Barbara L Cambridge. The paradigm shifts: Examining quality of teaching through assessment of student learning. *Innovative Higher Education*, 20(4):287–297, 1996.

[88] Chad D Ellett, Karen S Loup, RIta R Culross, Joanne H McMullen, and John K Rugutt. Assessing enhancement of learning, personal learning environment, and student efficacy: Alternatives to traditional faculty evaluation in higher education. *Journal of Personnel Evaluation in Education*, 11(2):167–192, 1997.

[89] Philip C Abrami. How should we use student ratings to evaluate teaching? *Research in Higher Education*, 30(2):221–227, 1989.

[90] Philip C Abrami, Sylvia d'Apollonia, and Peter A Cohen. Validity of student ratings of instruction: what we know and what we do not. *Journal of educational psychology*, 82(2):219, 1990.

[91] David Kember, Doris YP Leung, and KyP Kwan. Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education*, 27(5):411–425, 2002.

[92] John F Kremer. Construct validity of multiple measures in teaching, research, and service and reliability of peer ratings. *Journal of Educational Psychology*, 82(2):213, 1990.

[93] John A Centra. Evaluating the teaching portfolio: A role for colleagues. *New directions for teaching and learning*, 83:87–93, 2000.

[94] Isabeau Iqbal. Academics' resistance to summative peer review of teaching: questionable rewards and the importance of student evaluations. *Teaching in Higher Education*, 18(5):557–569, 2013.

[95] Bethany R Wilcox, Marcos D Caballero, Charles Baily, Homeyra Sadaghiani, Stephanie V Chasteen, Qing X Ryan, and Steven J Pollock. Development and uses of upper-division conceptual assessments. *Physical Review Special Topics-Physics Education Research*, 11(2):020115, 2015.

[96] Charles Henderson, Andrea Beach, and Noah Finkelstein. Facilitating change in undergraduate stem instructional practices: An analytic review of the literature. *Journal of research in science teaching*, 48(8):952–984, 2011.

[97] Charles Henderson, Noah Finkelstein, and Andrea Beach. Beyond dissemination in college science teaching: An introduction to four core change strategies. *Journal of College Science Teaching*, 39(5):18–25, 2010.

[98] Linda E Strubbe, Adrian M Madsen, Sarah B McKagan, and Eleanor C Sayre. Beyond teaching methods: Highlighting physics faculty's strengths and agency. *Physical Review Physics Education Research*, 16(2):020105, 2020.

[99] Angela R Penny and Robert Coe. Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of educational research*, 74(2):215–253, 2004.

[100] Judy McShannon, Pat Hynes, N Nirmalakhandan, Gadhamshetty Venkataramana, C Ricketts, April Ulery, and Robert Steiner. Gaining retention and achievement for students program: A faculty development program. *Journal of Professional Issues in Engineering Education and Practice*, 132(3):204–208, 2006.

[101] Sergio Piccinin and John-Patrick Moore. The impact of individual consultation on the teaching of younger versus older faculty. *The International Journal for Academic Development*, 7(2):123–134, 2002.

[102] Harry Hubball, John Collins, and Daniel Pratt. Enhancing reflective teaching practices: Implications for faculty development programs. *Canadian Journal of Higher Education*, 35(3):57–81, 2005.

[103] Edward F Redish, Jeffery M Saul, and Richard N Steinberg. Student expectations in introductory physics. *American journal of physics*, 66(3):212–224, 1998.

[104] Benjamin M Zwickl, Takako Hirokawa, Noah Finkelstein, and Heather J Lewandowski. Epistemology and expectations survey about experimental physics: Development and initial results. *Physical Review Special Topics-Physics Education Research*, 10(1): 010120, 2014.

[105] Lillian C McDermott and Edward F Redish. Resource letter: Per-1: Physics education research. *American journal of physics*, 67(9):755–767, 1999.

[106] Steven J Pollock and Noah D Finkelstein. Sustaining educational reforms in introductory physics. *Physical Review Special Topics-Physics Education Research*, 4(1):010110, 2008.

[107] Charles Henderson and Melissa H Dancy. Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. *Physical Review Special Topics-Physics Education Research*, 3(2):020102, 2007.

[108] Melissa Dancy, Charles Henderson, and Chandra Turpen. How faculty learn about and implement research-based instructional strategies: The case of peer instruction. *Physical Review Physics Education Research*, 12(1):010110, 2016.

[109] Elaine Seymour. Tracking the processes of change in us undergraduate education in science, mathematics, engineering, and technology. *Science Education*, 86(1):79–105, 2002.

[110] Charles Henderson and Melissa H Dancy. Impact of physics education research on the teaching of introductory quantitative physics in the united states. *Physical Review Special Topics-Physics Education Research*, 5(2):020107, 2009.

[111] Charles Henderson. The challenges of instructional change under the best of circumstances: A case study of one college physics instructor. *American Journal of Physics*, 73(8):778–786, 2005.

[112] Chandra Turpen and Noah D Finkelstein. Not all interactive engagement is the same: variations in physics professors' implementation of peer instruction. *Physical Review Special Topics-Physics Education Research*, 5(2):020101, 2009.

[113] Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[114] Charles Henderson, Melissa Dancy, and Magdalena 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(2):020104, 2012.

[115] Melinda T Owens, Gloriana Trujillo, Shannon B Seidel, Colin D Harrison, Katherine M Farrar, Hilary P Benton, JR Blair, Katharyn E Boyer, Jennifer L Breckler, Laura W

Burrus, et al. Collectively improving our teaching: attempting biology department–wide professional development in scientific teaching. *CBE—Life Sciences Education*, 17(1):ar2, 2018.

[116] Sara E Brownell and Kimberly D Tanner. Barriers to faculty pedagogical change: Lack of training, time, incentives, and... tensions with professional identity? *CBE—Life Sciences Education*, 11(4):339–346, 2012.

[117] National Research Council et al. *Improving undergraduate instruction in science, technology, engineering, and mathematics: Report of a workshop*. National Academies Press, 2003.

[118] K Anders Ericsson and Herbert A Simon. Verbal reports as data. *Psychological review*, 87(3):215, 1980.

[119] Marsha E Fonteyn, Benjamin Kuipers, and Susan J Grobe. A description of think aloud method and protocol analysis. *Qualitative health research*, 3(4):430–441, 1993.

[120] Otter.ai: Otter voice meeting notes. https://www.otter.ai/, 2020.

[121] Bethany R Wilcox and Steven J Pollock. Upper-division student difficulties with the dirac delta function. *Physical Review Special Topics-Physics Education Research*, 11 (1):010108, 2015.

[122] Bethany R Wilcox and Steven J Pollock. Upper-division student difficulties with separation of variables. *Physical Review Special Topics-Physics Education Research*, 11 (2):020131, 2015.

[123] Nandana Weliweriya, Justyna P Zwolak, Eleanor C Sayre, and Dean Zollman. Varied reasoning schema in students' written solutions. *arXiv preprint arXiv:1611.02262*, 2016.

[124] Hosun Kang, Jessica Thompson, and Mark Windschitl. Creating opportunities for

students to show what they know: The role of scaffolding in assessment tasks. *Science Education*, 98(4):674–704, 2014.

[125] Melanie M Cooper and Ryan L Stowe. Chemistry education research—from personal empiricism to evidence, theory, and informed practice. *Chemical Reviews*, 118(12): 6053–6087, 2018.

[126] Randy Elliot Bennett. Cognitively based assessment of, for, and as learning (cbal): A preliminary theory of action for summative and formative assessment. *Measurement*, 8(2-3):70–91, 2010.

[127] RE Bennett, Michael Kane, and Brent Bridgeman. Theory of action and validity argument in the context of through-course summative assessment. *Princeton, NJ: Educational Testing Service*, 2011.

[128] Randy Elliot Bennett. Formative assessment: A critical review. *Assessment in education: principles, policy & practice*, 18(1):5–25, 2011.

[129] RE Bennett. Theory of action and educational assessment. In *National Conference on Student Assessment, Orlando, FL*, 2011.

[130] Amali Priyanka Jambuge and James T Laverty. Assessing scientific practices in physics paper-based assessments. *arXiv preprint arXiv:2106.13028*, 2021.

[131] Linda A Meyer. Strategies for correcting students' wrong responses. *The Elementary School Journal*, 87(2):227–241, 1986.

