Investigating changes in scientists' ethical decision making and course design

by

Tyler Garcia

B.S., California State Polytechnic University Pomona, 2020

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

2024

Abstract

One way to bring about change in higher education is to introduce professional development programs for higher education, however these programs have been found to be ineffective at promoting positive change for individuals and departments. To address the need for better programs, I worked on two projects: one project attempts to identify a way to improve Responsible Conduct of Research training and the other project is an assessment designed to be distributed in a Thermal and Statistical Physics course that supports instructors on improving their curriculum.

Many scientists view science as value-free, despite the fact that both epistemic and non-epistemic values structure scientific inquiry. Current Responsible Conduct of Research training usually focuses on transmitting knowledge about high-level ethical concepts or rules and is widely regarded as ineffective. We argue that Responsible Conduct of Research training will be more effective at improving ethical decision making if it focuses on connecting values to science. Due to the investigation of research ethics education in physics being relatively new, we pull from philosophy and psychology to define ethical decision making using the Four Component Model. This model states that in order to make an ethical decision someone must consider four components: moral sensitivity, moral reasoning, moral motivation, and moral implementation. For this study we formed a moderated fellowship of fourteen science faculty from different disciplines who met for ten sessions over the course of a year, where they discussed the values embedded in different scientific norms. We then conducted interviews before and after the year-long fellowship that involved guided reflection of scenarios where there was some kind of ethical misconduct where the scientific practice required value judgements (e.g using unpublished data). From this data we looked at how the fellowship affected the scientists' ability to recognize ethical dimensions in their work. We found that this fellowship improves moral sensitivity, but their moral reasoning does not change. We then identified a more precise approach to looking at scientists' moral reasoning. This work can inform future ethical training to align better with what scientists value and introduce useful concepts from philosophy and psychology to education research in physics.

There are calls to create assessments that focus on gathering evidence that shows both knowledge of the desired subject and transferable skills between disciplines while providing useful feedback to instructors. To answer this call, we created a thermal and statistical physics assessment that provides evidence of student knowledge and skills in a thermal or statistical physics course that also provides actionable feedback to instructors. To create tasks, we use a knowledge-in-use framework that focuses on identifying the evidence we need to see in student answers to claim students are able to do physics, not just know physics. These "evidence statements" are the observable features students generate that show they have knowledge to complete a claim.

We need to determine a way to validate the tasks based on the focus towards obtaining evidence of student abilities when solving tasks. Current literature focuses on bringing in experts to validate whether the tasks are at the right level for the students. We are looking to expand on literature in Physics Education Research (PER) by articulating a way to validate tasks that use evidence-centered design through looking at students' evidence statements. To validate the assessment, we identified new components to gather evidence towards validation. Using these new components we introduced a new methodology to validate assessments that focus on delivering feedback through evidence. We have conducted and analyzed student think-aloud interviews answering the tasks in a free-response format or in a Coupled Multiple-Response format. We also conducted faculty interviews to see if the tasks are relevant to their courses. Through these interviews we developed a new methodology of contributing to the validation of assessments that focus on faculty feedback.

This dissertation introduces new methodologies for future researchers to improve on Responsible Conduct of Research trainings and assessment designed for supporting instructor curriculum. Through these new methodologies scientists can address the calls for better professional development programs in higher education.

Investigating changes in scientists' ethical decision making and course design

by

Tyler Garcia

B.S., California State Polytechnic University Pomona, 2020

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

2024

Approved by:

Major Professor
James T. Laverty

Copyright

Abstract

One way to bring about change in higher education is to introduce professional development programs for higher education, however these programs have been found to be ineffective at promoting positive change for individuals and departments. To address the need for better programs, I worked on two projects: one project attempts to identify a way to improve Responsible Conduct of Research training and the other project is an assessment designed to be distributed in a Thermal and Statistical Physics course that supports instructors on improving their curriculum.

Many scientists view science as value-free, despite the fact that both epistemic and non-epistemic values structure scientific inquiry. Current Responsible Conduct of Research training usually focuses on transmitting knowledge about high-level ethical concepts or rules and is widely regarded as ineffective. We argue that Responsible Conduct of Research training will be more effective at improving ethical decision making if it focuses on connecting values to science. Due to the investigation of research ethics education in physics being relatively new, we pull from philosophy and psychology to define ethical decision making using the Four Component Model. This model states that in order to make an ethical decision someone must consider four components: moral sensitivity, moral reasoning, moral motivation, and moral implementation. For this study we formed a moderated fellowship of fourteen science faculty from different disciplines who met for ten sessions over the course of a year, where they discussed the values embedded in different scientific norms. We then conducted interviews before and after the year-long fellowship that involved guided reflection of scenarios where there was some kind of ethical misconduct where the scientific practice required value judgements (e.g using unpublished data). From this data we looked at how the fellowship affected the scientists' ability to recognize ethical dimensions in their work. We found that this fellowship improves moral sensitivity, but their moral reasoning does not change. We then identified a more precise approach to looking at scientists' moral reasoning. This work can inform future ethical training to align better with what scientists value and introduce useful concepts from philosophy and psychology to education research in physics.

There are calls to create assessments that focus on gathering evidence that shows both knowledge of the desired subject and transferable skills between disciplines while providing useful feedback to instructors. To answer this call, we created a thermal and statistical physics assessment that provides evidence of student knowledge and skills in a thermal or statistical physics course that also provides actionable feedback to instructors. To create tasks, we use a knowledge-in-use framework that focuses on identifying the evidence we need to see in student answers to claim students are able to do physics, not just know physics. These "evidence statements" are the observable features students generate that show they have knowledge to complete a claim.

We need to determine a way to validate the tasks based on the focus towards obtaining evidence of student abilities when solving tasks. Current literature focuses on bringing in experts to validate whether the tasks are at the right level for the students. We are looking to expand on literature in Physics Education Research (PER) by articulating a way to validate tasks that use evidence-centered design through looking at students' evidence statements. To validate the assessment, we identified new components to gather evidence towards validation. Using these new components we introduced a new methodology to validate assessments that focus on delivering feedback through evidence. We have conducted and analyzed student think-aloud interviews answering the tasks in a free-response format or in a Coupled Multiple-Response format. We also conducted faculty interviews to see if the tasks are relevant to their courses. Through these interviews we developed a new methodology of contributing to the validation of assessments that focus on faculty feedback.

This dissertation introduces new methodologies for future researchers to improve on Responsible Conduct of Research trainings and assessment designed for supporting instructor curriculum. Through these new methodologies scientists can address the calls for better professional development programs in higher education.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First I would like to thank my advisor Dr. James Laverty for his mentorship throughout my grad school career. I remember having weekly meetings with you to discuss every little detail about my projects as I was insecure about working on my own research, but with your knowledge and guidance I was able to develop as the individual researcher that I am today.

I would then like to thank my committee members: Dr. Bethany Wilcox, Dr. Scott Tanona, and Dr. Jeremy Schmit for taking time out of their schedules to give me feedback on my dissertation. With all of your support I am able to make the best version of my dissertation and defense.

Thank you Bethany Wilcox for your guidance throughout my time on the TaSPA project and helping me organize my thoughts regarding validation. Your timely responses to my task development emails, thoughts regarding validation, and drafts of my journal rough drafts allowed me to develop my assessment developmet skills to what they are today.

Thank you Scott Tanona and Jonathan Herington for supporting my learning of philosophy and psychology. I had never taken a philosophy or psychology class before in my life, so I appreciate your patience and knowledge while working with me.

I would like to thank my undergraduate advisors Dr. Homeyra Sadaghiani and Dr. Qing Ryan. I never would have known about physics education research without your guidance in undergrad and I have always enjoyed the research since you introduced me into the field.

I would like to thank all of K-SUPER students of past and present who helped me all throughout grad school. I would also like to thank Dr. Dean Zollman for his wisdom into the field of PER. With all of your feedback I was able to create the best research that I possibly could.

I am especially grateful to grad students Bill Bridges, Lauren Carroll-Kibisov, Shams El-Adawy, and Amogh Sirnoorkar. I'll never forget all of the late nights working on homework together or all of our adventures ranging from Manhattan to Kansas City. Your friendship allowed me to thrive in grad school and become a better person.

I would also like to thank Michael Freeman for all his work he put in to making TaSPA work. I'll never forget all the late nights where we worked together to make sure the TaSPA was up to the highest of standards. TaSPA would not be where it is today without your help.

I would like to thank the most important person in the K-State physics department Kim Coy. You were always on top of my schedule more than I was and I genuinely would not be defending in the spring without your help.

Most importantly I would like to thank my family: Denise Garcia, Thomas Garcia, and Justin Garcia for their unwavering support throughout my life. Your continuous support throughout my entire life allowed me to grow and pursue what I love. I hope to continue researching in PER and making you all proud.

Chapter 1

Introduction

There is a need for change in some of the more important practices for educators in higher education.[7] For example, current higher education curriculum is reported as not relevant to students, students take a passive roll, and curriculum focuses on algorithmic problem solving which is driving off students from higher education.[7–9] Looking at responsible conduct of research in higher education shows that physicists are still committing various types of research misconduct in their work (e.g: falsely reporting authorship, misreporting data) and some physicists are experiencing harassment from people in their own research group.[10]

One way to bring about change in higher education is through professional development.[11] There are many types of professional development programs which include but are not limited to: workshops, written descriptions of effective practices, online programs, and use of expert or peer consultation with workshops being one of the more popular options.[7] Despite these many ways of introducing change into higher education, many of the current professional development programs have been found to be ineffective at producing positive change in higher education, whether it be the results of the programs are confusing to understand or the programs are not thorough enough to bring about change.[10;12–14]

To bring about change in professional development programs, we must look to see how professional development brings about change. Henderson et al. introduced a way to categorize change using two aspects: aspects of the system needed to be changed (individuals and environments and structures) and the intended outcome (prescribed or emergent).[15] Breaking down the aspects of the system, a change in

individuals refer to when the change directly impacts personal characteristics of individuals.[15] This usually refers to when instructors or students taking some sort of training to make a change. A change in the environments and structures refers to a change that impacts the system that are external to the individuals.[15] In education this often refers to the individual discipline departments where the instructors and students are affected or the university where everyone who is a member of the university are effected. For the intended outcomes, a prescribed change is when the desired final result of the individual or environment are determined at the beginning of the change process.[15] An example of this is training like Responsible Conduct of Research training where the expected outcome of this training is to reduce the instances of research misconduct. An emergent change is when the desired final result of the individual or environment is determined during the change process.[15] An example of this would be curriculum development as it is unknown if the change is successful until after the course is finished.

These aspects can then be used to create four different categories for creating change:[15]

1. Disseminating: Curriculum and Pedagogy
2. Developing: Reflective Teachers
3. Enacting: Policy
4. Developing: Shared Vision

The category of "Disseminating: Curriculum and Pedagogy" is where there is a prescribed change for the individual. This is usually where individuals like instructors or students would undergo a training designed to improve the individual (e.g: Responsible Conduct of Research Training, Title IX training, etc.). The category of "Developing: Reflective Teachers" is where there is an emergent change for the individual. An example of this would be instructors trying out a new course modification for the next semester due to assessments in the class eliciting the student knowledge from the previous class. The category of "Enacting: Policy" is where there is a prescribed change for the environment and structures. This is where there would be some sort of policy change within a department or school that would then go on to effect the individuals for the better. The last category of "Developing: Shared Vision" is where there is an emergent change for the environment and structures. This is where institutions support stakeholders to develop new environment features that focus on encouraging new teaching practices.

With the introduction of the categories of change for professional development, I introduce two projects designed to address improving professional development at the university level: the Goals and Values in Science (GVS) project and the Thermal and Statistical Physics Assessment (TaSPA) project. Since both of these projects involve instructors (individuals) changing over the course of the programs (emergent) we would categorize these projects in the change category of "Developing: Reflective Teachers". The GVS project was an attempt to address the ineffective Responsible Conduct of Research training by introducing values in science to faculty and seeing how their ethical decision-making improves based on the introduction of values. The TaSPA project is broken into two parts: the first part is the design and development of assessment tasks using an evidence-based framework and the second part is validating the task to make sure feedback given by the assessment is useful for course modifications.

The organization of this dissertation is as follows: Chapter I briefly introduces the projects and my role on them. Chapter II dives into the relevant literature regarding both of the projects. For the GVS project I discuss the literature regarding values in science and what are some of the approaches towards ethics training. For the TaSPA project I discuss the literature regarding designing an assessment with a theory-of-action framework, what are the historical components to validate an assessment, and what are the three components to classroom alignment. Chapter III goes in depth on what I did regarding the GVS project where I used the theory of Rest's Four Component Model to identify scientists' ethical decision-making. Chapter IV is where I discuss designing tasks for the TaSPA by using a knowledge-in-use framework. Chapter V is where I introduce and apply new components to the validation of the TaSPA. Chapter VI is the conclusions for the two projects.

1.1 Goals and Values in Science

The first project I worked on was an exploratory study to determine if and how scientists' ethical decision-making are affected by value laden discussions targeted for scientists. Throughout the fellowship, we never explicitly told the scientists how to make better ethical decisions or how should they think about values. Instead, we introduced the topic of values in science and had the scientists reflect on how values can affect them in their work. This is why we categorize the GVS project as "Developing: Reflective Teachers" as the target for the project was individual scientists and the intent of the training was for the

scientists to reflect more on how values affected ethical decision-making as they discussed the values that appear in science.

In 2019 the American Physical Society (APS) updated their Guidelines on Ethics to further clarify physicist's duties to avoid data fabrication, falsification, plagiarism, and abuse.[16] These updates were in response to the National Academies' Fostering Integrity in Research report, which identified a need to improve research integrity training due to the growing importance of information technology in research and the number of studies being retracted because of irreproducible data.[17] While these guidelines are an important step forward, we suggest that to implement these guidelines appropriately, physicists must cultivate their ethical decision-making.[18]

Currently, physicists are primarily exposed to ethics through Responsible Conduct of Research (RCR) training. This training, often offered in self-paced online modules, tends to focus on topics like authorship, conflicts of interest, research misconduct, and the protection of human subjects.[19] Studies have shown that this kind of online training is ineffective at improving scientists' ethical decision-making.[13;20;21] Specifically in physics, Kirby and Houle found in 2004 that around 39% of junior members of APS who responded to the survey had observed or had personal knowledge of at least one ethical violation, such as data falsification, plagiarism, or authorship showing that ethical training is needed, in physics.[12] This survey was re-administered in 2020 to junior level physicists and grad students. The survey showed mixed improvement: physicists were undergoing more ethics training (whether at the institution level or with their supervisors), but the number of misconduct violations did not significantly change.[10] Those that did notice violations reported more violations than in 2003.[10] The 2020 study also found that harassment or abuse (e.g. experiencing inappropriate remarks, treated differently, etc.) of physicists is a common experience, concentrated among women in physics.[10] In sum, there is more to be done to improve physicists' ethical conduct.