[132] Robert L Bangert-Drowns, Chen-Lin C Kulik, James A Kulik, and MaryTeresa Morgan. The instructional effect of feedback in test-like events. *Review of educational research*, 61(2):213–238, 1991.

[133] David J Nicol and Debra Macfarlane-Dick. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2):199–218, 2006.

[134] Margaret Price, Karen Handley, Jill Millar, and Berry O'donovan. Feedback: all that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, 35(3): 277–289, 2010.

[135] David Carless, Diane Salter, Min Yang, and Joy Lam. Developing sustainable feedback practices. *Studies in higher education*, 36(4):395–407, 2011.

[136] John Hattie and Helen Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[137] Bethany R Wilcox and Steven J Pollock. Student behavior and test security in online conceptual assessment. In *Physics Education Research Conference Proceedings*, PER Conference, Provo, Utah, July 24-25 2019.

[138] Bethany R Wilcox and Steven J Pollock. Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics. *Physical Review Special Topics-Physics Education Research*, 10(2):020124, 2014.

[139] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[140] Amali Priyanka Jambuge, Katherine D. Rainey, Amogh Sirnoorkar, Bethany R. Wilcox, and James T. Laverty. Designing research-based assessment feedback for physics faculty. *in prep to submit to PRPER.*

Appendix A

Supplemental Material for the "Assessing Scientific Practices in Physics Paper-Based Assessments" Study

Engineers are testing a new roller coaster-like ride before it starts functioning. The sandbags are strapped into the train to simulate passengers and the total mass of the train with sandbags is 1000 kg. It is supposed to start from rest at point A and stop at point E. The train starts braking at point D so that it will come to a stop at point E. If the brake system applies an average force of 6749 N, will it be enough to stop the train at point E? Under what conditions do you think your conclusion is valid? The heights from the ground to points A, B, C, D and E are 173, 145, 124, 95 *and* 95 (*in* m), respectively. The distance from D to E is 113 m.

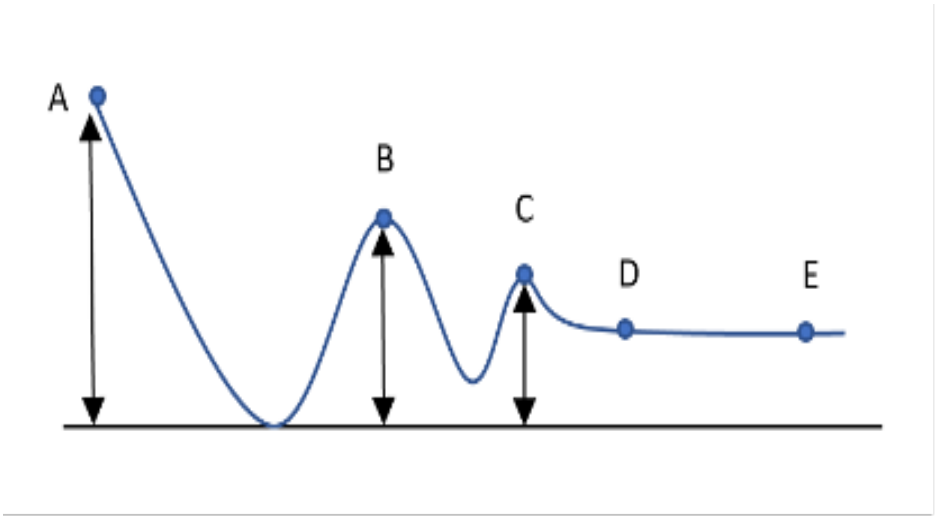

Figure A.1: *Roller coaster problem from the assessment.*

Consider a Ferris wheel in an amusement park in California. A Ferris wheel is a large circular machine with seats attached to the rim of it. The seats can freely rotate so that when the Ferris wheel is spinning, the seats hang downwards at all times. Assume the wheel is rotating with angular velocity ω and the diameter of the wheel is D. At what point in the motion does a rider feel "heaviest" and "lightest"?

Approximately how large would ω have to be for this to have a noticeable effect on your weight?

Figure A.2: *Ferris wheel problem from the assessment.*

The Ford manufacturers are testing their new version of the Ford Focus before releasing it to the market. The car has a mechanism for calculating its velocity, acceleration, and position and displaying those values on the dashboard. The manufacturers want to test whether this feature is functioning as intended. The case study is, the car will be travelling at a speed of 5.2 m/s and after 2 s, it starts accelerating at a constant rate of 2.2 m/s^2 until it reaches maximum speed of 16.6 m/s. If the dashboard readings for the velocities are 5.2, 5.2, 7.4, 9.6, 11.8 (in m/s) in each second (until 5 s), is the feature working correctly?

If the dashboard readings for the positions are 5.2, 10.4, 16.7, 25.2, 35.9 (in m) in each second (until 5 s), is the feature working correctly? Assume position is considered to be zero at time zero (0 s).

Figure A.3: *Testing Ford Focus problem from the assessment.*

You are in charge of designing a roller coaster for the county fair. To meet zoning requirements of the city your roller coaster can only be 20 m at the top of the first drop. Promptly afterwards the carts fall down a steep incline and go through a loop. Your friend Tom says that the loop should have a radius of 5 m. Is Tom's radius a safe radius, i.e., does the cart stick to the track as it goes around the loop?

Figure A.4: *Designing a roller coaster problem from the assessment. This task has been designed, administered, and recorded student responses by Katherine C. Ventura.*

There is an airplane moving with a constant speed of 820 km/h parallel to the ground (horizontally). The altitude and the wind speed (along the direction of the flight) displayed on the cockpit of the airplane are 12.4 km and 50 km/h. The pilot is intending to drop a bomb at a target 15 km ahead on the ground. When should the pilot drop the bomb?

He wasn't able to drop the bomb at the right time, but the target is still ahead. He got a message from the ground communication unit to release the bomb with a downwards speed of 50 m/s when the cockpit shows the target on the ground is 9.8 km ahead. Would it reach the target?

Figure A.5: *Airplane problem from the assessment.*

You are asked to design a Gravitron for the county fair, an amusement park ride where the rider enters a hollow cylinder, radius of 4.6 m, the rider leans against the wall and the room spins until it reaches angular velocity, at which point the floor lowers. The coefficient of static friction is 0.2. You need this ride to sustain mass between $25 - 160$ kg to be able to ride safely and not slide off the wall. If the minimum ω is 3 rad/s will anyone slide down and off the wall at these masses? Explain your reasoning using diagrams, equations, and words.

Figure A.6: *Gravitron problem from the assessment. This task has been designed, administered, and recorded student responses by Katherine C. Ventura.*

Table A.1: *Full codebook with examples from data.*

Codes	Subcodes	Symbol	Examples from data (Verbal)	Examples from data (Written)
Activation (A)	Identify appropriate physics concepts that can be used to understand the phenomenon	A1	"You feel 'weight' as your net force/accleration"	$F_{\text{net}} = m\frac{D}{2}\omega^2 - mg$
	Identify general physics equations to be applied	A2	"To find acceleration, V squared equals V note squared plus two a hmm change in x component"	$V^2 = V_0^2 + 2a(x - x_0)$
	Identify target parameters	A3	"So then just find the initial speed and compare to see if the driver is at fault"	$V@D$
Construction of the model (C)	Apply the general equations to a particular situation	C1	"Zero is equal to two a x minus x note plus V note squared"	$0 = V_0^2 + 2a(x - x_0)$
	Make assumptions	C2	"I'm assuming there's no friction between rest to E. No friction that kind of including drag. And energy is conserved and it will be sufficient. We're assuming that the train is attached to the track starting from rest"	I assumed the track is frictionless
	Develop representations (diagrams, free body diagrams)	C3	[Drawing a free body diagram] "You got force of friction, mg down hmm and then you got a velocity"	[Free body diagrams/representations of the physical system/modified diagrams given in the exam]
	Develop mathematical relations based on the physics concepts used	C4	"Thirteen [Fifteen] point two meters per second over twenty two point two [four] meters equals eighteen meters over x"	$\frac{15.2\text{ m/s}}{22.4\text{ m}} = \frac{18\text{ m/s}}{X}$ = 26.5 m

Table A.2: *Full codebook with examples from data (continued).*

Codes	Subcodes	Symbol	Examples from data (Verbal)	Examples from data (Written)
Execution of the mathematics (E)	Manipulate symbols	E1	"So I'm gonna use the Newton's law where F equals ma so a equals F over m"	$F = ma$, $a = \frac{F}{m}$
	Perform an arithmetic calculation	E2	[Input values to the calculator to calculate the values numerically]	$V_0^2 = 415.65$, $V_0 = 20.397$ m/s
	Execute math conceptually	E3	"m is just the same thing so m is cancelled out so a equals mu, k times g"	$ma = \mu_k mg$, $a = \mu_k g$
	Substitute expressions	E4	"Ok so, F equals ma which equals mu, k, m [g]"	$F = ma = \mu_k mg$
	Manipulate mathematical expressions	E5	"Ok, we want to track off one thousand, nine point eight times ninety five meters and multiply all by two and divided by one thousand to get V, D squared"	$0 = 15.2^2 + 2a(22.4)$, $\frac{-231.04}{2*22.4} = a$
Reflection of the results (R)	Make sense of the answer with the information given in the prompt	R1	"I mean the average skid mark at this point would be twenty six point five meters and given forty point three which is like insane"	$X = 31.088$ m $<$ 40.3 m, Yes the driver was at fault
	Make sense of the answer found in an intermediate/final step	R2	"The acceleration is equal to negative five point two one six which once again is a reasonable answer since they are going from a faster speed down to a lower speed"	$F = ma$, $a = \frac{F}{m}$, $a = \frac{6749\ \text{N}}{1000\ \text{kg}}$, $a = -6.749\ \text{m/s}^2$
	Make sense of the result for use in a subsequent step	R3	"Hmm, if that's the acceleration then that should also be the acceleration of the crash actually occurred on"	[Calculated during previous part of the questions] $a = -5.16\ \text{m/s}^2$, $V^2 = V_0^2 + 2a(X - X_0)$, $\sqrt{V^2} = \sqrt{2 * 5.16 * 40.3}$

Appendix B

Supplemental Material for the "Designing Research-Based Assessment Feedback for Physics Faculty" Study

B.1 Interview Protocol for Semi-Structured Interviews with Faculty

1. Welcome to the interview.
   (a) First of all, thank you for volunteering for this interview and we appreciate your help on this research.
   (b) I'm [NAME] and a [POSITION] at [INSTITUTION].
      i. E.g., I'm Amali and a graduate student at Kansas State University.
2. Brief Introduction about the research.
   (a) We're doing research to explore how we can give faculty actionable feedback to

modify their courses as part of developing a standardized assessment for upper-division thermal physics.

(b) Let me give you an overview about the standardized assessment and the feedback that accompanies the assessment.

   i. This standardized assessment can be conducted as an end of the semester assessment as a diagnostic tool.

   ii. The assessment includes a set of learning goals/claims (I think it might be better to use the learning goal term here as that might be familiar to many faculty).

      A. Learning goal description: Assessable statements about what students should know and be able to do.

      B. Example learning goal: (SI-U1-A-f) Use a model to determine what will happen to the internal energy of a thermodynamic system given information about the energy flow into and out of the system.

   iii. Faculty have the autonomy to pick up the learning goals that they value the most, from the available pool of learning goals.

   iv. Students will take up the assessment with that selected set of questions in an online-setting.

   v. After students finished up the assessment, faculty will receive feedback that contains how well their instruction supports students to achieve the selected learning goals.

      A. This feedback contains actionable information that faculty can adopt to modify their courses to support students in meeting the selected learning goals.

(c) Therefore, this interview is to explore your perspectives [as a faculty who teaches upper-division/intermediate thermal physics courses] on this researcher-generated feedback.

3. Overview about the interview

    (a) You will be asked several questions during the interview about your classroom practices particularly how you go about course modifications, perspectives on some researcher-generated learning goals and the feedback associated with those learning goals. We would like to know your thoughts, benefits, and challenges associated with them assuming you would use them in your own classroom.

4. Consent to conduct interviews.

    (a) Thank you for sending us the signed consent form. Do you have any questions for us about the consent form, the interview, or something broader about the project?

        i. By this time, interviewee responded with a signed consent form.

5. Recording the interviews.

    (a) I will begin recording now.

        i. "The interviews will be recorded" has already been mentioned in the solicitation emails and consent forms.

6. Any questions before proceeding.

    (a) Your participation in this interview is completely voluntary and you are allowed to withdraw this interview at any point you decided to do so. Upon your withdrawal, all the data acquired will be destroyed.

    (b) I'll be taking notes during the interview, but I'm listening when you're speaking.

    (c) Do you have any questions before we proceed?

7. Interview questions:

    (a) General Questions.

i. To get a better understanding about your teaching responsibilities, could you provide us some information such as your academic rank, the thermal course you teach/have taught (level of the course, textbook, who typically enroll in the course), the typical/average number of students in that class, how often you teach that course, etc.?

(b) Addressing theoretical aspects of feedback.

Capturing Self-Regulated learning aspects to promote faculty agency:

i. How, if at all, do you go about course modifications to your own course?

A. (If said THEY DID) What motivates you to do those modifications?

B. (If said THEY DID) How often do you do course modifications?

C. (If said THEY DID) Do you measure the success/impact of your course modification?

D. (If said YES to C) How do you measure a modification to your course as successful?

E. (If said NO to C) Move to part F.

F. What rewarding experiences and challenges did you encounter during the process of course modifications?

G. (If said THEY DID NOT) Are there any particular reasons you have not made/rarely make modifications to your course?

H. (If said THEY DID NOT) What barriers/challenges do you face in making course modifications?

I. (If said THEY DID NOT) If you were to make modifications, how would you do those modifications?

ii. Can you remember a time when you got feedback from external sources such as researcher-generated feedback after conducting standardized assessment in your classroom, or a peer evaluation?

A. (If said YES) What determines that you receive external feedback?

B. (If said YES) Did you ask for it?

C. (If said YES) Did your department want or require you to have?

D. (If said YES) How, if at all, did you use that external feedback, specifically for course modifications?

E. (If faculty said that they incorporated the feedback) Did you find any rewarding experiences, benefits or challenges when incorporating that feedback?

F. (If faculty said that they incorporated the feedback) What features of that feedback were most helpful for you?

G. (If faculty said that they did not incorporate the feedback) Would you like to share why you didn't incorporate that feedback into your course modifications? It's ok if you do not wish to mention any names associated.

H. (If said NO to ii) Move to the next question.

iii. Overall, what does effective feedback look like to you? This could be based on something that you generated yourself or by someone else on your behalf.

(c) Addressing practical aspects of feedback.

Capturing practicality of the researcher-generated feedback:

SHARE THE SET OF LEARNING GOALS AND CORRESPONDING FEEDBACK

i. We would like to provide you a list of learning goals and we would like that you rate (1- least likely to be assessed , 10- most likely to be assessed) them 1-10 based on your preference to assess them in your classroom.

A. Give faculty 5 learning goals where we already have the assessment tasks that address them.

ii. (After faculty rated the learning goals that they valued the most)

A. What do you think about these learning goals?

B. Would you want to assess these types of learning goals in your class? Why or why not?

C. Are there any specific reasons about why you rated learning goals in that way?

D. What kind of feedback would you like to receive around these learning goals?

E. Why do you feel that kind of feedback is the most useful for you?

After faculty responded to the above question, provide and explain faculty, the feedback reports that we expect them to demonstrate.

iii. What do you think about this feedback? Do you think this would be useful to you?

A. (If said YES) How would you use this to make course modifications?

B. (If said NO) Move to part v below.

iv. What are the things that you like about this feedback?

v. What are the things that seem challenging about this feedback?

vi. What changes do you feel that might be needed for this feedback to make it more effective?

vii. Do you feel this feedback is informative enough to make course modifications in your classroom?

A. (If said YES) Why do you think so?

B. (If said NO) What changes do you feel that might be needed for this feedback to align with your needs?

Move to the next learning goal and repeat the process for part (iii) up above.

viii. Overall, what do you think about giving faculty the option to pick learning goals so that they could assess learning goals that they prefer the most to be assessed in their classrooms?