In our study of scientists' ethical decision-making, we focus on the fact that scientists routinely make value judgements in science. We define "values in science" to refer to aims, goals, or principles that direct or influence scientific work and "value judgements" to refer to evaluations of priorities between values (e.g., deciding trade offs) or of how a value is to be implemented in a decision or action. The idea that scientists make value judgements in science contrasts with the commonly held belief that social and ethical values ought not affect the agenda, methodology or dissemination of research.[18] However, more recently

philosophers of science have argued persuasively that science can never be value free, and pretending otherwise distorts scientist's sense of their own ethical responsibilities.[22;23] Ethics, after this recognition of values in science, requires attention to those values and reasoning well about them.

For instance, Douglas has argued that scientists need to attend to non-epistemic values (e.g. consequences on people's lives) as well as epistemic values (e.g. gaining knowledge) to address the "inductive risk" of making an error in either accepting or rejecting a claim.[23] When talking about epistemic values vs. non-epistemic values, we define epistemic values as indicators of truth or improving knowledge in some capacity while non-epistemic values are defined as everything else that is not an epistemic value.[24] We keep a very broad definition of non-epistemic values due to the wide range these values cover. Scientists often report the significance of statistical findings by using a 'p-value' to indicate the evidential strength of their data. The choice of a statistical significance level for a p-value (i.e. $p < 0.01$, $p < 0.05$, etc.) implies a trade off between the risk of a false positive result and the risk of a false negative result. Moreover, choices in experimental design, and sometimes during experiments themselves, influence the chances of positive or negative results. These trade offs are not merely epistemic, as false positives and false negatives can differentially affect peoples lives, directly or cumulatively, both within the scientific community and in the public more broadly (e.g., consider the replication crisis, or the consequences of requiring definitive evidence of a health risk before action). This highlights both that scientists cannot avoid making value judgements and that the scope of ethics in science involves a broader range of topics than is traditionally covered in RCR training. Building on this work, we argue that we can improve physicist's ethical decision-making by focusing on the values invoked in actual scientific practice.

We define "ethics" to refer to the domain of inquiry concerning *ultimately doing the right thing*, and "ethical decision-making" as a process (individual or collective) aimed at doing the right thing.[25] While some invoke a distinction between the adjectives "ethical" and "moral", we align with the convention in the philosophical and psychological literature on ethical decision-making that uses these words interchangeably.

Rest's theory of moral development states the Four Component Model of ethical decision-making identifies four necessary components to making an ethical decision: moral sensitivity (identifying morally important ideas), moral reasoning (identifying morally right action), moral motivation (intending to do the morally right action), and moral implementation (acting on morally right action).[26;27] (Note that while

the original name for the second component is "moral judgement", in this dissertation we call the second component "moral reasoning", consistent with some of the literature on this topic).[28] In this dissertation, we focused on the moral sensitivity and moral reasoning components in ethical decision-making. These components and their definitions will be further explained in Chapter III.

This dissertation reports on an intervention aimed at learning how to use the presence of value judgements in science to help improve scientists' ethical decision-making. This intervention took the form of a two-semester fellowship in which scientists from different fields met together to talk about scientific topics rooted in value judgements. To see if their ethical decision-making changed over the course of the fellowship, we conducted (one-on-one) interviews with the faculty before and after the fellowship to talk about their own experience of research ethics and their responses to vignettes describing situations with multiple ethical concerns identified by RCR training.

The goal of the GVS project was to identify the process of how scientists make ethical decisions in their work with the end goal of exploring how value laden discussions among scientists improved their ethical decision-making skills. My role in this project was how to determine if scientists improved on moral sensitivity and moral reasoning as this is how Rest described how people improve on their ethical decision-making.[29] I determined a new methodology for how to categorize scientists' moral reasoning as we believe we needed a more precise methodology. This work adds to the scarce literature regarding how to improve on scientists' (including physicists) ethical decision-making.

1.2 Creating Tasks for Thermal and Statistical Physics Assessment

The TaSPA aims to better support faculty instruction in a thermal or statistical physics classroom. Our target audience for developing the assessment is instructors, who are considered individuals in Henderson et al. categories of change. We also expect faculty to take the changes recommended by the TaSPA and implement these changes with the goal of attempting to improve their classroom. These changes are categorized as an emergent result as the instructors are creating curriculum changes themselves. The TaSPA is considered a "Developing: Reflective Teachers" change.

One of the main ways to measure student educational progress is to use standardized tests or assessments.[4;30;31] This can be seen in Physics where there currently are 117 assessments for all of the topics

taught in physics.[32]

While assessments are one of the main ways to evaluate student learning, assessment design can be improved.[4] One way to improve current assessment design is to focus on how a student can do science instead of of determining how much students learned in a course.[33] Another way to improve assessment design is to make scores easier to understand as many instructors have trouble with interpretation of assessment scores or the topics being assessed are not relevant to what they teach.[14;34]

One of the advancements for improving assessments is to focus assessments towards an evidence centered design framework (ECD). ECD states that when making a claim about student learning from assessment, there must be observable and defensible evidence to support the claim.[3;4;35] There are three main components to this approach: identifying the knowledge students should have (claim space), identifying what is the evidence we need to see that shows the students have the intended knowledge (evidence space), and what is the task the students will perform to elicit this evidence (task).[3] Once construction of the claims to be made from the assessment and the evidence from the students are determined, assessment tasks are designed to elicit evidence from the students.

One framework that incorporates ECD is the Knowledge-in-Use framework (KiU).[36] When designing assessments with KiU, there needs to be an identified learning performance that highlights the proficiencies required of student performance, focal knowledge, skills, and abilities that articulate the proficiencies of an assessment task, evidence statements that articulate what to look for in student answers that provide a high level demonstration of the learning performance, and task features that highlight the aspects of the task that will elicit the evidence statements from the students.[36] By designing tasks through Knowledge-in-Use, assessment developers can identify claims, evidence, and tasks to improve on instruction in a course.

In this dissertation, I discuss the design process for developing the TaSPA using the KiU framework. We designed the TaSPA with the goal of developing a thermal and statistical physics assessment that focuses on giving instructors useful feedback they can use to improve their course. This is done by the instructors selecting the learning performances that align with their learning goals for their own course and then receiving feedback with evidence on how well the students performed on each of the selected learning performances. This feedback is intended to be timely and useful by being fully online and graded automatically. Evidence of student proficiencies will be highlighted and course modifications are recommended to fix any of the areas where students may need more support. Once finished, the TaSPA will have

tasks that are relevant to the course and measure students' knowledge-in-use of thermal and statistical physic.[32;37]

In Chapter IV, I highlight how to create tasks for an assessment designed to illicit evidence for student knowledge-in-use that is translated into useful feedback for instructors. This includes the process for developing the tasks, but also developing feedback from the tasks that instructors will receive based on the evidence collected from the TaSPA. The goal of the work in Chapter IV is to address the need for assessments to give clear feedback to instructors, introduce the KiU framework to physics, and establish a theoretical approach for developing feedback.

The process for creating tasks was created by multiple people working on the project and was documented in a 2020 journal article and briefly written about in a 2023 conference paper.[1;37] My role in the development of tasks was to finish up the development of the assessment, collaborate with others in the project who were also working on tasks, and finalize in writing the process of developing tasks for TaSPA as the process has undergone multiple changes all the way up to creating the final tasks. I also finalized the process for developing feedback for the assessment by identifying which aspect of KiU correlated with the feedback statements given to instructors.

1.3 Validating Thermal and Statistical Physics Assessment

This part of the dissertation builds upon the previous section's claim of TaSPA being an approach to professional development that brings about change regarding reflective teachers. In this section we introduce the concept of validation and how we intend to use validation to bring about the best possible feedback for instructors.

In order to give useful feedback to instructors, we need to make sure that the tasks on the TaSPA are able to elicit what knowledge-in-use we want the students to have in thermal and statistical physics. This is usually done through the process of validation. Validation is using evidence and theory to interpret scores for the intended use of a test.[38] For example, if someone were to give an assessment to their students with the intention of taking the scores from the assessment and using it to improve their classes, one question they should ask is "do the exam scores accurately reflect my students' knowledge?"

Historically, validation of assessments contains four components: face validation (assessment is mea-

suring what is intended at a glance), content validation (assessment measures knowledge on topic it intends to measure), criterion-related validation (assessment scores are compared to another measure), and construct validation (assessment scores show relation between concept and consequences of scores).[39] In a typical process of validating an assessment, the face validation and content validation of the assessment are typically conducted during the development phase of the assessment while criterion-related and construct validation are conducted once the exam is complete.[38]

As we are currently working on the design of the TaSPA, we will focus on how content validation is typically considered. Content validation of an assessment means how well do the test items cover the content domain.[39] Content validation usually involves getting around 4-5 experts in the field to determine if the tasks are at the right level for students and that the tasks cover the topics in the field adequately.[39] Additionally student interviews are conducted to make sure the students can answer the tasks and that the tasks are at the right level for the students.[3]

While these are the generally accepted components to validate an assessment, these components don't exactly work for the main goal of TaSPA which is designing an assessment that will provide useful feedback for instructors. For instance, one of the key components of TaSPA is that we are designing with the idea of identifying student knowledge-in-use in the tasks. The four validation components listed above *only* considers student knowledge when validating an assessment. However, we want to make sure the assessment is validated for assessing student *knowledge-in-use*. Another way the components need to be modified is that the original four components consider how student scores reflect student knowledge, but there is no component that reflects how instructors take these scores and use it to improve on instruction.

Instead of modifying the components ourselves, we looked to the literature and found that there is a modification to the four original validation components that incorporates both knowledge-in-use into the validation as well as how the assessment affects instruction. Pelligrino et. al. introduced three new components of validation: cognitive validation (assessment measures knowledge-in-use of students), instructional validation (assessment provides timely and useful instructional information), and inferential validation (assessment scores provide accurate information about student performance).[6] In order for an assessment to be considered valid, the assessment must contain evidence of all three of the new validation components.[6] Note that while these are different components than the original four components of validation, they still contain parts of the original components. For example, cognitive validation contains

some aspects of content validation where both of the components of validation consider evidence of student knowledge. However, cognitive validation considers evidence of both knowledge-in-use of students instead of just knowledge.

To incorporate these new components to validate assessments, we are conducting student and expert interviews for the tasks to determine if the tasks are comprehensible for students and follow NGSS' three-dimensional learning.[36;40] For student interviews, we ask the students to complete the tasks while explaining their thoughts out loud. After the interviews are done, we analyzed the results to see if the students' responses align with the predetermined knowledge, skills, and abilities needed to solve the problem (evidence statements). If their work aligns with the evidence statements, then we can say that the student is able to use the correct skill to solve the problem. If the students are able to use the right knowledge, skills, and abilities that were predetermined to solve the problem and get the problem correct, or the students are not able to use the right skills to solve the problem and they get the problem wrong, then we can say that the tasks are valid. This validation evidence falls under the component of cognitive validation as we are determining if the assessment tasks are eliciting both knowledge-in-use from the students.

For expert interviews, we gave physicists who research or taught in an area of thermal or statistical physics all of the assessment tasks and asked them if the tasks were relevant to the course they taught and if there are any changes they would make to the tasks themselves. These expert interviews fall under both cognitive and instructional validation. We consider these interviews cognitive validation as we received insight into what they believe students should know or do when working through the problems based on how they taught their class. The interviews are also considered instructional validation as we received feedback to if the tasks and learning performances were relevant to their classroom which means that the feedback received from the TaSPA will be relevant to their course.

In Chapter V of this dissertation, I discuss a new methodology to analyze evidence from the TaSPA that achieves cognitive validation and intructor validation for TaSPA. I introduce a validation table, which is a novel way to visually show if an assessment task is eliciting both knowledge-in-use from students. I highlight the questions we asked experts to insure that the tasks are relevant for instructors. This work is meant to pave the way for future validation of assessments containing tasks that elicit both knowledge-in-use from students.

Chapter 2

Literature Review

2.1 GVS

2.1.1 Values in Science

Science cannot be "value free": in a recent study, Elliot noted that decisions in science are *necessarily* made with values in mind.[41] Values are involved in science when choosing research questions, handling data, drawing conclusions, communicating findings, etc. Scientists conducting vaccine trials must make tradeoffs between production speed, safety for human participants, and the reliability of the efficacy data.

One of the main arguments that science should remain value free is that people will not trust science if scientists place their own personal values in their research.[41;42] While this might be an ideal worth striving for, it is simply not possible to conduct science without making explicit and implicit value judgements. Elliot mentions in their paper that there are multiple arguments on why science can never really be value free with the gap argument (evidential gaps in data are filled by value-laden background assumptions), the error argument (when scientists face epistemic risks they ought to factor non-epistemic values into decisions), the aims argument (achieving a non-epistemic aims in science means taking into account non-epistemic values when using models, theories, and hypothesis), and the conceptual argument (non-epistemic values are relevant to assessing value-laden concepts).[41]

Findings from the developmental moral psychology literature show that changes in moral reasoning are correlated with improvements in ethical conduct in the profession.[43] Currently most ethics training

focuses on applications of ethical theories or codes of ethics, instead of discussing the values embedded in ordinary scientific practice.[13] While scientists receive training in methodologies for distinguishing factual or theoretical questions, they receive little training in how to adjudicate questions about values such as whether to replicate data through a third party or spend more money training a graduate student.

In a project looking at the values expressed by scientists' when discussing decisions within their own work, Linville et al. found that scientists view the value judgements in science as involving more than just ethical/epistemic values.[5] For example, when tasked with redoing a study or publishing the study with omitted data, the scientists considered possible funding (economic values) problems with redoing the study. The conclusions for this paper were that ethical training should focus more on the overall values in science rather than focusing on ethical theories or guidelines.

2.1.2 Assessing Ethical Decision Making

Historically someone's ethical decision making is measured through standardized assessments.[28] These assessments are mainly targeted towards the medical/dentistry field. Since these assessments have been targeted towards the medical/dentistry field, there is little literature on how scientists interact with ethical decision making.

There are a few assessments that gauge ethical sensitivity. For example, there is the Dental Ethical Sensitivity Test (DEST) and Hebert et al's medical ethical sensitivity test.[44;45] In both of these measures, the person taking the test reads a few vignettes and is asked to identify the moral issues in each vignette.