8. Closing remarks.

    (a) It's almost 50 mins and at this time, we would like to wrap up the interview.

        i. Do you have anything that you would like to share before we wrap up?

        ii. Would you like to be contacted if we have follow-up interviews. That will not be sometime soon.

        iii. Would you be interested in piloting a version of this standardized assessment where you can choose learning goals to assess in your class in the future?

        iv. Do you think anything we discussed during this interview would impact the way you approach instruction or course modifications?

        v. Do you have any questions for us?

        vi. If you would like, it would be helpful for us to have your demographic information. This will help us ensure that we are incorporating input from a wide range of backgrounds.

            A. Put the demographic survey in chat.

        vii. Thank you so much for your participation and your insightful thoughts on this work. We really appreciate that.

9. Stop recording.

10. Take any additional notes after wrapping up the interview before getting forgotten.

11. Store data and consent forms at a secure place. Use pseudonyms to protect the identity of the participants.

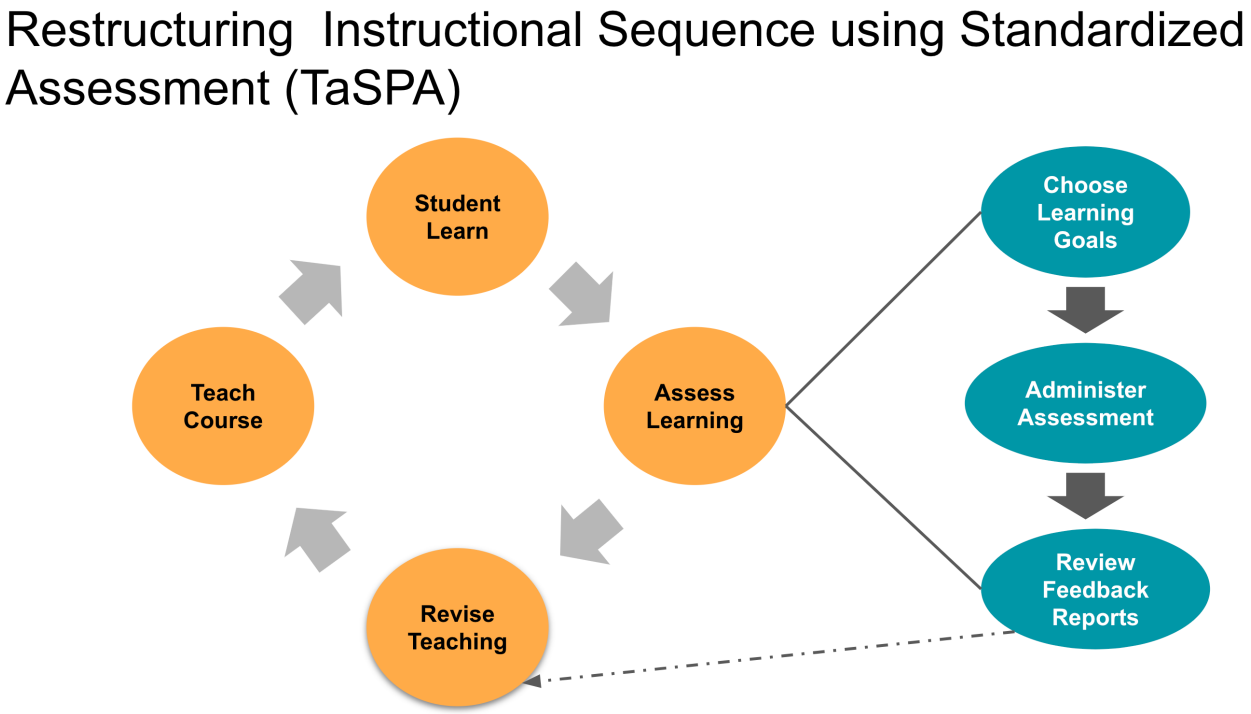

Figure B.1: *This presentation slide was used during the semi-structured interviews to help faculty understand how TaSPA can facilitate course modifications in classrooms.*

Example Learning Goals You Could Choose:

1- Least likely to be Assessed **10- Most Likely to be Assessed**

1. 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.

2. Analyze and interpret data to justify or refute claims about temperature of a system using information about changes in entropy and internal energy.

3. Use a representation of a physical system to determine the number of microstates for a given macrostate to predict the system's macroscopic property of entropy.

4. Analyze and interpret data to determine whether a thermodynamic process will happen spontaneously using the idea that entropy of the universe is maximized for spontaneous processes.

5. Use mathematics to determine the number of microstates within a system to deduce the macroscopic quantity of entropy for that system and make a conclusion about the system.

Figure B.2: *This presentation slide was used during the semi-structured interviews to demonstrate to faculty a set of sample learning goals the TaSPA includes. We also asked them to rate from 1 to 10, the likelihood of them assessing these learning goals in their classrooms.*

Template for the Feedback Reports

Learning Goal:

Expected Performance	Evidence of Expected Performance			Students' Performance based on Overall Class Trends	Recommendations for Course Modifications to Address the Expected Performance
	No	Some	Yes		
1					
2					
3					

Figure B.3: *This presentation slide was used during the semi-structured interviews to demonstrate to faculty a template of the feedback that will be provided to them.*

Learning Goal: 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.

Expected Performance 1	Evidence of Expected Performance			Students' Performance based on Overall Class Trends	Recommendations for Course Modifications to Address the Expected Performance
	No	Some	Yes		
Students relate changes in internal energy of a system to **both** heat and work as forms of energy flow into and out of that system.	0%	50%	50%	**50%** of students related changes in 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.	• Students can be given more opportunities to explore how **both** heat and work as forms of energy flow into and out of a system can be related to changes in internal energy of that system. • These opportunities can be contextualized in complex real-world scenarios which include concurrent changes to temperature, pressure, and volume of a system.

Figure B.4: *A sample feedback used during the interviews to obtain perspectives of faculty about the generated feedback.*

Learning Goal: 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.

Expected Performance 2	Evidence of Expected Performance			Students' Performance based on Overall Class Trends	Recommendations for Course Modifications to Address the Expected Performance
	No	Some	Yes		
Students construct arguments about the changes to internal energy of a system. These arguments are composed of coherent reasoning that takes into account the contributions from **both** heat and work as forms of energy flow into and out of that system.	50%	50%	0%	**50%** of students did not take into account the contributions of either heat or work as forms of energy flow into and out of that system to construct an argument about the changes to internal energy of a system during the full duration of the considered process.	• Students can be given more opportunities to generate coherent explanations about changes in internal energy of a system when the considered process involves concurrent contributions from both heat and work. • Situating these opportunities in real-world scenarios that include systems undergoing multiple processes (e.g., expansion and compression of gases) could be helpful for this population of students.

Figure B.5: *A sample feedback used during the interviews to obtain perspectives of faculty about the generated feedback.*

Learning Goal: 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.

Expected Performance 3	Evidence of Expected Performance			Students' Performance based on Overall Class Trends	Recommendations for Course Modifications to Address the Expected Performance
	No	Some	Yes		
Students make an accurate claim about the overall changes to internal energy of a system.	15%	-	85%	**85%** of students met the expected performance.	• No course modifications are suggested to address the expected performance.

Figure B.6: *A sample feedback used during the interviews to obtain perspectives of faculty about the generated feedback.*

Learning Goal: Use a representation of a physical system to determine the number of microstates for a given macrostate to predict the system's macroscopic property of entropy.

Expected Performance 1	Evidence of Expected Performance			Students' Performance based on Overall Class Trends	Recommendations for Course Modifications to Address the Expected Performance
	No	Some	Yes		
Students identify the relation that connects the macroscopic property of entropy to the number of microstates of a system.	34%	30%	36%	**34%** of students did not identify a relation that connects the macroscopic property of entropy to microscopic properties of a system.	• Students can be given more opportunities to explore when and how macroscopic and microscopic properties of a system can be linked. • Such opportunities can be situated in lessons that include discussions connecting entropy (macroscopic property) of a system to the number of microstates (microscopic property) of that system. • Including real-world scenarios that help visualize this connection in discussions could help students more easily grasp the connection between macroscopic and microscopic properties of a system.

Figure B.7: *A sample feedback used during the interviews to obtain perspectives of faculty about the generated feedback.*

Learning Goal: Use a representation of a physical system to determine the number of microstates for a given macrostate to predict the system's macroscopic property of entropy.

Expected Performance 2	Evidence of Expected Performance			Students' Performance based on Overall Class Trends	Recommendations for Course Modifications to Address the Expected Performance
	No	Some	Yes		
Students determine the number of microstates of the given system using an appropriate representation of that system.	50%	-	50%	**50%** of students did not determine the number of microstates of the given system using appropriate representation of that system.	• Students can be given more opportunities to use a representation of a physical system to determine the number of possible microstates of that system. • These opportunities could be situated in real-world scenarios that help students model and conceptualize concurrent changes to both macroscopic (volume) and microscopic (number of microstates) properties of a system. • This could enable students to effectively use a representation of a system in determining the number of microstates of that system, when presented with new problems capturing real-world contexts.

Figure B.8: *A sample feedback used during the interviews to obtain perspectives of faculty about the generated feedback.*

Learning Goal: Use a representation of a physical system to determine the number of microstates for a given macrostate to predict the system's macroscopic property of entropy.

Expected Performance 3	Evidence of Expected Performance			Students' Performance based on Overall Class Trends	Recommendations for Course Modifications to Address the Expected Performance
	No	Some	Yes		
Students make accurate predictions about the entropy of the system that align with their chosen representation.	73%	-	27%	**73%** of students did not make accurate predictions about the entropy of the system.	• Students can be given opportunities to make predictions about the macroscopic property of a system such as entropy using appropriate representations. • These predictions could be embedded in real-world scenarios which include representations to determine microscopic properties of a system. • These predictions should also be made along with appropriate connections between macroscopic property (entropy) and microscopic property of a system (number of accessible microstates). • These opportunities could enable students to make appropriate predictions about the entropy of a system when presented with new real-world contexts.

Figure B.9: *A sample feedback used during the interviews to obtain perspectives of faculty about the generated feedback.*

Table B.1: *Feedback aligned with the internal energy task. This task addresses the* ***LP:*** *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. Feedback corresponds to* ***ES1:*** *Relations that connect change in internal energy to heat and work. Ratings of 2, 1, and 0 align with the criteria of proficiency -met, -partially met, and -not met.*

Rating	ES1
2	i. Students relate changes in internal energy of a system to **both** heat and work as forms of energy flow into and out of that system. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
1	i. Students relate changes in internal energy of a system to **both** heat and work as forms of energy flow into and out of that system. ii. Students related changes in 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. iii. Students can be given more opportunities to explore how *both* heat and work as forms of energy flow into and out of a system can be related to changes in internal energy of that system. These opportunities can be contextualized in complex real-world scenarios which include concurrent changes to temperature, pressure, and volume of a system.
0	i. Students relate changes in internal energy of a system to **both** heat and work as forms of energy flow into and out of that system. ii. Students did not relate changes in internal energy of a system to either heat or work as forms of energy flow into and out of that system. iii. Students can be given more opportunities to explore factors that contribute to changes in internal energy of a system using ideas of conservation of energy, such as the first law of thermodynamics. These opportunities can be contextualized in real-world scenarios to specifically identify the contributions from both heat and work to changes in internal energy, while also paying attention to the physical properties of a system and the process(es) it undergoes. This can enable them to prompt appropriate factors such as heat and work as relating to changes in internal energy of a system when presented with new real-world scenarios.

Table B.2: *Feedback aligned with the internal energy task. This task addresses the **LP:** 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. Feedback corresponds to **ES2:** Generated explanation about the change in internal energy of the system using relations that include heat and work. Ratings of 2, 1, and 0 align with the criteria of proficiency -met, -partially met, and -not met.*

Rating	ES2
2	i. Students construct arguments about the changes to internal energy of a system. These arguments are composed of coherent reasoning that takes into account the contributions from **both** heat and work as forms of energy flow into and out of that system. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
1	i. Students construct arguments about the changes to internal energy of a system. These arguments are composed of coherent reasoning that takes into account the contributions from **both** heat and work as forms of energy flow into and out of that system. ii. Students constructed an argument about the changes to internal energy of a system during the full duration of the considered process by taking into account the contributions from **either** heat or work as forms of energy flow into and out of that system, but not both. iii. Students can be given more opportunities to generate coherent explanations about how both heat and work as forms of energy can concurrently contribute to the changes in internal energy of a system. Embedding these opportunities in real-world scenarios that include systems undergoing multiple processes could be helpful for students. Such processes could include isobaric, isochoric, adiabatic or isothermal expansions or compressions.
0	i. Students construct arguments about the changes to internal energy of a system. These arguments are composed of coherent reasoning that takes into account the contributions from **both** heat and work as forms of energy flow into and out of that system. ii. Students did not take into account the contributions of either heat or work as forms of energy flow into and out of that system to construct an argument about the changes to internal energy of a system during the full duration of the considered process. iii. Students can be given more opportunities to generate coherent explanations about changes in internal energy of a system when the considered process involves concurrent contributions from both heat and work. Situating these opportunities in real-world scenarios that include systems undergoing multiple processes (e.g., expansion and compression of gases) could be helpful for this population of students.

Table B.3: *Feedback aligned with the internal energy task. This task addresses the* ***LP:*** *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. Feedback corresponds to* ***ES3:*** *Statement about the change in internal energy of the system. Ratings of 1, and 0 align with the criteria of proficiency -met, and -not met.*

Rating	ES3
1	i. Students make an accurate claim about the overall changes to internal energy of a system. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
0	i. Students make an accurate claim about the overall changes to internal energy of a system. ii. Students did not make an accurate claim about the overall changes to internal energy of a system. iii. Students can be given more opportunities to make conclusions about the overall changes to internal energy of a system based on properties of a system and/or characteristics of a particular process. These opportunities can be situated in real-world scenarios with systems undergoing multiple processes such as expansion and compression of gases. Thus, these opportunities could enable students to make conclusions about overall changes to internal energy of a system based on changes to internal energy during each process.

Table B.4: *Feedback aligned with the simulation task. This task addresses the* ***Learning Performance:*** *Analyze and interpret data to justify or refute claims about temperature of a system using information about changes in entropy and internal energy. Feedback corresponds to* ***ES1:*** *Relations connecting temperature, internal energy, and entropy. Ratings of 2, 1, and 0 align with the criteria of proficiency -met, -partially met, and -not met.*

Rating	ES1
1	i. Students identify an accurate relation that links temperature to entropy and internal energy. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
0	i. Students identify an accurate relation that links temperature to entropy and internal energy. ii. Students did not identify an accurate relation that links temperature to entropy or internal energy. iii. Students can be given more opportunities to explore factors that temperature of a system can be attributed with. They could benefit from more opportunities contextualized in real-world scenarios to learn and effectively apply the explicit link between temperature, entropy, and internal energy of a system. This could enable students to prompt the connection between temperature, entropy, and internal energy of a system when presented with new real-world scenarios.

Table B.5: *Feedback aligned with the simulation task. This task addresses the* ***Learning Performance:*** *Analyze and interpret data to justify or refute claims about temperature of a system using information about changes in entropy and internal energy. Feedback corresponds to* ***ES2:*** *Use of mathematical relations connecting temperature, internal energy, and entropy to determine values of temperature from the data provided. Ratings of 2, 1, and 0 align with the criteria of proficiency -met, -partially met, and -not met.*

Rating	ES2
2	i. Students use the provided data about the changes to entropy and internal energy of a system to determine the temperature values. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
1	i. Students use the provided data about the changes to entropy and internal energy of a system to determine the temperature values. ii. Students used the provided data to reason about temperature, without explicitly calculating any temperature values (e.g., trend analysis). iii. Students can be given more opportunities to determine explicit values of temperatures when data describing both entropy and internal energy of a system are provided, as opposed to only considering trends about the temperature from the provided data describing entropy and internal energy.
0	i. Students use the provided data about the changes to entropy and internal energy of a system to determine the temperature values. ii. Students did not use the provided data to determine the temperature values. iii. Students can be given more opportunities to determine temperature values using provided data about changes to entropy and internal energy of a system. This could help students utilize provided data about the changes to entropy and internal energy in calculating temperature values of a system, when presented with similar contexts.