To gauge someone's moral reasoning, moral psychologists have developed the Defining Issues Test (DIT) and subsequently developed the DIT-2.[13;43] The DIT-2 is the accepted multiple choice format to rank a set of items based on the scenarios test to measure moral reasoning, which contains 5 moral dilemmas including ones like stealing food from someone hoarding food and a doctor has to decide to give a frail and suffering patient an overdose on medicine.[46;47]

In Rest's formulation of the DIT-2 there is a hierarchy between the schema; The DIT-2 is scored based of of someone's N2 score, which is determined by the degree to which post-conventional items are prioritized and personal interest items are deprioritized.[46] Another way to measure ethical reasoning capacity is to use the Ethical Decision Making Measure (EDM). This assessment is similar to the DIT

where the participant is given summaries of ethical scenarios and asked to rate the severity of the violation in the scenario on a scale of 1 (low) to 7 (high).[48]

Klinker and Hackmann used a modified version of the DIT-2 to measure school principals' moral reasoning. The study looked at 64 different state school principals with different years of employment to see their level of reasoning when asked about scenarios that could happen to them at their schools. Klinker and Hackmann found that the principals' values is one of the main factors that they use in order to make ethical decisions regarding students' welfare.[49]

Moral motivation has received less attention from the psychology literature than the other components. You and Bebeau measured moral motivation through the Professional Role Orientation Inventory (PROI). The PROI assesses a person's commitment to prioritize professional values over personal values by measuring their authority/responsibility and a person's perceptions of self-efficacy by measuring their agency/autonomy.[28;50]

Moral implementation is the least investigated component. The sole attempt at measurement was by You and Bebeau, who used dental students' scores in a professional problem solving (PPS) class. In the class students were required to implement action for complex cases that present difficult human interaction problems that can arise in dental practice.[28]

While these components can interact with each other, these components are distinct from one another, meaning that mastery in one of the components does not mean mastery in other components. You and Bebeau found that after testing students on their proficiency in each of the four components, there was no correlation between a student's competence in one component compared to another.[28]

2.1.3 Approaches to Ethics Training

One of the main ways for scientists to receive ethical training is through self-paced online modules like the CITI program.[48] There are also hybrid options available, but a recent study of university RCR training requirements found that 82% of research universities with available institutional plans for RCR training require an online-only ethics training.[51] This training has been found to be ineffective improving ethical decision-making, likely because the online training is wholly didactic, and focuses on applying regulatory rules to simple, context-free cases.[21] Current RCR training is often restricted to passive online readings

or lectures while it has been found that active learning modalities are more effective than passive learning modalities.[52;53] While it was found that longer ethical training did cause people to have better ethical decision making, many institutions do not have the resources to have these extended training sessions.[13]

Some studies have sought to change RCR training to be more grounded in actual scientific practice instead of passively focusing on what are the rules and regulations of RCR. One of these new methods is to have a virtue-based approach to teaching ethics. This kind of training involves looking at the ethical considerations scientists' encounter on a daily basis and connects RCR to this everyday ethical decision making.[54] This would allow scientists to see the connection between ethical values and the practice of science, hopefully leading to more robust and consistent ethical conduct.

One way of changing RCR training to be more grounded in scientific practice is to see the causes of unethical conduct first. Cairns et al.
conducted a phenomenography where scientists were asked what makes other scientists act unethical. She found that the two biggest perceived factors for unethical conduct are pressure and personal gain.[55] These two factors they found line up with what other literature says about what motivates people to act unethically.

2.2 Creating TaSPA

2.2.1 Theory of Action

There are calls to move the design of assessments towards gathering evidence to make inferences about student knowledge or skills.[3;36] One way to design an assessment towards gathering evidence is by using a Theory of Action framework. Typically a theory of action contains three major components: assessment components (intended components to bring about change), action strategies (actions to bring about change), and intended effects (results of intended change).[56] A theory of action in assessment is the detailed mechanisms that bring about change in the education system.[57] A theory of action for an assessment might contain the following elements: Intended effects of the assessment, components of assessment with rationale for the components, interpretive claims of assessment, action mechanisms, and potential unintended negative effects and what might be done to mitigate these effects.[57]

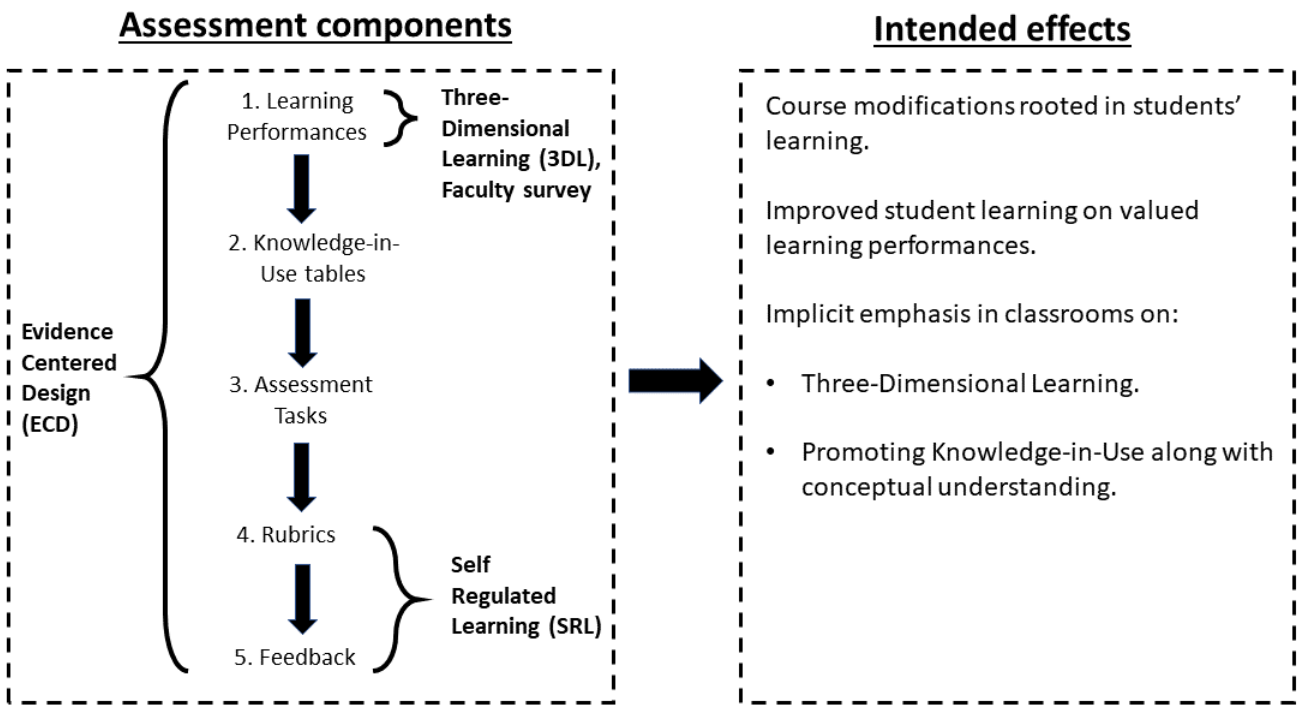

Figure 2.1: *The theory-of-action for the Thermal and Statistical Physics Assessment (TaSPA) highlighting the assessment components (on the left) and the intended effects (on the right). The arrow in between signifies faculty's uptake of our generated feedback. Note this was created before I started working on the project, but I still used this to create tasks and feedback.*[1]

Looking now at relating the theory of action literature to the TaSPA, we see in Figure 2.1 the components of the theory of action. The Assessment Components are the aspects of the assessment designed through evidence centered design. Note that in the assessment components, we design our assessment based on three dimensional learning, knowledge-in-use, and self-regulated learning. We address these topics in Chapter III. The arrow in Figure 2.1 refers to our action strategies which is the instructors will receive feedback from the assessment and suggested course modifications to address the scores from the students. The intended effects are what we intend instructors to do when they receive the feedback. These changes should create positive changes for everyone involved.

2.3 Validating TaSPA

2.3.1 Past Validation

The *Standards for Educational and Psychological Testing* define validity as the degree to which evidence and theory support the interpreted test scores for the intended use of tests.[38] In general, there are five different kinds of evidence to look for when trying to validate an assessment: evidence based on test content (description of how the test was made), evidence based on response processes (assessing the responses to see if the nature of the response is engaged by test takers), evidence based on internal structure (degree to which items/components conform to construct), and evidence based on relations to other variables (comparing scores to another standard indicator).[38;58]

One way to collect evidence listed above is to identify four different components in the assessment: face validation, content validation, criterion-related validity, and construct validity.[39] The four components are briefly described in Table 2.1.

Table 2.1: *Table showing the four components of validation used in PER and their definitions*

Validation Method	**Description**
Face Validation	Assessment measures topic at a quick glance
Content Validation	Assessment adequately measures the topic that it intends to measure
Criterion-Related Validation	Assessment scores are compared to another standard measure
Construct Validation	Assessment scores shows understanding of concept measured and consequence of uses and interpretations of results

Face validation is the extent to which a test or measure appears to measure what it purports to measure.[39;59] Since the definition of face validation is similar to the definition of content validation, some literature combines the two components into one.[39]

Content validation is defining the domain of the content being covered by the assessment and then determining if the domain of content is adequately covered by the instrument, or how well the items cover the content domain in the test.[39;59] This kind of validation is done early on in the assessment development as the assessment creators are still creating items for the assessment.[39] Content validation is usually conducted by forming a panel of experts to look at the items and determine if the instructional objectives of

the assessment are adequate enough for the targeted audience.[39;60]

Criterion-related validation is comparing measurements obtained from the assessment and comparing this to an accepted standard indicator of the concept being measured.[59] There are two types of criterion-related validation: predictive validity and concurrent validity.[39] Predictive validity indicates how well an individual's performance on the criterion compares to their performance on the assessment.[39] Experiments regarding predictive validation typically use the scores received from the assessment and compare them to the grades received in the class to see if the assessment can be used as a way to predict the success of the students. Concurrent validity indicates how well an individual's performance on the criterion at the time (e.g: grade on an in class assessment) and their performance on the assessment under validation compare.[39]

Construct validation is the understanding of the characteristic being measured as well as the consequences and uses of the results.[39;59] There are three methods to collect evidence of construct validation: intervention studies, differential population studies, and related measures studies. Intervention studies look at how does some intervention effect the assessment scores.[39] Differential population studies look at the differences in scores for different populations.[39] Related measures studies look at positive and negative correlation scores between two different assessments.[39]

2.3.2 Classroom Alignment

There are three central components that effect student learning: assessment, curriculum learning goals, and instruction.[2;61] These components all work in tandem with each other and have a reciprocal relation highlighted in Figure 2.2 (e.g: If an instructor decides to change their curriculum's learning goals for their course, then the assessment and instruction of the course should change). Classrooms with a high degree of alignment between curriculum learning goals, assessment, and instruction provide students with the most opportunities to grow their abilities in their STEM field.[61]

Curriculum learning goals are the knowledge and skills in a subject matter that teachers teach and students learn throughout a course.[2] *Instruction* usually consists of defining content area and the progressions on how students should learn the content.[2] *Assessment* is the means to measure student outcomes and the achievements in regards to the students learning in the course.[2] In undergraduate physics, this is usually

done with in-class exams required for a grade (midterms, finals, etc.).

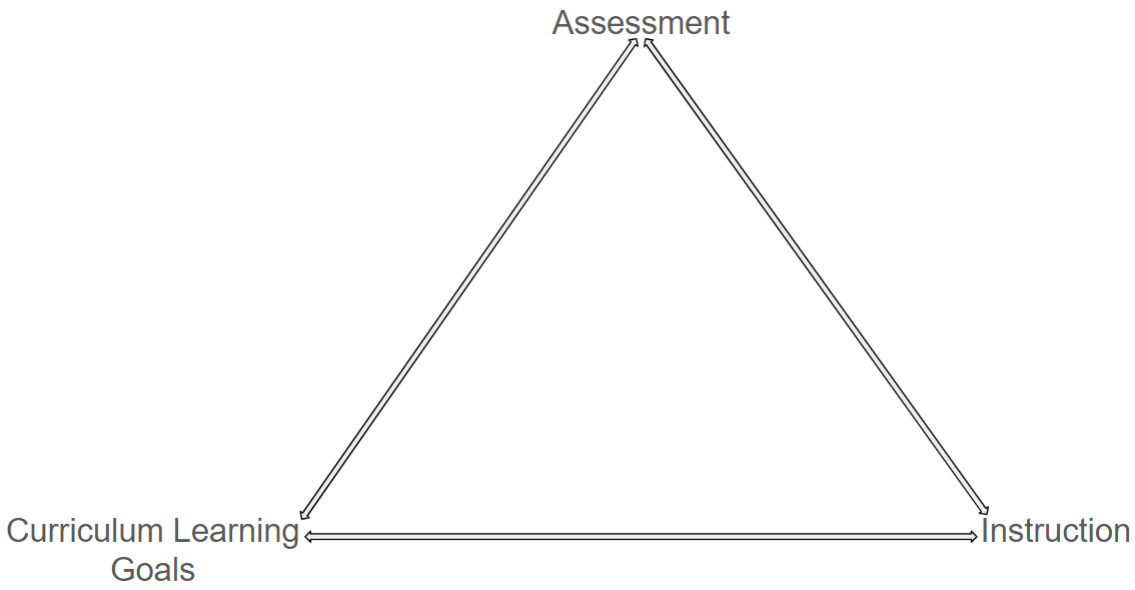

Figure 2.2: *Education system triangle highlighting how learning goals of curriculum, assessment, and instruction connect inspired by Pelligrino's curriculum-instruction-assessment triad.*[2]

2.4 Summary of Literature Review

Values are important for science and form the basis for how we looked at scientists' ethical decision making. Researchers have looked at ethical decision making changing by identifying the components of the Four Component Model as this has been done by scientists in the medical field, but it is a novel approach in PER.

The Theory of Action describes how an assessment will bring about change. While this process was developed for the TaSPA before I was involved, I used the Theory of Action approach to design feedback. For TaSPA validation, I introduced the typical components of validation collected for an assessment designed to test students' conceptual understanding in Physics. Collecting evidence using these components are I then end the literature review with the Classroom Alignment which states that the three components that effect student learning in a classroom are assessments, curriculum learning goals, and instruction as these components are the basis to the design of the knowledge-in-use framework.

In Chapter III, I discuss how we applied values and the Four Component Model to determine how scientists' ethical decision making changed after an intervention. In Chapter IV, I discuss a new methodology for creating an assessment that is designed to support instructors by giving feedback intended to improve curriculum. I highlight my role in developing feedback for the instructors. In Chapter V, I discuss the new validation methodology designed to collect evidence of both cognitive validation and instructional validation.

Chapter 3

Goals and Values in Science

3.1 Theory

3.1.1 Values

Findings from the developmental moral psychology literature show that changes in moral reasoning are correlated with improvements in ethical conduct in the professions.[43] Whereas most ethics training focuses on applications of ethical theories or codes of ethics, we believe that discussing the values embedded in ordinary scientific practice is a key pathway to improving moral reasoning among scientists. In this research, we rely on the "values in science" literature that notes the many ways decision-making about scientific methods are laden with non-epistemic (e.g. ethical, legal, or economic) values.[22;23] There are few studies about the impact of intentionally engaging scientists in these value-laden discussions.