Table B.6: *Feedback aligned with the simulation task. This task addresses the* ***Learning Performance:*** *Analyze and interpret data to justify or refute claims about temperature of a system using information about changes in entropy and internal energy. Feedback corresponds to* ***ES3:*** *Statement about the validity of a provided claim or hypothesis using the given data about internal energy and entropy by utilizing the mathematical relationship between temperature, internal energy, and entropy. Ratings of 1, and 0 align with the criteria of proficiency -met, and -not met.*

Rating	ES3
1	i. Students make an accurate claim about the temperature of a system using reasoning based on data. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
0	i. Students make an accurate claim about the temperature of a system using reasoning based on data. ii. Students did not make an accurate claim about the temperature of a system using reasoning based on data. iii. Students can be given more opportunities to make claims about the temperature of a system with appropriate reasoning that incorporates factors such as entropy and internal energy. This could enable students to make accurate claims about the temperature of a system that is consistent with their analysis of data that incorporates properties such as entropy and internal energy when presented with new problems.

Table B.7: *Feedback aligned with the rubber balls in a box task. This task addresses the* ***Learning Performance:*** *Use a representation of a physical system to determine the number of microstates for a given macrostate to predict the system's macroscopic property of entropy. Feedback corresponds to* ***ES1:*** *Relation that connects entropy to the number of microstates of a system. Ratings of 2, 1, and 0 align with the criteria of proficiency -met, -partially met, and -not met.*

Rating	ES1
2	i. Students identify the relation that connects the macroscopic property of entropy to the number of microstates of a system. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
1	i. Students identify the relation that connects the macroscopic property of entropy to the number of microstates of a system. ii. Students only identified a relation that connects the macroscopic property of entropy of a system to another macroscopic property of that system, but not to a microscopic property. iii. Students can be given more opportunities to explore the connection between macroscopic and microscopic properties of a system. These opportunities could be embedded in discussions where the connection between entropy (macroscopic property) and the number of microstates (microscopic property) is emphasized. Including real-world scenarios where entropy can be concurrently related to both microscopic properties (such as number of microstates) and macroscopic properties (such as volume of a system), could help students explore ways in which entropy of that system can be explained using such properties.
0	i. Students identify the relation that connects the macroscopic property of entropy to the number of microstates of a system. ii. Students did not identify a relation that connects the macroscopic property of entropy to microscopic properties of a system. iii. Students can be given more opportunities to explore when and how macroscopic and microscopic properties of a system can be linked. Such opportunities can be situated in lessons that include discussions connecting entropy (macroscopic property) of a system to the number of microstates (microscopic property) of that system. Including real-world scenarios that help visualize this connection in discussions could help students more easily grasp the connection between macroscopic and microscopic properties of a system.

Table B.8: *Feedback aligned with the rubber balls in a box task. This task addresses the* ***Learning Performance:*** *Use a representation of a physical system to determine the number of microstates for a given macrostate to predict the system's macroscopic property of entropy. Feedback corresponds to* ***ES2:*** *The number of microstates of the given system determined from their chosen representation. Ratings of 1, and 0 align with the criteria of proficiency -met, and -not met.*

Rating	ES2
1	i. Students determine the number of microstates of the given system using an appropriate representation of that system. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
0	i. Students determine the number of microstates of the given system using an appropriate representation of that system. ii. Students did not determine the number of microstates of the given system using appropriate representation of that system. iii. Students can be given more opportunities to use a representation of a physical system to determine the number of possible microstates of that system. These opportunities could be situated in real-world scenarios that help students model and conceptualize concurrent changes to both macroscopic (volume) and microscopic (number of microstates) properties of a system. This could enable students to effectively use a representation of a system in determining the number of microstates of that system, when presented with new problems capturing real-world contexts.

Table B.9: *Feedback aligned with the rubber balls in a box task. This task addresses the* ***Learning Performance:*** *Use a representation of a physical system to determine the number of microstates for a given macrostate to predict the system's macroscopic property of entropy. Feedback corresponds to* ***ES3:*** *Prediction/Explanation about the entropy of a system. Ratings of 2, 1, and 0 align with the criteria of proficiency -met, -partially met, and -not met.*

Rating	ES3
1	i. Students make accurate predictions about the entropy of the system that align with their chosen representation. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
0	i. Students make accurate predictions about the entropy of the system that align with their chosen representation. ii. Students did not make accurate predictions about the entropy of the system. iii. Students can be given opportunities to make predictions about the macroscopic property of a system such as entropy using appropriate representations. These predictions could be embedded in real-world scenarios which include representations to determine microscopic properties of a system. These predictions should also be made along with appropriate connections between macroscopic property (entropy) and microscopic property of a system (number of accessible microstates). These opportunities could enable students to make appropriate predictions about the entropy of a system when presented with new real-world contexts.

Table B.10: *Feedback aligned with the solids in thermal contact task. This task addresses the* ***Learning Performance:*** *Analyze and interpret data about interacting systems to determine whether a thermodynamic process will happen spontaneously using the idea that entropy of the universe is maximized for spontaneous processes. Feedback corresponds to* ***ES1:*** *Statements that identify entropy as the quantity which governs spontaneous processes. Ratings of 2, 1, and 0 correspond to the criteria of proficiency -met, -partially met, and -not met.*

Rating	ES1
2	i. Students identify entropy as the fundamental thermodynamic quantity that governs spontaneous processes. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
1	i. Students identify entropy as the fundamental thermodynamic quantity that governs spontaneous processes. ii. Students identified other thermodynamic quantities as governing spontaneous processes, but not entropy (e.g., temperature differences leading to spontaneous heat flow). iii. Students can be given more opportunities to explore how entropy fundamentally governs the spontaneity of a process (e.g., spontaneous heat flow), in addition to considering factors that are consequences of this concept (e.g., temperature differences causing heat flow). These opportunities could help students activate entropy as a fundamental quantity that governs spontaneous processes while other quantities that could contribute to such processes are present.
0	i. Students identify entropy as the fundamental thermodynamic quantity that governs spontaneous processes. ii. Students did not identify entropy as the fundamental thermodynamic quantity that governs spontaneous processes. iii. Students can be given more opportunities to identify factors that govern spontaneous processes (e.g., spontaneous heat flow) by taking entropy into account. These opportunities can be situated in lessons that discuss the second law of thermodynamics which could help students prompt entropy as a quantity that governs such processes when presented with new problems.

Table B.11: *Feedback aligned with the solids in thermal contact task. This task addresses the* ***Learning Performance:*** *Analyze and interpret data about interacting systems to determine whether a thermodynamic process will happen spontaneously using the idea that entropy of the universe is maximized for spontaneous processes. Feedback corresponds to* ***ES2:*** *Statements that include the use of given representation of data by extracting the information required to determine entropy. Ratings of 2, 1, and 0 correspond to the criteria of proficiency -met, -partially met, and -not met.*

Rating	ES2
2	i. Students use a representation of data to extract information required to determine entropy. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
1	i. Students use a representation of data to extract information required to determine entropy. ii. Students used the representation of data to extract information to determine other thermodynamic properties of the system (e.g., temperature and/or internal energy), but not entropy. iii. Students can be given more opportunities to explore the ways in which information relating to different thermodynamic properties can be extracted from a representation of data. These opportunities can be embedded in a representation of data which describes concurrent changes to various thermodynamic quantities describing systems (e.g., entropy, internal energy, and temperature). Including spontaneous processes in these opportunities (e.g., spontaneous heat flow) could help students explore both how various thermodynamic quantities evolve and the most relevant quantities to extract from a representation of data that help explain such processes.
0	i. Students use a representation of data to extract information required to determine entropy. ii. Students did not use the representation of data to extract information required to determine entropy. iii. Students can be given more opportunities to use and extract information from a representation of data that describes various thermodynamic quantities of a system. Situating these opportunities in real-world scenarios that include spontaneous processes (e.g., spontaneous heat flow) could help students explore how data describes thermodynamic properties involved in such processes (e.g., entropy, and internal energy). Explicit involvement of representations of data to represent and describe the evolution of thermodynamics properties of a system could help prepare students to analyze and interpret data when presented with new problems.