There are multiple definitions of the concept of "value". For this study, we adopt the view that an individual values something, if the individual *prefers* that thing over some other thing.[62] Values can be "goals", "aims", "principles" or anything else individuals want to be realized. Things can be valued intrinsically, because they are preferred on their own, or instrumentally, because they help achieve other values. Values that might guide scientific decision-making include finding the truth, getting a publishable result, saving time or money, helping graduate students, providing benefits to society, or a variety of other personal, professional, epistemic, or ethical goals. In general, we distinguish "values" from "beliefs" by suggesting that the latter involves claims about purely factual matters (i.e. what *is* the case), while the

former involves claims about normative ideals (i.e. what *ought* to be the case). Note that values can be held by a community as well as individuals, and they need not be mere "opinion" or "taste". Ethical values are values, and many people hold some ethical values to apply universally.[63]

Our main focus in this chapter is on the different values scientists invoke when discussing ethics in science. While we should generally expect that scientists will be attentive to aspects of their work relevant to knowledge production, their ability to reason well depends also on their attentiveness to ethical values. Therefore, we investigated the *epistemic*, *ethical*, and other *non-epistemic* values invoked by scientists during discussions about research ethics. In this research, we relied on the "values in science" literature that notes the many ways decision-making about scientific methods are laden with non-epistemic (e.g. ethical, legal, or economic) values.[22;23]

We define "epistemic values" as motivating reasons related to the advancement or improvement of knowledge in some way: e.g., truth, objectivity, empirical accuracy, and testability.[24] These can be either directly related to knowledge (e.g., truth) or promote the attainment of it (e.g., testability).[64] We define "ethical" values as reasons related to the realization of moral principles: e.g., the impact of scientific work on human welfare, autonomy, or respect for persons. Finally, we define "other" values to encompass all other values that are not either epistemic or ethical: e.g., communitarian, adherence to legal standards or codes of conduct, economic, self-interest, and other practical concerns that were otherwise non-epistemic and non-ethical actions that are necessary for science.

3.1.2 Rest's theory of moral development

The Four Component Model (FCM) states that in order for someone to behave morally in a situation, they must demonstrate each of the following components: moral sensitivity, moral reasoning, moral motivation, and moral implementation.[26] These components are not steps that have to be followed in a linear sequence but are rather processes that collectively and interactively produce ethical action.[29] A summary of the four components for ethical decision making are found in Table 3.1.

Moral sensitivity is the ability to identify the morally salient features of the scenario. This component is where the person making an ethical decision identifies what is going on in a scenario and how a decision might affect others.[26] For example when scientists want to reproduce their results for a study, they must

Component	Description
Moral Sensitivity	Identifying morally salient ideas
Moral Reasoning	Identifying morally right action
Moral Motivation	Intending to do morally right action
Moral Implementation	Acting on morally right action

Table 3.1: *This table shows brief definitions of the four components for making an ethical decision*

evaluate the ethical dimensions of different alternative actions: training grad students to reproduce the data or send the data to a third party to reproduce the data and save money.

Moral reasoning is the capacity to identify the morally right action.[29] This component interacts with moral sensitivity since once someone has identified the morally relevant factors of the situation (people possibly affected, actions available, etc), they must decide which action or actions are morally justifiable.[26] Rest et al. developed a set of"Neo-Kohlberg" schemas to classify moral reasoning based on Kohlberg's theory of moral development.[65] Rest et al. organized moral reasoning into three schemas summarized in Table 3.2: (1) personal interest, (2) maintaining norms, and (3) post-conventional.[65] The personal interest schema focuses on the personal gains and losses that result from moral action. The maintaining norms schema involves: generally accepted social norms as a collective, norms apply society-wide, need for norms to be clear, norms establish a reciprocity, and there is an establishment of a hierarchy. The post-conventional schema involves reasoning about moral obligations based on shared ideals, fully reciprocal, and are open to scrutiny.[43;66]

Moral motivation is the component where someone intends to do what is morally right.[29] After someone identifies the morally justifiable decision, they must want to act on this decision. Rest argues that in order for someone to have a high moral motivation, they must prioritize ethical values over other values.[26]

Moral implementation is the component where a person acts on their plan of action. A person may be able to identify an action, come up with a morally justifiable action, and place a priority on moral values, but in the end they could end up not doing any action because of a deficiency in moral implementation (i.e. non-cognitive barriers to action such as weakness of will, depression, anxiety, post-traumatic stress, etc.).[26]

Neo-Kohlberg Schema	Definition
Personal Interest Schema	Doing something in hopes of personal gain
Maintaining Norms Schema	Doing something to follow set guidelines (e.g. following the law, following set science standards)
Post-Conventional Schema	Doing something because it is ultimately the right thing to do

Table 3.2: *This table shows brief definitions of the the schema for moral reasoning*

3.1.3 Improving Ethical Decision Making

The main reason for using the FCM is that it provides a schematic for identifying the components in a person's ethical decision-making. While we can identify values to identify the connection someone makes between ethics and science, we need to also use the FCM to identify how someone makes a decision in science. Identifying a person's components of their ethical decision will allow us to focus research and interventions to strengthen any deficiencies in any of the components. For example, some research has identified that an improvement in someone's moral sensitivity will lead to an overall improvement in ethical decision making, indicating that training should focus primarily on improving someone's moral sensitivity.[67;68]

Narvaez and Rest claim that a person who fails to make the morally right decision has a deficiency in one or more of the ethical decision making components.[27] For example, someone can identify all of the morally salient ideas in a scenario but ultimately cannot come up with a reason to make a decision. Since all components are a necessary part of the process for making an ethical decision, an improvement in one component **does not guarantee** an improvement in someone's ethical decision making.[27] However, we take the position that improving in one or more components is a step in the right direction towards morally acceptable decisions.

We identify how scientists' are improving with respect to individual ethical decision making components. Since Rest defined moral sensitivity as identifying the morally salient ideas in a scenario and the literature states that epistemic and non-epistemic values are important for making decisions in science, we define an improvement in scientists' moral sensitivity as identifying a greater number of values when discussing ethically-laden cases.[22;23;41] "More values" in this case means both quantitatively more values but also more relevant values (e.g. instead of invoking purely epistemic values, participants should invoke

ethical, epistemic, and other values in their sensitivity).

We define an improvement in moral reasoning as reasoning with a higher level schema. As mentioned above, in the DIT-2, the post-conventional schema is defined the as the highest level reasoning schema, followed by the conventional and the personal interest schemas.[46] For example, someone's initial reasoning for a scenario starts in the self-interest schema but over time the reasoning changes into the post-conventional schema. This means that their moral reasoning improved. In Rest's case, he applied these schema to multiple choice questions. In this study, we applied the same schema to participant's open-ended responses to ethical vignettes.

We applied these theories to answer the following research questions:

1. How do discussions of values embedded in science shift scientists' *moral sensitivity*?
2. How do discussions of values embedded in science impact scientists' *moral reasoning*?
3. How effective is the value-focused training in improving scientists' overall ethical decision making?

3.2 Methods

3.2.1 Fellowship

Our project is structured around interviews conducted as part of a year-long fellowship for 15 faculty members in the basic sciences at a midwestern research university. Participants were recruited by email, word of mouth, and explicit invitations by the fellowship organizers. Recruitment intentionally tried to promote diversity in terms of gender and academic status.

The fellowship consisted of 10 sessions over the course of an academic year. Each fellowship session was one and a half hours long, and centered on reading materials given to participants prior to the session. The fellowship did not explicitly teach the participants about ethics, but was designed to foster directed exploration and discussion of the role that values play in decisions made within science. The contexts ranged from foundational methodological decisions, such as the choice of statistical methods, to decisions about the application of science, such as whether and how to advocate for policy changes. Throughout the fellowship, the participants were made aware that they were explicitly addressing value questions.

They knew the context of the research project was about ethics in science, but the research goals were not discussed.

One moderator ran the sessions and helped guide discussions. Each session focused on a particular topic meant to highlight the role of values in scientific decision-making. The sessions were designed to provide some basic conceptual tools for understanding and potentially addressing issues. Participants were asked to read and come prepared to discuss pre-readings they were provided for each session. During the sessions, the participants were prompted to discuss topics with open-ended, general questions inviting them simply to react or discuss what they thought was interesting, as well as with directed questions meant to elicit discussion of particular topics. At the end of the session, they were asked to write a follow-up reflection, which asked for general thoughts about the session as well as for answers to specific follow-up questions. The 15 fellowship participants were split into two groups of 8 and 7 respectively for both for ease of scheduling and to ensure more chance for discussion for each participant. Halfway through the fellowship (after five sessions), the groups switched some members for both for scheduling purposes and to provide broader exposure to other faculty. Meetings were held in person, until the COVID-19 pandemic required meetings by Zoom for the last three sessions.

The first introductory session introduced participants to ethical theories and other philosophical concepts (such as descriptive and normative distinction). After that, each session focused on a particular topic or two that highlighted the potential role of values in scientific decision-making, including a number of topics from the "values in science" literature (e.g., inductive risk, choice of statistical methods, social structure of science, reproducibility) and topics that highlighted the connection between science and society (e.g., advocacy, diversity, public participation in science, public sharing of pre-prints). In the final sessions, discussion turned inevitably to the COVID-19 pandemic, during which previous concepts were applied.

These topics were chosen, and sessions were designed to encourage participants to reflect on and communicate both the long-term, immediate, intermediate, and instrumental goals of their own work. These topics highlighted epistemic values (including choices between them, e.g., between avoiding type I or type II error, or between modeling for predictive accuracy or representational accuracy), as well as relationships between epistemic goals and ethical values (e.g., in "inductive risk"). This allowed discussions about the goals of science as a whole, disciplinary standards, general disciplinary goals, and impacts on society.

This same fellowship is explored in other works.[5;55]

3.2.2 Data Collection

Our data for this project comes from 29 individual, semi-structured interviews of fellowship participants conducted both before and after the fellowship series, respectively called the pre-interview and the post-interview. All interviews were video and audio recorded. Due to one of the fellowship participants dropping out and being unable to make time for a post-interview, we only analyzed the pre- and post- interviews for 14 of the 15 participants so we only looked at data from 28 interviews.

In these interviews, the interviewee was asked about the ethics they consider in their own work as well as several questions to respond to a series of fictional vignettes involving research misconduct (named "RCR Vignettes"). The interviewer asked questions related to the vignettes to elicit more detailed responses from the interviewee (e.g. "Can you explain more about ..."). In total, there are three vignettes with six different ethical dilemmas within these vignettes (dilemmas: student concern with the experiment, publishing possibly wrong data, faculty taking authorship over student, working a similar experiment with another group, student getting a hold of unpublished data, hiring diversity). Two of the three vignettes are adapted from the ethical decision making measures.[48] The names and brief descriptions of these vignettes and dilemmas can be found in Table 3.3. Note that while these vignettes are not specifically related to physics, these concerns are relevant to physicists.[12]

We looked at both the categories of interview questions regarding their answers to the ethical considerations in the scientists' work as well as the answers to the vignettes. We called the first category of questions *general experience* questions since they are questions that ask about the participant's views on, and experience with, research ethics. The questions for the *general experience* section included questions like "What does ethics mean to you", "What kinds of ethical issues do you run into in your research?", "What about publishing? Do you run into any ethical issues in that area?", and "What about working with your students/postdocs/mentees? Any sorts of ethical concerns there?"

We called the second category of questions the *RCR vignette* section, which involved fictional scenarios where scientists were faced with an ethical choice regarding the conduct of their research. The interviewee was asked what would they do in that situation, and to describe their reasoning. The RCR vignettes

Title	Description
Informed consent	A professor wants to conduct an experiment where he gives a placebo shock to participants, but does not tell the participants about the placebo. The graduate student working on the project does not want this project to go on.
Publishing concern	The previous experiment ends up being conducted and two groups are working on the data. One group believes that their data is incorrect, and both groups want to publish the correct data.
Authorship concern	The professor promises a grad student that they would be first author on the project, but he then takes back this promise because he needs more first authorship.
Review similar experiment	A professor is asked to review a project, but the project is very similar to the one he is working on.
Unpublished data	The professor found out that one of his students got a hold of unpublished data from another student and is trying to use the data in their own research.
Hiring diversity	A professor department is hiring a new faculty member. They can only bring in three candidates, and they decide to bring in three male candidates and leave out a female candidate.

Table 3.3: *This table shows the vignettes and ethical dilemmas from the interviews.*

included three different scenarios with questions regarding the scenarios: the first scenario was about faulty data, the second scenario was about teams working on a similar project with authorship concerns, and the third scenario was about diversity in hiring.

These two categories provide different ways of eliciting views about ethics, and could elicit different aspects about the participants' reasoning. The first category asks questions that are likely to generate real examples from their scientific practice but with no focus on particular types of concerns. The second category might have common reasoning across all of the participants, but because of the artificiality of the fictional vignettes, might not be as meaningful to participants.

3.2.3 Data Analysis

We transcribed all 28 (14 pre-interview plus 14 post-interview) of the video and audio recordings using otter.ai as the initial transcriber. We then corrected the responses starting from the first question of the interview and ended as soon as the questions about vignettes were done. We checked to see if time was a factor in the changes between moral sensitivity and moral reasoning, and we found that the average inter-

view length for the vignettes was about thirty minutes for both the pre-interviews and the post-interviews (the average interview length for the pre-interviews was 29:41 while the average interview length for the post-interviews was 30:30).

Rest defined that the Four Component Model is the process that people undergo when given a *scenario* and the general experience questions did not involve asking the participants a scenario. We looked at the responses for the RCR vignettes instead of both the general experience questions and the RCR vignettes. While this method of only looking at the number of values was refined later to include analyzing both number of values and relevant types of values, I report the results of just the numbers analysis as it helped me further refine my method for analyzing the scientists' ethical decision making.

In order to see what *types* of values scientists appealed to when thinking about research ethics, we identified the explicit and implicit content of the values they appealed to when discussing the vignettes. We are primarily interested in the extent to which scientists invoke epistemic, ethical, or other non-epistemic values. We first took all of the responses from the interviewee and split them up into "quotes". One "quote" was the part of the response that expressed a value or values.

We identified values from quotations by identifying expressions that contained normative or evaluative judgements. A feature of the statement was judged as valued in a statement if it could be fit into a form such as "The interviewee values..." or "the interviewee has a goal of ...", whether or not the person explicitly held that value themselves or was discussing someone else's values or possible values. For example, if one of the participants said "I guess that means doing what's right, in particular situations and what, what's, what's good and right, and most beneficial to whoever might be impacted by whatever the decision is, or whatever the process is, that's being done" in response to the question "What does ethics mean to you?", we would say that this person values *doing what's right* which would be an example of an ethical value.

Moral Sensitivity

We focused on changes in scientist's "moral sensitivity" by looking at the number of values the scientists invoked since these are the "morally salient features" of the scenarios that we gave them. We categorized the values into ethical (appeal to what is the right thing to do), epistemic (appeal to improving knowledge), RCR (appeal to regulations and guidelines), legal (appeal to threat or punishment by governing entity),

Category	Subcategory	Definition
Ethical	Ethical, misc.	Doing what is right. This subcategory was used when the interviewee valued doing what is right, but it did not fit in the other subcategories
	Rights	Treating people in a certain way due to some intrinsic feature
	Fairness	Treating others in the same way
	Social Good	Doing something that will benefit society or a group of people
	Virtue	Fulfilling a character trait that the interviewee found desirable
	Interpersonal Care	Maximizing welfare for one specific person
Responsible Conduct of Reasoning (RCR)	RCR	Favoring regulations set by RCR training (plagiarism, authorship, etc)
Legal	Legal	Worrying about threats by government regulations.
Communitarian	Communitarian	Doing something good for the desire for peer/social approval.
Epistemic	Epistemic, misc	Advancing or improving science in some way. This subcategory was used when the interviewee valued improving science, but it did not fit in the other subcategories
	Alethic	Pursuing or clarifying knowledge about something
	Explanatory	Understanding some process
	Methodological	Understanding a part of the scientific process
	Aesthetic	Having a theory or explanation that is pleasing to the interviewee
	Predictive	Conducting science to make predictions about the future
	Empirical	Disseminating data or proper use of data
	Technological	Using science to have some technological application
Economic	Economic	Using fixed resources (time, money, etc)
Self-Interest	Self-Interest	Action providing some benefit to the interviewee
Practical	Practical	Actions that must be done for science to be conducted

Table 3.4: *This table shows the categories, sub-categories, and definitions for the values used in the project. This table is inspired by Linville's values table.*[5]

communitarian (appeal to social approval), economic (appeal on using resources), self interest (appeal that benefits self), and practical (appeal to actions that must be done for science). Before we even looked at the interviewee transcripts, we identified a subset of values identified in Kohlberg's stages of moral development, which went on to influence the moral reasoning schema. These values are ethical, RCR, legal, communitarian, and self interest. As we coded the quotes, other values appeared that did not fit within the original values identified. These values are epistemic, economic, and practical values are not referenced by Kohlberg. Within the ethical and epistemic categories are sub-categories that further defined the categories of these values. The definitions for the categories and sub-categories are in Table 3.4.

We also investigated the number of *unique* values (i.e. distinct, separate values not mentioned previously in the scenario). We started by looking at the quotes to see the sub-category of the value for the quote. If there are two values with the same sub-category within the same scenario, then we would return to the original transcript and look at the context for the value. For example, if someone values "taking care of their students" followed by valuing "publishing accurate data" and then said they valued "taking care of their students" in the same scenario, there would be three total values since stated values three times, but there would only be two unique values since they are valuing one of them twice. For this analysis, we considered unique values *per person*, so if one of the participants said that they value "taking care of their students" and another participant said that they also value "taking care of their students", then these would be considered two different unique values when looking at the data since these values come from two different people. Since we are saying that Moral Sensitivity is being able to identify the morally salient features in a scenario, we looked to see if the participants are invoking a wide range of values instead of repeating the same value over again. Below are example quotes for coding values:

Ethical Example: In this quote the interviewee is responding to the question "What should Dr. Kaylee do?" about trying to convince other people to bring in a woman to the hiring process:

> *"So I mean, I guess if they're all roughly equal I mean he should probably talk to the chair and get the woman in as a candidate...**It's kind of really important to have the same sort of representation as the general population**. It's gonna be problematic if you don't have that."*
> - Interviewee 2

The interviewee's response shows that they think having a diverse representation in the department is important and that the department should bring in the woman candidate. This shows that they value the importance in having proper and equal representation in a department and since this would ultimately benefit other people in the department this would be an ethical value with the subcategory of "fairness".

Epistemic Example: In this quote, the interviewee is responding to the interviewer's statement "Tell me why you shouldn't throw out data" when following up on the question "What should they do?" regarding the discrepancy between the two group's data:

> *"So having data sets that **acknowledge all of your data, how it was collected, and then that it's been thoroughly vetted** so that you've done your best effort possible to make sure that*

there haven't just been mistakes in the synthesis of these data sets. ***I think it's integral to getting that data****."* - Interviewee 14

The interviewee responded by saying that all data and the collection and analysis of the data is "integral" and should not be thrown out. From this quote we see that the interviewee values having proper collection/analysis of all data which would fall under the epistemic category with the subcategory of "empirical".

Economic Example: In this quote the interviewee is responding to the interviewer's statement

"So I would do that no matter what, it's just then, you know, is there ***some aspect of that that's fundable and worthy****. . . "* -Interviewee 9

We interpret that there is something that is "fundable" for the person in the scenario which is an emphasis on the economic value.

Since we define an improvement in moral sensitivity as participants invoking more epistemic and non-epistemic values, we are interested in the difference in the *number* of values between the pre-interviews and post-interviews. We did not assume any hierarchy among the values, and thus focused on how many values are invoked and how many values of each category are invoked instead of comparing what it means for one category of values to be invoked. An analysis of invoking epistemic/ethical values versus RCR values can be found in Linville et al.[5]

Due to the low numbers of participants and the uncertainty of our normalized data, we considered running a parametric statistical test (t-test) vs. a non-parametric statistical test (Wilcoxon signed rank test) to determine if the changes between the pre-interviews and post-interviews are significant. After running both kinds of the statistical tests on the data, we found no significant difference in the results so we report the statistically more powerful parametric statistical tests mentioned below.

We used a paired t-test to determine if there are any significant differences between the pre-interview values invoked and post-interview values invoked. This test determines if there is a mean difference between observations for two pairs of data (in our case the pairs would be the pre and post vignettes).[69] Due to the low numbers of scientists, we added up all of the scientists' invoked values per ethical dilemma for the pre and the post interviews. We then ran a paired t-test on the total values invoked per dilemma and unique values invoked per dilemma to find the p-value for the differences in values invoked. This allows us to see if there are any significant differences between the pre and post invoked values. If there are any

significant differences between the pre and post invoked values, we can then make claims about how the value focused discussions effected the participants' moral sensitivity.

Moral Reasoning

Originally we focused on changes in scientist's "moral reasoning" by looking at how the participants are invoking Neo-Kohlberg Schema in their reasoning to questions about the vignettes. We identified the participants' moral reasoning by classifying a claim made by the participants followed by the reasoning for that statement.

An example of this identification is found in "they should stop the experiment because harming people is not the right thing to do". In this quote, the claim made is that "they should stop the experiment". The reasoning following the claim is "harming people is not the right thing to do". We only applied the Neo-Kohlberg Schema to the participants' reasoning and not the claims they were making.

After coding each of the individual participant's responses, we combined all of the interviewee responses to see how the scientists' moral reasoning as a whole changes similar to what we did for the values coding. Below are examples of coding participant's moral reasoning:

Post-Conventional Example: In this quote the interviewee is responding to the question "What should Fowler do?" about being first author on the paper instead of his graduate student:

> *"I mean, this is why students are doing the work in people's labs, it's not for the stipend they received day to day. It is for the professional accomplishment that they get from authorship on paper in terms of its ability to step them up into a good post or a good faculty position is of critical importance to his trainees"* - Interviewee 7

In this quote we classify the claim as "students are doing work in people's labs". The **reasoning** is the "professional accomplishment" they get from being first author. This accomplishment will be for the benefits of their grad students and is an example of post-conventional reasoning.

Maintaining Norms Example: In this quote the interviewee is responding to the question "What should Stavenick do?" regarding a grad student using unpublished work in their research:

> *"So I mean, that would be a discussion along the lines of you, you know, you can't do this because you're not allowed to see this proposal in the first place."* -Interviewee 12

The claim in the quote is "[the student] can't do this" study. The **reasoning** is the grad student is not allowed to do see the proposal in the first place. The use of "not allowed" is evidence the interviewee is referencing some rules that prevent the grad student from seeing the unpublished data in the first place.

Since we are identifying an improvement in moral reasoning as participants using higher level schema in their reasoning after the fellowship, we are analyzing the participants' *percent distribution* of schemas to determine if the participants' reasoning changed. Since we are now looking at which specific schemas the participants are using due to the hierarchy of the schemas mentioned in the theory section, a percentage breakdown of each schema the participants use will more accurately reflect on their change in moral reasoning. We acknowledge that for improvement in moral sensitivity we are looking at *total number and types of values* and for improvement in moral reasoning we are looking at *percent distribution*, however due to the independence of improvement on the components we argue that we can analyze the two components for improvements in different ways. We are identifying the participants' moral reasoning this way as we are attempting to identify a way of identifying moral reasoning in a more asset based model, which the DIT and EDM do not fit in.

We used Fisher's exact test to compare the distributions of schema between pre-interviews and post-interviews.[70] Fisher's exact test is used to compare the distribution of a categorical variable in a group with the same variable of another group.[71] In our case, the categorical variable is the schema invoked and the two groups we're comparing are pre-interviews and the post-interviews. If the Fisher's exact test shows a significant difference in the distribution of schemas between the pre- and post-interviews, then we argue the change was a result of the fellowship.

We also looked at the changes in total reasonings after the fellowship to confirm the independence of the Four Components through identifying values. Once we get the number for the total schemas invoked between pre and post-interview per dilemma, we ran a paired t-test to find the p-value for the differences in the schemas invoked. If the p-value is less than 0.05, we consider the difference significant and conclude that the number of reasonings between the interviews increased.

Inter-rater reliability

We determined the agreement of the values and Neo-Kohlberg schema through inter-rater reliability (IRR). Three people coded the interviewee's responses. One coded all of the quotes for both values and Neo-Kohlberg schema and will be labeled the *original coder*. The other coders received a subset of these quotes to determine the value and schema of these quotes. For IRR for coding the values in general, the original coder sent out around 20% of the total value quotes coded (around 121 quotes) and the other two researchers coded values for these quotes.

We used Fleiss Kappa to determine the rater agreement between the three researchers for the individual values.[72] Calculating out the Fleiss Kappa gives a number between 1.00 and 0.00 where 1.00 means the coders had perfect agreement between codes and 0.00 means the coders had no agreement between codes.[73;74]

The Fleiss Kappa value for the reliability is around 0.53, which according to Landis and Koch is moderate agreement between the coders.[75] The two researchers met up with the original coder and discussed their codes. After discussing the values the coders agreed on the same value for each quote until there was a 100% agreement between the coders for the values for each quote.

For IRR for the unique values, the coders went through each ethical dilemma and discussed what they thought were unique values. We looked for agreements in the total number of unique values by discussing what were the unique values in each dilemma. After going through one of the interviewee dilemmas that were sent out and the coders reached 100% agreement on each of the unique values, the coders continued to the next one until 100% agreement was reached for each unique value on all of the dilemmas.

For the IRR for the moral reasoning schema, we used a similar method to the total number of values where the original coder sent out around 20% of the total reasoning quotes (around 84 quotes) and the other two researchers coded schema for these quotes. The Fleiss Kappa value for the reliability is around 0.39, which according to Landis and Koch would be a fair agreement between the coders.[75] After the Fleiss Kappa value was calculated, the two researchers met up with the original coder and discussed the schema they coded. The coders then discussed the agreed upon a schema for a quote and disputed any differences. After discussion there was a 100% agreement between the coders for each each schema for the quotes.

3.2.4 Limitations

The scope of our conclusions are limited because we have a relatively small number of faculty from one university in this study. This only allows us to look at a small section of the scientific community and we therefore hesitate to make broad claims from this. Additionally, we acknowledge that the values participants' invoked could have changed due to factors outside the fellowship.

We also acknowledge that due to the one-on-one nature of the interviews, the participants may have displayed some sort of social approval bias where they mainly talked about doing what is right. This may raise the total post-conventional reasonings which is outside of our control though we expect both pre and post interview reasonings are modified the same way.

3.3 Results

3.3.1 Moral Sensitivity

These are the results from when we analyzed both the general experience questions and the RCR vignettes at the same time. Originally we looked at three value categories: ethical, epistemic, and *other*. We looked at ethical and epistemic values while grouping up the other values due to the small amount of times these other values were invoked individually. This then led to discussions on what does it mean to invoke moral sensitivity when making a discussion and we came to the conclusion that moral sensitivity includes both the number of values and the type of values, so we included the individual categories in later work.

GE	Pre-Interviews	Post-Interviews	Total
Ethical	79	144	223
Epistemic	173	97	270
Total	319	284	493

Table 3.5: *Contingency table for the general experience questions values.*

RV	Pre-Interviews	Post-Interviews	Total
Ethical	105	249	354
Epistemic	63	121	184
Total	168	370	583

Table 3.6: *Contingency table for the RCR vignette values.*

Tables 3.5 and 3.6 show the distribution of values that we found for both the *general experience* section and the *RCR vignettes*. The total number of values for the post-interviews is higher than the pre-interviews. There are more values invoked in the vignettes when compared to the general ethics questions. Breaking down the numbers, there were more values invoked when talking about general research questions in the pre-interviews. Our Fisher's exact test revealed that the percentage of ethical values compared to epistemic values increased significantly in the *general ethics* section of the interview after the fellowship ($p < 0.001$).

When it comes to the vignettes, the post-interviews have over double the values that were invoked when compared to the pre-interviews. The Fisher's exact test showed that the percentage of ethical values compared to epistemic values did not have a significant change in the *RCR vignettes* ($p = 0.28$).

Refined Results

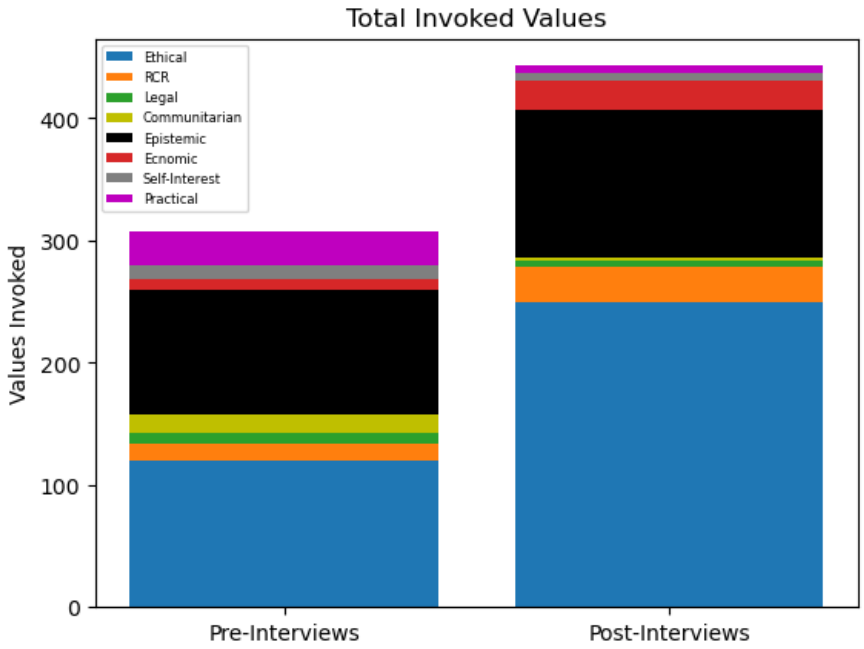

Figure 3.1: *Total number of invoked values for the pre-interviews and post-interviews.*

Table 3.7 shows that the most prevalent values to show up are the ethical values and the epistemic values. While these two categories show up the most, there is only a change between the pre-interviews and the post-interviews for the ethical category. The ethical category had the largest increase in number of values invoked while every other category stayed roughly the same. There is an increased number of

Participants	Pre Value	Post Values
Participant 1	30	25
Participant 2	19	28
Participant 3	34	34
Participant 4	23	29
Participant 5	25	45
Participant 6	33	39
Participant 7	21	50
Participant 8	23	32
Participant 9	17	26
Participant 10	26	24
Participant 11	16	22
Participant 12	15	33
Participant 13	16	29
Participant 14	10	28

Table 3.7: *Total values invoked per participant from the interviews.*

values invoked in the post-interview (443 vs. 308).