Table B.12: *Feedback aligned with the solids in thermal contact task. This task addresses the* ***Learning Performance:*** *Analyze and interpret data about interacting systems to determine whether a thermodynamic process will happen spontaneously using the idea that entropy of the universe is maximized for spontaneous processes. Feedback corresponds to* ***ES3:*** *Statements that conclude spontaneous processes occur such that entropy is maximized to make judgements about the given claim or hypothesis. Ratings of 1, and 0 correspond to the criteria of proficiency -met, and -not met.*

Rating	ES3
1	i. Students make a conclusion about the occurrence of a spontaneous process such that the entropy of the universe is maximized. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
0	i. Students make a conclusion about the occurrence of a spontaneous process such that the entropy of the universe is maximized. ii. Students did not make a conclusion about the occurrence of a spontaneous process such that the entropy of the universe is maximized. iii. Students can be given more opportunities to make conclusions about the occurrence of a spontaneous process (e.g., spontaneous heat flow) such that the entropy of the universe is maximized. Embedding these opportunities in interacting systems surrounding real-world contexts which also include other conditions determining spontaneity (e.g., temperature differences) could help students compare and contrast the most relevant concepts in determining the occurrence of a spontaneous process.

Table B.13: *Feedback aligned with the semiconductors task. This task addresses the* ***Learning Performance:*** *Use mathematics to determine the number of microstates within a system to deduce the macroscopic quantity of entropy for that system and make a conclusion about the system. Feedback corresponds to* ***ES1:*** *The mathematical relationship between the number of microstates and entropy of the system. Ratings of 1, and 0 correspond to the criteria of proficiency -met, and -not met.*

Rating	ES1
1	i. Students identify the mathematical relationship (i.e., Boltzmann's equation) between the number of microstates and entropy of the system. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
0	i. Students identify the mathematical relationship (i.e., Boltzmann's equation) between the number of microstates and entropy of the system. ii. Students did not identify the mathematical relationship between the number of microstates and entropy of the system. iii. Students can be given more opportunities to explore how microscopic properties of a system determine the macroscopic features of that system. In particular, students should be given more opportunities to explore the underlying mathematical relation connecting microstates to entropy of a system (i.e., Boltzmann's equation). This could enable students to identify such mathematical relations and their applicability when presented with new problems.

Table B.14: *Feedback aligned with the semiconductors task. This task addresses the* ***Learning Performance:*** *Use mathematics to determine the number of microstates within a system to deduce the macroscopic quantity of entropy for that system and make a conclusion about the system. Feedback corresponds to* ***ES2:*** *Accurate calculation of number of microstates using appropriate mathematical tools. Ratings of 1, and 0 correspond to the criteria of proficiency -met, and -not met.*

Rating	ES2
1	i. Students accurately calculate the accessible microstates of a system using appropriate mathematics. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
0	i. Students accurately calculate the accessible microstates of a system using appropriate mathematics. ii. Students did not accurately calculate the accessible microstates of a system using appropriate mathematics. iii. Students can be given more opportunities to explore and calculate the accessible microstates of a system (i.e., multiplicity) using appropriate mathematics. These opportunities can be situated to calculate accessible microstates in both simple systems (e.g., accessible microstates for a single particle of a system) and complex systems (e.g., accessible microstates for multiple, indistinguishable particles). That could help students accurately calculate accessible microstates when presented with complex systems.

Table B.15: *Feedback aligned with the semiconductors task. This task addresses the* ***Learning Performance:*** *Use mathematics to determine the number of microstates within a system to deduce the macroscopic quantity of entropy for that system and make a conclusion about the system. Feedback corresponds to* ***ES3:*** *Accurate determination of entropy of the system using the unpacked relation and number of microstates. Ratings of 1, and 0 correspond to the criteria of proficiency -met, and -not met.*

Rating	ES3
1	i. Students accurately determine entropy of the system using the mathematical relation that connects entropy to accessible microstates of that system. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
0	i. Students accurately determine entropy of the system using the mathematical relation that connects entropy to accessible microstates of that system. ii. Students did not determine entropy of the system using the mathematical relation that connects entropy to accessible microstates of that system. iii. Students can be given more opportunities to determine entropy of the system based on the mathematical relation that links entropy to microstates of that system (e.g., Boltzmann's equation). These opportunities can be embedded in systems where accessible microstates for particles in that system are changing (e.g., thermal fluctuations can change the number of sites available for particles to occupy). Giving students opportunities to determine entropy in such systems could help them transfer these abilities when presented with new contexts. Further, giving students opportunities to engage with mathematical functionality of logarithm could also be useful when using equations such as Boltzmann's equation.

Table B.16: *Feedback aligned with the semiconductors task. This task addresses the* ***Learning Performance:*** *Use mathematics to determine the number of microstates within a system to deduce the macroscopic quantity of entropy for that system and make a conclusion about the system. Feedback corresponds to* ***ES4:*** *Statement on the interpretation of the obtained entropy value consistent with calculations. Ratings of 1, and 0 correspond to the criteria of proficiency -met, and -not met.*

Rating	ES4
1	i. Students make an interpretation about the calculated entropy. ii. Students met the expected performance. iii. No course modifications are suggested to address the expected performance.
0	i. Students make an interpretation about the calculated entropy. ii. Students did not make an interpretation about the calculated entropy. iii. Students can be given more opportunities to make interpretations or reflections about obtained entropy values. Giving students these opportunities could help them make sense of implications of obtained quantities that go beyond just calculating them (e.g., whether a process is entropically favorable or not). This could enable students to reflect on calculations in order to make conclusions about systems when presented with new contexts.

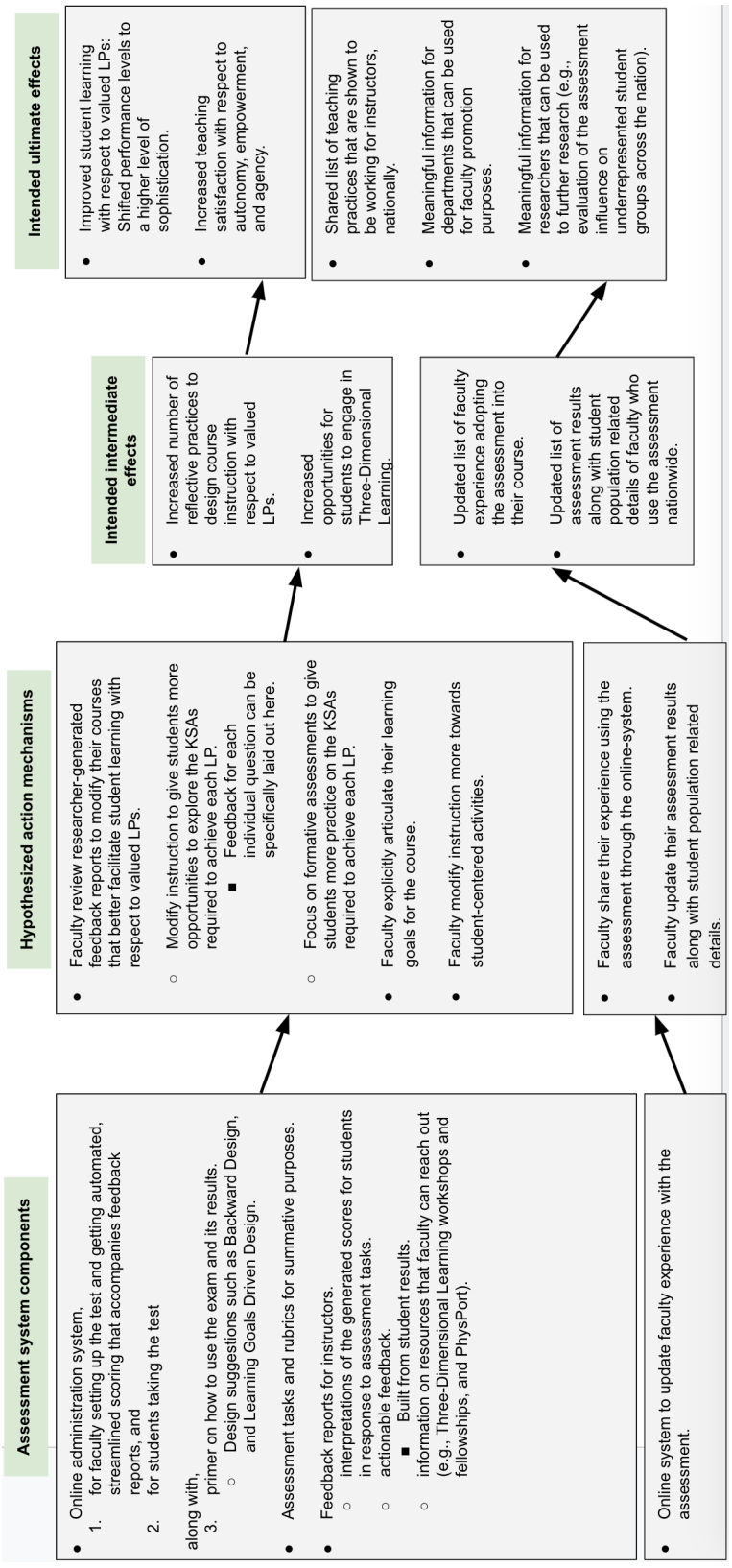

Figure B.10: *Theory-of-action for the Thermal and Statistical Physics Assessment (TaSPA).*