We found that the paired t-test for the total number of values from Table 3.7 gives us a p-value of 0.002. More explicitly, the number of post-interview values invoked are significantly higher than the pre-interview values invoked.

Looking at Table 3.8, we can see that the participants are invoking a variety of values instead of repeating the same values. The trends here are very similar to what we see when looking at the total number of values invoked. There is an increase in the total number of unique values invoked (360 vs. 270). Once again, the biggest increase in unique values was in the ethics category while the rest of the categories stayed relatively the same.

We found that the paired t-test for the unique values gives us a p-value of 0.008. This means that there is a significant difference between the post-interview unique values and pre-interview unique values and that the participants are invoking different unique values after the fellowship.

Looking at both Figure 3.1 and Figure 3.2 we see that the total number of values invoked by the participants increased. More specifically, the total number of ethical values invoked is more than double in the post-interview than the pre-interview. To see if the change in ethical values is significant, we ran a t-test on the ethical values invoked found in Table 3.9 and we found a p-value of 0.00003. This shows that the participants are noticing more ethical values in these ethical scenarios.

Participants	Pre Value	Post Values
Participant 1	26	20
Participant 2	15	22
Participant 3	28	30
Participant 4	21	21
Participant 5	21	35
Participant 6	29	31
Participant 7	20	43
Participant 8	20	23
Participant 9	16	22
Participant 10	22	21
Participant 11	14	17
Participant 12	13	27
Participant 13	16	25
Participant 14	9	24

Table 3.8: *Unique values invoked per interviewee from the interviews.*

Participants	Pre Value	Post Values
Participant 1	10	12
Participant 2	7	13
Participant 3	13	27
Participant 4	9	20
Participant 5	8	28
Participant 6	15	23
Participant 7	6	25
Participant 8	10	17
Participant 9	5	12
Participant 10	11	12
Participant 11	6	11
Participant 12	7	18
Participant 13	7	16
Participant 14	6	15

Table 3.9: *Ethical values invoked per interviewee from the interviews.*

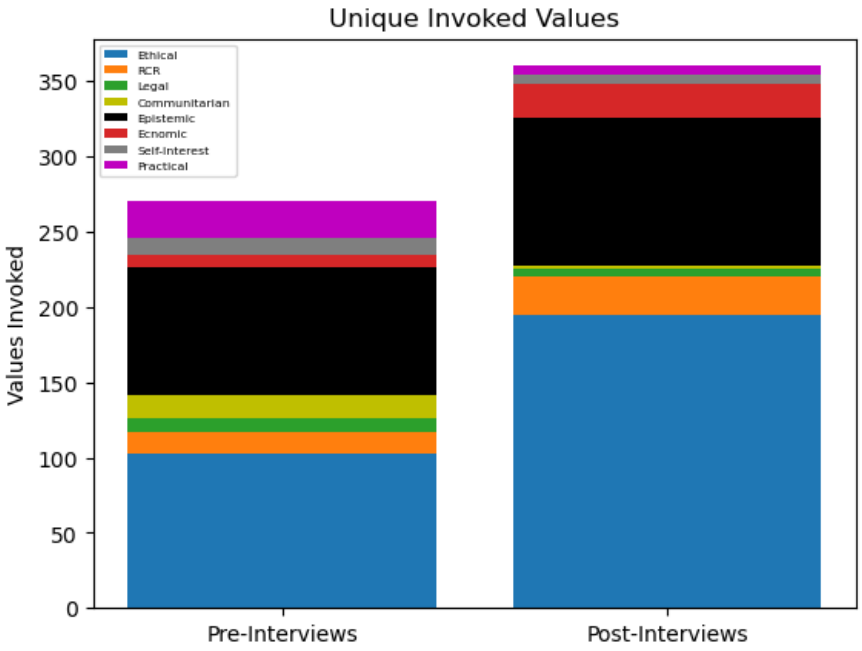

Figure 3.2: *Unique invoked values for the pre-interviews and post-interviews.*

3.3.2 Moral Reasoning

Figure 3.3 and Figure 3.4 show there is no significant change in the distribution of schemas between the pre-interviews and the post-interviews for the participants' moral reasoning. We can see from these figures that the Self-Interest schema did not change with staying at 7% for both the pre-interviews and post-interviews, while the Maintaining Norms schema and the Post-Conventional schema showed no significant change between the interviews.

The Fisher's exact test for the *distribution* of schemas invoked gives us a p-value of 0.70. This means that the difference between the distribution of pre-interview schemas and post-interview schemas is not statistically significant.

The paired t-test for the *total number* of schemas invoked gives us a p-value of 0.004. This means that the difference between the the total post-interview schemas invoked and total pre-interview schemas invoked are statistically significant. More explicitly, the number of post-interview schemas invoked are significantly higher than the pre-interview schemas invoked.

Figure 3.3: *Pie chart for percent breakdown of schema for levels of reasoning for the pre-interviews. We combined all participants into one graph since we are looking at scientists as a whole.*

3.4 Discussion

One of the goals for this project is to introduce theories from philosophy and moral psychology to Physics Education Research in order to deconstruct scientists' ethical decision making. We found that one of the better ways to categorize ethical decision making is by using the Four Component Model due to the distinct and easy to recognize components. We believe that as interest in research ethics education for physicists grows, the Four Component Model (and theories like it) will provided much-needed theoretical and methodological structure to attempt to understand, measure, and improve research ethics education.

As discussed in Linville et al's paper, scientists invoke both epistemic and non-epistemic values when discussing ethics *in their work* before the fellowship.[5] If the scientists truly valued RCR training, then we would have seen more invoked RCR values from the participants.[5] This is further highlighted in this chapter as the participants are invoking significantly more ethical values and not RCR values when discussing vignettes instead of their own work after the fellowship. This further highlights the need to improve on current RCR training by focusing more on values in science.

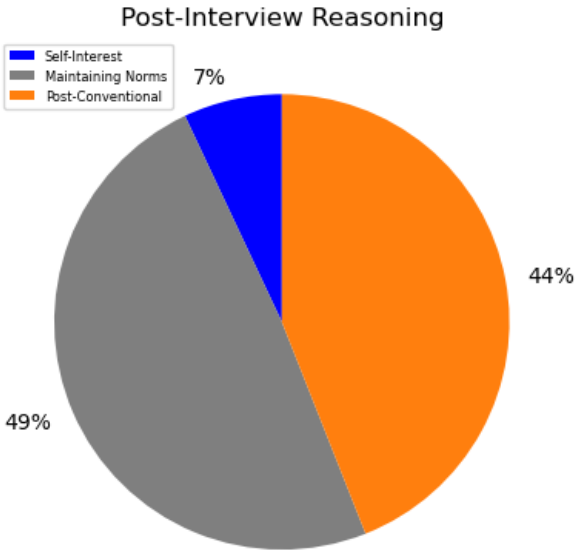

Figure 3.4: *Pie chart for percent breakdown of schema for levels of reasoning for the post-interviews. We combined all participants into one graph since we are looking at scientists as a whole.*

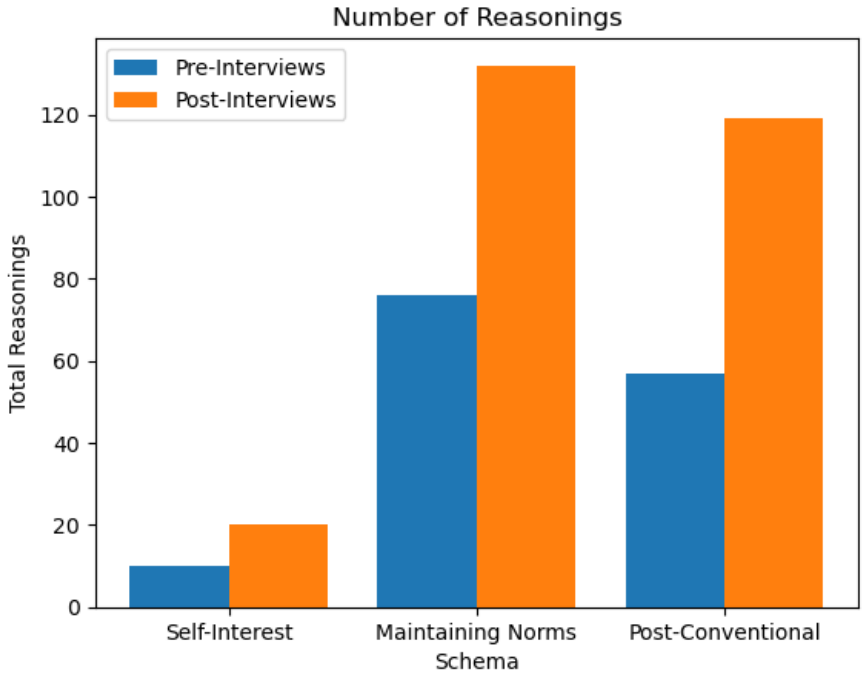

Figure 3.5: *Bar chart for total number of reasonings categorized into the different schemas for all of the scientists.*

3.4.1 Moral Sensitivity

Our first conclusions were the scientists in the fellowship invoked more values overall after the end of the fellowship (654 vs. 487). Since the average time for both the pre and the post-interview were around the same, we can say that time was not a factor in this increase in the frequency of value claims. Thus, the participants were invoking values ~33% more often at the end of the fellowship than at the beginning.

In our group of participants, the kinds of values the scientists invoked when discussing their personal experience with research ethics changed from predominantly epistemic to predominantly ethical. While much of the fellowship focused on various epistemic goals in science, and trade-offs between different epistemic goals, nonetheless the participants seemed more attentive to ethical dimensions of their work after the fellowship.

There was not a statistically significant change in the kinds of values invoked when talking about RCR vignettes. We hypothesize that this is because the goal of the vignettes was to invoke mainly ethical values, so it would make sense for a majority of the values invoked for the vignettes to be more ethical than epistemic.

After our first analysis we discussed how to improve on how we are looking at the scientists' ethical decision making. Our research question asked, "How do discussions of the values embedded in scientific practice change scientists' reasoning about responsible conduct of research?" We found that after having discussions focused on the values in science, scientists' reasoning changes to invoking more ethical values when talking about their own work, but when it comes to ethical scenarios outside of their own work, the values they invoked did not change. We then modified the original research question to discover more about scientists' moral sensitivity and moral reasoning.

To answer the question of "how do the discussions of values embedded in science shift scientists' ethical sensitivity", we claim that the value-focused training improves scientists' ethical sensitivity. We can see from the data that the participants evoked more values in the vignettes since there was a significant increase in the *number* of ethical values invoked after the fellowship.

We also saw that the participants invoked a wider range of values after the fellowship which is seen in the unique values invoked. Since time is not a factor in the change in number of values, the participants are noticing more values in the vignettes after the fellowship sessions. Since we are saying that the morally

salient items in these scenarios are the values, this means that the participants' moral sensitivity increased. This is a step in the right direction since to start improving in ethical decision making people need to improve on each individual component.

3.4.2 Moral Reasoning

The fellowship originally did not appear to improve scientists' moral reasoning according to our measure. From the data, there is a significant change in the total number of schemas invoked, but there is not a change in the distribution of schemas participants used. While there is a significant increase in the *number* of distinct reasons brought up from pre to post fellowship (as seen in Figure 3.5), the overall trend between the pre and the post-interviews shows that the schema of reasoning used remains about the same.

This then raises concerns that the way we measured the scientists' moral reasoning did not fully capture significant changes in moral reasoning. Although the participants' reasoning schema did not change, we believe something about their reasoning must have changed. Reasons for these speculations are because of the *increase* in participants' total reasonings in the post-interviews and the unique values invoked are higher in the post-interviews. This suggests that the participants are able to come up with not only more claims and reasonings for scenarios but they are reasoning with different ideas as well. One suggestion for this is that looking at just the Neo-Kohlberg Schema is not enough to fully analyze someone's moral reasoning as this does not fully encapsulate what is being said in the scientists' moral reasoning.

This raised the question "How do we look more in depth at the participant's reasoning?" Answering this new question will allow us to ultimately answer how the participants' reasoning changed due to the fellowship more accurately. While we believe that the Neo-Kohlberg Schema should not be disregarded, we need more than just the schema to categorize scientists' reasoning.

3.4.3 Improving Moral Reasoning Analysis

In order to fully analyze how these physicists are reasoning in ethical scenarios, we attempted to expand on what we analyzed the scientists' moral reasoning. Originally we only used the theories of Neo-Kohlberg Schema, but we further analyzed the scientists' moral reasoning by also including principles and stakeholders. Rather than saying that there is a correct way to ethically reason like some methods of measuring

Principle	Definition	Example
Autonomy	The ability to make one's own decisions. There are two conditions that are essential for autonomy: liberty (independence from controlling influences) and agency (capacity for intentional action).	Participant gives informed consent
Non-maleficence	Non-maleficence is the obligation to not harm anyone intentionally. This can be both an action or a non-action that is made to prevent any harm.	Avoiding unnecessary harm towards patients
Beneficence	Beneficence is the obligation to contribute to someone's overall welfare. This is different from non-maleficence since beneficence is about taking positive acts to help others instead of refraining from harmful acts.	Improving someone's health through surgery
Justice	Justice is the fair, equitable, and appropriate treatment of other people. This principle states all people must be treated equally and must provide benefits to all people regardless of features like gender or race.	Everyone receives life saving supplies regardless of social class

Table 3.10: *Definitions and examples of autonomy, non-maleficence, beneficence, and justice from the Four Principles.*

ethical reasoning, these theories allow us to view ethical reasoning in a more asset-based way since we are looking at how the physicists are reasoning with schema, principles, and stakeholders in their own way. However, we need to refine our definition of a stakeholder more. We would also like to note that the application of this combination of theories to open-ended responses is novel. Since we have shown that we can apply these theories to categorize the physicists' reasoning, we can attempt to use these theories for the rest of the scientists and eventually look to see any changes in their reasoning.