Q-2	Heat supplied to the system	1	Q-3	First law of thermodynamics	1
	Work done by and/or on the system	2		Equipartition theorem	2
	Temperature of the system	3		Internal energy as a state function	3
	Kinetic energies of the gas molecules	4		Ideal gas law	4
	Pressure of the system	5		PV diagram for this scenario	5
	Volume of the system	6		None of the above	6
	None of the above	7		Other: _____	7
	Other: _____	8			

Figure B.11: *Q-2: Which of the following physical quantities did you use to reason about the changes to the total internal energy of the system? Q-3: Which of the following did you use to reason about the changes to the total internal energy of the system? Answer options 1-8, and 1-7 are the reasoning elements available for students to demonstrate proficiency for ES1 in the CMR task developed based on the FR task in Fig. 4.4 in Chapter 4.*

Proficiency for ES1--- Q-2,3	
2	• **If** (1 & 2) are selected in Q-2 (without 7 being selected) along with 1 in Q-3 (without 6 being selected) (Work **AND** heat MUST be selected along with First law of thermodynamics without "none of the above" being selected in both Q-2,3)
1	• **If** (1 | 2) are selected in Q-2 (without 7 being selected) along with 1 in Q-3 (without 6 being selected) • **Else If** 1 is selected in Q-2 (without 7 being selected) along with 2 in Q-3 (without 6 being selected) (Work **OR** heat MUST be selected along with First law of thermodynamics) **OR** (heat MUST be selected along with Equipartition theorem without "none of the above" being selected in both Q-2,3)
0	• **Else** (any response patterns not captured above)

Figure B.12: *Students' selections to achieve "performance-met," "performance-partially met," and "performance-not met" with respect to ES1 are also noted in the criteria provided for "2," "1," and "0" respectively. Notations: & – AND, ˜ – NOT, and | – OR.*

If both option 1 and option 2 are selected in Q-4

5a	The expansion was isothermal	1	5b	The compression was isothermal	1
	The expansion was isobaric	2		The compression was isobaric	2
	The internal energy increased during the expansion	3		The internal energy increased during the compression	3
	The internal energy decreased during the expansion	4		The internal energy decreased during the compression	4
	The internal energy remained the same during the expansion	5		The internal energy remained the same during the compression	5
	Heat entered the system	6		Heat entered the system	6
	Heat left the system	7		Heat left the system	7
	Temperature increased during the expansion	8		Temperature increased during the compression	8
	Temperature decreased during the expansion	9		Temperature decreased during the compression	9
	Energy flowed into the gas due to work	10		Energy flowed into the gas due to work	10
	Energy flowed out of the gas due to work	11		Energy flowed out of the gas due to work	11
	No work was done during the expansion	12		No work was done during the compression	12
	Other: _____	13		Other: _____	13

Figure B.13: *Q-4: Which of the following arguments did you primarily use to reason about the changes to the total internal energy of the system? Option 1: Changes in internal energy depend on the portion of the process of expansion of the gas. Option 2: Changes in internal energy depend on the portion of the process of compression of the gas. 5a: How did you reason about the expansion? 5b: How did you reason about the compression? Answer options 1-13 are the reasoning elements available for students to demonstrate proficiency for ES2 in the CMR task developed based on the FR task in Fig. 4.4 in Chapter 4.*

Figure B.14: *Q-4: Which of the following arguments did you primarily use to reason about the changes to the total internal energy of the system? Option 3: Changes in internal energy only depend on the initial and final state of the system. 5a: Which of the following did you use to reason about the initial and final states of the system? 5b: Which of the following did you use to make your conclusion about the changes in internal energy of the system using the initial and final states? Answer options 1-10 are the reasoning elements available for students to demonstrate proficiency for ES2 in the CMR task developed based on the FR task in Fig. 4.4 in Chapter 4.*

Proficiency for ES2--- Q-4,5	
2	• **If both "option 1 AND option 2 AND/OR option 3" are selected in --- Q-4** ○ **Q-5a**: 3 & 6 & 11 & (~1,4,5,9,10,12,13) (Internal energy increased during expansion AND heat entered AND energy flowed out of the gas due to work AND NOT(the expansion was isothermal, the internal energy decreased during the expansion, the internal energy remained the same during the expansion, temperature decreased during the expansion, energy flowed into the gas due to work, no work was done during the expansion, other)) **AND** ○ **Q-5b**: 3 & 6 & 10 & (~1,2,4,5,9,11,12,13) (Internal energy increased during compression AND heat entered AND energy flowed into the gas due to work AND NOT(the compression was isothermal, the compression was isobaric, the internal energy decreased during the compression, the internal energy remained the same during the compression, temperature decreased during the compression, energy flowed out of the gas due to work, no work was done during the compression, other))
1	• **If both "option 1 AND option 2 AND/OR option 3" are selected in --- Q-4** ○ **Q-5a**: {3 & [6 | 11] & (~1,4,5,9,10,12,13)} {Internal energy increased during expansion AND [heat entered OR energy flowed out of the gas due to work] AND NOT(the expansion was isothermal, the internal energy decreased during the expansion, the internal energy remained the same during the expansion, temperature decreased during the expansion, energy flowed into the gas due to work, no work was done during the expansion, other)} **OR** ○ **Q-5b**: {3 & [6 | 10] & (~1,2,4,5,9,11,12,13)} {Internal energy increased during compression AND [heat entered OR energy flowed into the gas due to work] AND NOT(the compression was isothermal, the compression was isobaric, the internal energy decreased during the compression, the internal energy remained the same during the compression, temperature decreased during the compression, energy flowed out of the gas due to work, no work was done during the compression, other)}
0	• **Else** (any response patterns not captured above)

Figure B.15: *Students' selections to achieve "performance-met," "performance-partially met," and "performance-not met" with respect to ES2 are also noted in the criteria provided for "2," "1," and "0" respectively. Notations: & – AND, ˜ – NOT, and | – OR.*

Q-1	total internal energy decreased	**1**
	total internal energy increased	**2**
	total internal energy was unchanged	**3**
	a conclusion about the total internal energy cannot be made from the given information	**4**
	I don't know	**5**

Proficiency for ES3--- Q-1	
1	• **If** "2" is selected
0	• **Else** (any other selection)

Figure B.16: *Q-1: What has happened to the total internal energy of the gas since the beginning of the experiment? Answer options 1-5 are the reasoning elements available for students to demonstrate proficiency for ES3 in the CMR task developed based on the FR task in Fig. 4.4 in Chapter 4. Students' selections to achieve "performance-met," and "performance-not met" with respect to ES3 are also noted in the criteria provided for "1," and "0" respectively.*

Appendix C

List of Abbreviations

Table C.1: *Abbreviations used in this dissertation and explanations for them.*

Abbreviation	Explanation
3D-LAP	Three-Dimensional Learning Assessment Protocol
ACER	Activation-Construction-Execution-Reflection
CMR	Coupled-Multiple Response
ECD	Evidence-Centered Design
ES	Evidence Statement
FR	Free Response
ICLS	International Conference of the Learning Sciences
ISSS	International Student and Scholar Services
HSI	Hispanic-Serving Institutions
KSA	Knowledge, Skills, and Abilities
LP	Learning Performance
MSI	Minority-Serving Institutions
PEER	Professional-Development for Emerging Education Researchers
PER	Physics Education Research
PWI	Predominantly-White Institutions
TA	Teaching Assistant
TaSPA	Thermal and Statistical Physics Assessment