Principles

Beauchamp and Childress identified four principles to adhere to when it comes to ethical decisions in clinical medicine.[76] These principles are designed for making ethical decisions in biomedical ethics, we suggest that we can use them to categorize someone's context of reasoning in the STEM field. Definitions and examples can be found in Table 3.10:

Stakeholders

For our definition of stakeholder, we will be using Freeman's definition of a stakeholder: a stakeholder is any group or individual who can affect or is affected by the achievement of the organization's objectives.[77] This theory comes from the philosophy of business management. Freeman then identified two groups that are classified as stakeholders: "can affect" and "affected". In this chapter, we will refer to these as "involved stakeholder" and "affected stakeholder" respectively.[78]

To further break down the classifications of stakeholders, Ulrich identified the different roles stakeholders can occupy.[78;79] These roles can be applied to both involved and affected stakeholders:

- **Source of motivation/client:** The stakeholder whose values are being served
- **Source of control/designer:** The stakeholder who has the power to decide
- **Source of expertise/planner:** The stakeholder who contributes to the overall expertise of the project
- **Sources of legitimation:** The stakeholder that contributes the necessary responsibility of those involved

Improving Moral Reasoning Data Analysis

Once we decided on using Rest's schema, principles, and stakeholders to use to further analyze participants' moral reasoning, we selected two of the physicists post-interviews to test if our proposed theories were viable for application to all participants. We put these transcripts into Excel with the interviewee questions in one column and the interviewee responses in the column next to the questions. We did this for every question and response in the vignette section of the transcripts. Once we put all the questions and responses in the Excel document we separated everything by the vignette scenario.

To code the responses, we highlighted where they were reasoning. For their Neo-Kohlberg schema, we looked to see if they were reasoning about themselves (self-interest), reasoning about how things should be based on social, legal, or scientific norms (maintaining norms), or reasoning for the betterment of others (post-conventional). For their principles invoked, we looked to see if their reasoning involved respecting other's choices (autonomy), avoiding intentional harms (non-maleficence), maximizing benefits (beneficence), or wanting what's fair for everyone (justice).

Using Ulrich's groups to help us classify stakeholders, we created a list of specific groups of stakeholders for ourselves which are: **self, students, scientists, administration, and public**.[78] Note that stakeholders could be categorized in more than one group depending on the context of the response. We did this for every vignette and for both physicists. A more detailed breakdown on how we coded the responses is in the "Vignette Responses" section below.

For inter-rater reliability, four of the authors of this project met together and discussed each response and what we should code for each response. We came to an agreement on all of the responses.

sectionVignette Responses

First Vignette

The following is a question-response from one of the physicists for the vignette regarding a grad student bringing up the concern that a scientist did not get informed consent to a project. In this, the italicized words indicate the question from the interviewer; the response is in normal text:

> (*How should [the scientist] respond?*) If he can really help improve the reliability of drugs, then maybe it's worth inflicting some pain on people...

For this response, the physicist's justification for inflicting pain is "If he can really help improve the reliability of drugs". While the action in this quote is about causing pain, the reason for inflicting pain is to increase benefits for other people (reliability of drugs), so this would be classified in the post-conventional schema.

As mentioned in the previous paragraph, the reasoning here is to increase benefits for other people. The principle here would be beneficence since the interviewee wants to increase benefits of the drugs.

In this response, the "he" would be the scientist from the vignette in charge of the project and the "reliability of drugs" affects the populace. The groups for stakeholders in this response would be "scientists" for the involved stakeholder and the "public" for the affected stakeholders since the interviewee is reasoning for the scientist in charge of the project to run the study to improve drugs for the people.

This next quote is from the same physicist from the same vignette:

(*How do you think about causing pain to people?*) Well, I mean, if people are dying from the bad drugs, then I guess it's okay. Maybe to give some people shocks...

For this response, the action might seem unethical since they are saying that the scientist should perhaps shock people, which would cause harm to them. However, their reasoning that "if people are dying from the bad drugs, then I guess it's okay" implies that the shocks are to improve the safety of the drugs, which would ultimately prevent less deaths. This puts this quote into a post-conventional schema.

For the principle of this response, the interviewee is implying that they want to reduce the deaths from the bad drugs, which would be minimizing the harms from the drugs. The avoidance of harm would make this an application of non-maleficence. However, they also are suggesting shocking people would be justified by this consequence, which, since it is intentionally causing harm to individuals not offset by benefit to them, would go against the principle of non-maleficence.

The groups for stakeholders in this response would be "scientists" for the involved stakeholder, since they are still talking about the scientist from the vignette running the experiment, and the "public" for the affected stakeholder, since the scientist is running the project improve drugs for the people.

Second Vignette

The following is a question-response from a different physicist for the vignette regarding a scientist is in charge of hiring a new faculty member and could only bring in three male candidates but not the female candidate:

(And so why what's the obligation [for bringing in the female candidate]?) Where's it come from? It's better for science, you're getting different, you're getting different perspectives, the perspective of her life, everything her life has told her everything about her environment, as a woman could be significantly different from that of a male.... And so you need this diversity in personnel in order to better guarantee diversity in how we solve problems, how we see the data, how we treat the data, etc. I think that's, that's essential.

In this vignette response, the physicist is appealing to post-conventional reasoning since in this response the physicist is appealing to the fact that promoting diversity through hiring practices would be

better for science, as is seen from their reasoning quotes of "It's better for science" and "you need this diversity in personnel in order to better guarantee diversity in how we solve problems".

This is the one of the responses we saw in the vignettes where there were more than one principle in each response and there was a supporting principle in this response. The interviewee responds with "It's better for science" where the "It's" in this response is diversity. This response would be a beneficence response since diversity is improving science (maximizing the benefits of science). Later on in this response, they say that diversity is needed for all of these problems and do not explicitly state that diversity would maximize benefit or reduce any harms from these problems. This might then be a justice statement. In this whole response, the beneficence in the first part with diversity would then go on to support the justice of diversity in solving those problems.

The stakeholders here are more complicated than the other vignette responses. The issue lies where the interviewee states "It's better for science". The "It's" would be the diverse group of scientists, so the involved stakeholder would be "scientists", but the affected stakeholder in this response would be "science". Since "science" is not an individual or group, we would then categorize this as "beneficiaries of science", which does not fit in any of the groups as mentioned above since all of the groups of stakeholders we found would benefit from science.

Neo-Kohlberg Schema: From the responses, we can identify the invoked Neo-Kohlberg schema. The physicists invoked a mixture of post-conventional responses and maintaining norms responses with no self-interest responses. This could possibly mean that the level of reasoning is at a high level already for these physicists, or they could also be trying to make themselves look better in front of the interviewer.

Four Principles: The principles we do see in the above responses are beneficence, justice, and non-maleficence with none of the responses in this project containing the principle of autonomy. This is not to say that the physicists don't invoke autonomy at elsewhere in the interviews. Beauchamp and Childress's principles of biomedical ethics can be applied to our setting; however, there were responses where it was not entirely clear what principle would align with the interviewee's reasoning. This may be due to lack of detail in the interviewee's responses, especially for non-maleficence, which is more clear in a clinical setting where one person has substantial responsibility and the potential to harm others directly.

Stakeholders: Identifying who the physicists are reasoning towards is not as simple as the traditional stakeholder theory would make it. The issue with our current theory of stakeholders is what happens when

the the stakeholder is an abstract topic or object like in the last vignette response where the stakeholder is identified as "science" since our definition of stakeholder is an individual or group. This shows that our current use of stakeholder theory needs to be refined and our groups of stakeholders needs to be adjusted since the physicists reasoned with an abstract idea rather than an individual.

Looking at the effectiveness of the value-focused fellowship on improving the overall ethical decision making of the scientists, we see that there is a start for improving scientists' ethical decision making. We see an improvement in ethical sensitivity and inconclusive results for improvement in moral reasoning. More work will be needed in looking at motivation and implementation in order for us to fully claim that this value-focused training is effective at improving ethical decision making.

With this work, we want to highlight that this kind work requires collaboration between physics and philosophy. This is due to the fact that physics in general has little work when it comes to making ethical decisions in science. Because of this, we have to turn to the field of psychology and philosophy to learn how people make ethical decisions and adapt this to science as they are the experts in this field. We encourage future ethics work to be interdisciplinary with philosophy to have a more solid foundation in ethics work.

Chapter 4

Development of Tasks

4.1 Theory

4.1.1 Three-Dimensional Learning

There is a call for instruction to focus more on promoting transferable knowledge and skills across topics/disciplines instead of just focusing on teaching and assessing specific content knowledge in a course.[3;36;80;81] The transferable knowledge and skills means that instead of just memorization of the content knowledge taught in a course, students should be taught the content knowledge and skills so that they can be applied to other topics.[36] This should now be done by focusing on motivating new approaches to instruction and assessment.[3]

One way to focus promoting transferable knowledge is by developing learning goals and assessment through the lens of Three Dimensional Learning. The National Research Council stated that in order to be proficient in science, science must be viewed as both a body of knowledge and as an evidence-based model with theory building that continuously expands and revises knowledge.[82;83] In order to keep science in line with the sentiments of being a body of knowledge and evidence that revises knowledge, standards must be designed with three dimensions: scientific and engineering practices, crosscutting concepts, and disciplinary core ideas.[82;83]

The first dimension of scientific and engineering practices are the behaviors that scientists and engineers engage in to do science.[82;83] These practices incorporate what scientists do and help students un-

derstand how scientific knowledge develops (e.g: developing and using models to explain a phenomenon, use mathematics to calculate an important variable, etc.).[83] It's important to highlight the practices being applied in learning goals because showing what scientists and engineers do can pique student interest and keep students motivated in science or engineering.[84;85] In Table 4.1, we highlight the scientific and engineering practices used while developing tasks for the TaSPA. Note that while we use most of the scientific and engineering practices to develop tasks, we found it nearly impossible to include the practices of "asking questions" and "planning and carrying out investigations" in a multiple-choice assessment.

The second dimension of crosscutting concepts involves the ideas that link across the other scientific disciplines.[82;83] These concepts give students an organizational framework that connects the knowledge from different disciplines into a coherently and scientific based view of the world.[83] These concepts should be explicit and common enough so that students develop a usable understanding of how science and engineering works.[83] A list of all crosscutting concepts used is found in Table 4.1.

The disciplinary core ideas are the foundational concepts of the specified discipline of the course or assessment.[81;83;86] In order for something to be a disciplinary core idea, they must meet at least two of the four following criteria: have a broad impact over multiple sciences or a key idea in one discipline, provide a key tool for understanding complex problems, relate to the likes and interests of the students, and be teachable and learnable over multiple grades at increasingly levels of depths.[86] While the National Research Council only highlights disciplinary core ideas for physical sciences, we created core ideas for the TaSPA that are more relevant to a thermal or statistical physics class. These disciplinary core ideas for thermal and statistical physics can be found in Table 4.1.

Scientific Practices	**Croscutting Concepts**	**C**
Developing & Using Models	Patterns	E
Analyzing and Interpreting Data	Cause and Effect: Mechanism and Explanation	T
Using Mathematics	Scale, Proportion, and Quantity	E
Constructing Explanations	System and System Models	H
Obtaining, Evaluating, and Communicating Information	Energy and Matter: Flows, Cycles, and Conservation	M
Engaging in Argument from Evidence	Structure and Function	P
	Stability and Change	

Table 4.1: *List of components from three-dimensional learning used to create tasks for TaSPA*

4.1.2 Knowledge-in-Use

To incorporate all of the evidence centered design framework including the components of the evidence centered design triangle, Harris et al. created a Knowledge-in-Use framework that incorporates defining evidence of students' knowledge and skills for a certain domain. The aspects of Knowledge-in-Use include identifying learning performances, knowledge skills and abilities, evidence statements, and task features that will elicit student understanding in a course.[36].

Learning performances refers to a set of proficiencies that are used to determine a performance expectation set in the classroom.[36] These are related to the learning goals of the classroom where both learning performances and learning goals identify the expected knowledge students are intended to learn in the classroom. The learning goals created should be rooted in the three dimensional learning framework.[36] This means that each learning goal should contain a scientific practice, crosscutting concept, and disciplinary core idea as appropriate to the course.

Knowledge, skills, and abilities are the proficiencies that are targeted by the assessment task.[36] Evidence statements are the observable features of student performance that capture the proficiencies highlighted by the learning performance.[36] Task features are how the task is designed that will illicit the proficiencies of the learning performance.[36] When designing tasks through this Knowledge-in-Use framework the goal is to make it easier to identify students' knowledge and skills that can then be used to improve course curriculum.

4.1.3 Self-Regulated Learning

Self regulated learning is defined in terms of self-generated thoughts, feelings, and actions oriented towards attaining one's own goals in a constructive and self-directed process.[87–89] In order for self regulated learning to be successful, someone must be able to direct their own learning which includes the skills of orienting, planning, monitoring, evaluating, and correcting.[88] People who have had success with self-regulated learning have been found to swiftly transfer knowledge and strategies learned to other situations.[88]

One of the more crucial parts of learning from self-regulated learning is through the use of feedback.[87] In a formative assessment such as the TaSPA, feedback is a key element as it defines the information about

the success of the students on the assessment.[90] Nicol and Macfarlane-Dick established a model connecting self-regulated learning and internal feedback for instructors.[90;91] Through this model they established seven principles that support and develop self-regulation in students, but for developing tasks we only used three of the principles: clarify goals, deliver quality feedback, and provide opportunities to close the gap. We go into how we apply these principles to create feedback for instructors in the Generating Feedback Reports Section.

In the next section, we highlight how these aspects are being addressed in the TaSPA by: a) allowing instructors to select the topics that are assessed, b) highlighting the scores of the students, and c) providing possible course modifications depending on the scores of the students.

4.2 Designing Tasks

We now articulate the development of our assessment components by highlighting how each component guides the development of further components. The components include: learning performances, knowledge-in-use tables, assessment tasks, rubrics, and feedback reports. This section will go over the development of a "Heat Engines" task by identifying key aspects of the components.

4.2.1 Deciding Task Topics

In order to develop topics that are relevant for a thermal and statistical physics class, we distributed a survey to university faculty across the country. The results of this survey can be found in Rainey et al's paper discussing the creation of this survey and outcomes.[37] This survey's results gave us an insight to what faculty around the country believed were the ideal topics covered in thermal and statistical physics courses. While we initially focused on creating tasks with an 85% agreement on the survey, we eventually attempted to create tasks from all of the topics highlighted in the survey but had to not include some topics due to time.

4.2.2 Creating Learning Performances

Once we identified the topics for the TaSPA, we created sub-ideas for the topics that we identified that should be a task on the assessment. These sub-ideas are what we believe students should be able to know or do about each topic. Sub-ideas for each topic allows us to easily incorporate the three-dimensional learning components described in Table 4.1 to make learning performances for each topic.

For example, a sub-idea for the Heat Engine task would be:

> The variation of pressure and volume of an engine provides information to determine the efficiency for a given amount of heat supplied over one cycle. .

This sub-idea states that if students are given the pressure and volume of an engine (in many cases this would be a pressure vs. volume diagram) then they would be able to solve for the efficiency of the engine. This is just one of the sub-ideas we could have created for this topic and we acknowledge there could be other sub-ideas for each topic. After discussions we accepted that the sub-idea mentioned above should be the accepted sub-idea for heat engines and we then decided to make a learning performance out of this.

To develop a learning performance out of a sub-idea, we first had to identify which of the aspects of three-dimensional learning paired best with the sub-idea of a heat engine. For the crosscutting concepts we pictured that students would be given a pressure vs. volume diagram and would have to figure out the heat flow from the diagram, so the crosscutting concept would be "Energy and Matter: Flows, Cycles, and Conservation". Due to a heavy focus on determining the efficiency from the heat flow, the core idea for the heat engine task would be "Heat Flow". For the scientific practices, we had multiple options for learning performances such as:

1. Analyze and interpret data about the variation of pressure and volume inside an engine to determine the efficiency for a given amount of heat supplied over one cycle.

2. Construct an argument about the variation of pressure and volume inside an engine to determine the efficiency for a given amount of heat supplied over one cycle.

3. Use mathematics to determine how the variation of pressure and volume inside an engine to determine the efficiency for a given amount of heat supplied over one cycle.

Note that when deciding learning performances we attached different scientific practices to the sub-idea in order to create a learning performance. The benefit of identifying a sub-idea for each topic is it allows us to identify multiple different scientific practices that could potentially be used for each learning performance. This allows us to decide which scientific practice would be ideal for each learning performance and we can adjust the learning performance to account for different scientific practices that are not covered in other topics. The crosscutting concepts and core ideas for each of the topics would remain static when deciding the learning performance, only the scientific practice should be able to be changed when deciding on the knowledge/skills students should have for a particular topic. After discussions, we determined that the most appropriate learning performance for the Heat Engines topic should contain the scientific practice of "Analyze and interpret data" as this allows us to have the task focus on having the students take information from a pressure vs. volume diagram and determine what to do with that information.

4.2.3 Designing Knowledge in Use Table

Once we created the learning performance highlighting the knowledge and skills we wanted to assess, we created the KSAs, ESs, and task features to further highlight what we are looking for in students' answers. First is determining the KSAs for the task as this is what we believe to be the proficiencies required to complete the learning performance for the task. The number of required KSAs ranges between 2-3 depending on the complexity of the task (e.g: Students using mathematics to calculate a value could include only two KSAs if the task requires the student to make a calculation and then conclusion based on the calculation, but analyzing and interpreting data could include three KSAs depending on the data presented and claims required from students).

These KSAs guided the creation of the evidence statements since the evidence statements are what we need to see in student answers to claim students meet the KSAs. For example, in Table 4.2 the second evidence statement starts with the phrase "calculation of" the efficiency required for the task. This shows that we want the students to fully show their calculation for the efficiency based on the data given in the task.

The final aspect of the Knowledge-in-Use Table is the task features. These are what should be in

Table 4.2: *KiU table for the Heat Engine task where we explicitly show the Learning Performance, KSAs, ES, and Task Features*

Aspects of Knowledge in Use	Description
Learning Performance	Analyze and interpret data about the variation of pressure and volume inside an engine to determine the efficiency for a given amount of heat supplied over one cycle
Focal Knowledge, Skills, and Abilities (KSAs)	KSA1: Identify the relationship between efficiency (e) and heat (Q), i.e., $e = \frac{Q_H - Q_C}{Q_H}$ KSA2: Calculate the efficiency from the given data KSA3: Make a conclusion based on the calculated efficiency vs. efficiency given
Evidence Statements (ES)	ES1: Identification of the relationship between efficiency and heat ES2: Calculation of the efficiency of the engine ES3: Identification if the calculated efficiency is reasonable given the context of the task
Task Features	Question gives a scientific question, claim, or hypothesis to be investigated. Question gives a representation of the data (e.g., table or graph, or list of observations) provided to answer the question or test the claim or hypothesis. Question gives an analysis of the data or asks student to analyze the data. Question asks student to interpret the results or assess the validity of the conclusions in the context of the scientific question, claim, or hypothesis.

the task to elicit the evidence listed in evidence statements. These task features are dependent on which scientific practice is selected for each task (e.g: a task with the scientific practice "using mathematics" will have the same task features as another task with the "using mathematics" task features). Once we identify the KSAs, ESs, and task features, we get the KiU table shown in Table 4.2.

Note that while most of the evidence statements are in terms of general tasks (e.g: there is no context revolving around them), the third evidence statement mentions making a conclusion about the context of the task. While the actual task for the Heat Engines task was not created before creating the KiU table, there was a general idea on what the task should be when designing the KiU table. This allows a smooth transition from creating the KiU table and creating the task as the task will better be in line with what we were looking for in the KiU table. These KSAs and evidence statements are not a permanent fixture before creating the task and can (and should) be adjusted depending on the feedback received from students and

instructors. The process of KiU tables were created before I joined the project, but I created this specific KiU table. The list of all of the KiU tables I created can be found in Appendix A.

4.2.4 Creating Assessment Task

Using the KiU table, we created the task that focused on specified topic outlined in Rainey et al's survey.[37] For the Heat Engines task we wanted students to analyze and interpret data so we created a pressure vs. volume cycle of an ideal engine. For all tasks on the TaSPA, we wanted to make sure that the tasks were as close to real world applicable as we could possibly make. Making tasks "relevant to real world" is to make the subject matter more relatable for the students taking the assessment and allows students to fully apply their knowledge and skills to each task.[92;93]

In the Heat Engines task, we decided to give the context of the task to be determining the efficiency of an Otto Cycle engine, as this engine is typically found in everyday cars.[94] While keeping the need for making the task as "real world" as possible and incorporating the aspects outlined in the KiU table, we created the Heat Engines task found below:

> You are tuning the Otto engine (engine typically found in modern cars) of your car and you use a dynamometer to measure the efficiency of the engine. Based on the dynamometer readings you calculate that the efficiency of your car's engine is 60%. You show this to your colleague and they think that your dynamometer is off, so you record the pressure/volume readings of your engine for one cycle and develop a PV graph for your engine (shown below). After finding the efficiency of this graph, is it possible that your dynamometer is working right? Explain your reasoning. (This task is accompanied by Figure 4.1 along with the pressures and volumes at points 1, 2, 3, and 4)

In this task, we give the students a pressure vs. volume diagram of an Otto Cycle engine as well as readings from a dynamometer which gives the students the real efficiency. We then ask the students to calculate the efficiency from the given diagram and then compare this efficiency to the dynamometer efficiency (ideal efficiency vs. real efficiency). From this task, we believe that the way it is phrased we will be able to see student responses that contain the proper identity relation, the calculated efficiency, and

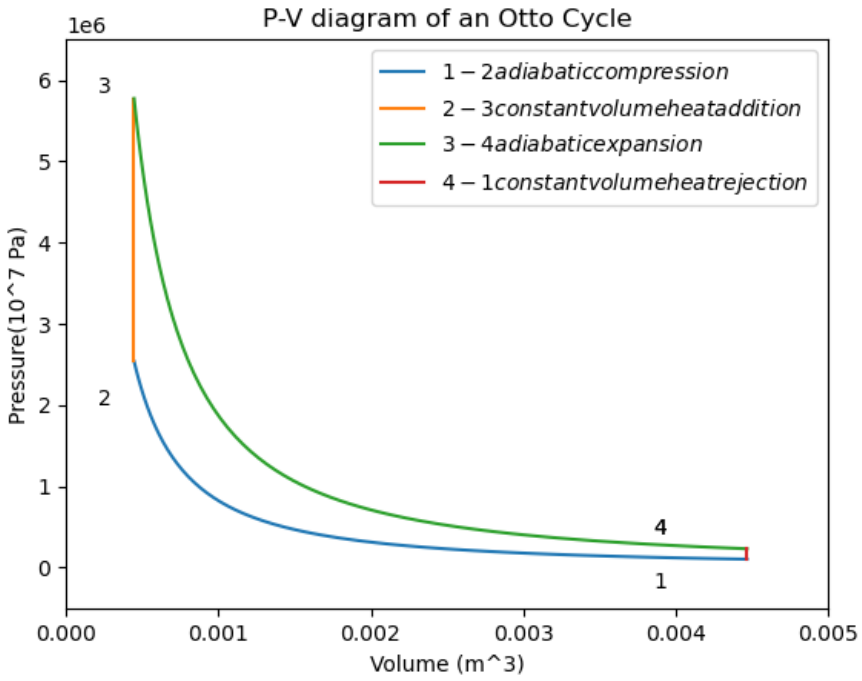

Figure 4.1: *Image to go along with the Heat engine task where students are given the processes of an Otto engine as well as the data from the plot.*

a conclusion based on the scenario. In order to verify if this is true, we gave students the free response version of this task as well as other tasks.

4.2.5 Creating Coupled Multiple Response

Once we created a task, we gave the newly created task in its free response format as well as a few other newly created tasks in an online survey to students who were currently taking a thermal and statistical physics course. The students received the survey towards the end of the semester. Specifically for the heat engines free response task, we received data from 95 students total from five different universities across the United States.

Through the student responses we created questions for the tasks in a coupled multiple-response format (CMR). CMR format is when the questions are multiple choice, but the students are allowed to select their reasoning to go along with their selected answers.[95] In CMR format, students can potentially receive partial credit depending on the answers and reasoning they selected in each task. The benefits for distributing an assessment in CMR format is it allows for automated grading.

We utilized the student responses to the free response tasks to create the different responses students would see in the CMR format. We looked at all the students' answers and selected the correct option(s) the students wrote down to include as options in the CMR version of the questions. We also included incorrect options given from the students to include as many possible ways to answer a question as possible. For example, if an evidence statement for a task includes "Calculation of a quantity", we would take the student responses where they made the calculation and include it in the CMR version of that evidence statement.

Each task took the format as follows: students would give their final answer on the first question of the task and then the students would answer a series of questions to express their reasoning for the correct answer. After the student answers a question for a given task, the student cannot go back to the previous question to change their answers. This makes it so the students have to work through the answer first and cannot look at future questions regarding reasoning for help.

When students look at each task question online the questions will be on separate pages and the student can only move on to the next question after they answer the current question. Students taking the assessment cannot go back to a previous question to change their answer. This ensures that the students completely work through the task first before making their answer and reasoning to ensure that the instructor gets the best possible feedback from the assessment.

Note that for the entire design of the task questions the order of the evidence statements appearing does not matter if they are in numerical order. The only order that matters is that the student conclusion comes as the first question and then questions regarding reasoning come after as mentioned above.

Looking at the Heat Engines task, we start by getting them to answer if the dynamometer in question is working. Note that for all of the questions on this document the correct answer or answers will be italicised:

1. Is it possible that the dynamometer is working correctly?

    A. Yes

    B. *No*

    C. Cannot determine from the information

    D. I don't know

For this question we give them four options: the dynamometer is working, it is not working, it cannot be determined, and they don't know. The option "I don't know" is included in all sixteen of the tasks on the first question as it gives students an option to skip the question entirely if they did not know how to do it in the first place. This is intended to discourage guessing. For example if the student believes that they have not covered the topic in their course or they cannot remember how to complete the topic, the student would select the "I don't know" option and they would move on to the next task that the instructor selected to give the students.

For evidence statement three for the Heat Engines task, we want to see if the students are able to identify if the calculated efficiency is reasonable compared to the real efficiency given from the task. As for this question, we address the students' conclusions about if the calculated efficiency is reasonable, and we want more evidence from students. We address this in another question. For the next question, we want to see what the students' calculated efficiency actually was from the pressure vs. volume diagram given:

2. What was your calculated efficiency from the graph?

    A. 40%

    B. *57%*

    C. 79%

    D. 82%

    E. 98%

    F. 100%

In this question, we give the students a wide variety of potential efficiencies to choose from. While the correct answer is the calculated efficiency should be 57%, we also included efficiencies that other students calculated and answered on the surveys. We only put responses that we saw at least two students put on their surveys as to make the question give realistic answers, but also not put too many responses from the students to select from.

For the second evidence statement of the Heat Engines task, we asked the students to actually calculate the efficiency of the pressure vs. volume diagram. In this question, we ask for their final answer for the

calculation. This is enough evidence to show that they are able to take the data from the diagram and correctly calculate the efficiency from the data. For the next question, we ask the students how they identified the relationship to calculate the efficiency:

3. What quantities from the graph did you use to calculate efficiency?

    A. *Heat flow along the isochoric heat addition (Path 4-1 on graph)*

    B. *Heat flow along the isochoric heat rejection (Path 2-3 on graph)*

    C. Work along the adiabatic compression (Path 1-2 on graph)

    D. Work along the adiabatic expansion (Path 3-4 on graph)

    E. Pressures from the processes

    F. Volumes from the processes

In this question, we highlight that students can select multiple responses as answers to their questions. Throughout most of the questions in TaSPA students have the option to select multiple responses if they like or they can select none at all depending on if any of the responses elicit their reasoning. This question also shows that while the heat flow is what the students need to consider for this question, they could also select "Volumes" as another answer since the students could use the volumes from the diagram to answer this task. We saw from the student responses that they calculated the efficiency using the volumes from the diagram, but students will only be considered correct if they first made the connection between efficiency and heat.

For the first evidence statement we wanted students to demonstrate that they understood the connection between heat and efficiency. By asking which quantities they used to calculate the efficiency, we get the students reasoning on how they calculated the efficiency without explicitly giving them the actual equation highlighted in KSA1 that is the required equation to calculate efficiency. We now have questions that get students to show their answer and reasoning for all of the evidence statements, but we mentioned before that the first question was not enough to consider if the students could fully make the connection between the ideal efficiency and the real efficiency, so we created one more question:

4. What is/are the reason(s) you used to determine if the dynamometer is working?

A. The calculated efficiency from the graph should be about the same as the dynamometer efficiency

B. *The calculated efficiency from the graph should be significantly higher than the dynamometer efficiency*

C. The calculated efficiency from the graph should be significantly lower than the calculated efficiency

D. The calculated efficiency cannot be used to determine if the dynamometer is working

E. Other

With this final question in the task, we believe that the questions in this task will illicit student evidence that answers the proficiencies highlighted by the KSAs and in turn elicits the knowledge and skills highlighted in the learning performance.

In a summary of the order for the Heat Engines task, we asked the students for their conclusion first (ES3), then their calculated efficiency for when they made their conclusion (ES2), then the connection between heat and efficiency (this is asked after the calculation for the efficiency as to not influence how the students calculate the efficiency) (ES1), and finally we asked the students if the real efficiency given is correct based on the calculated efficiency (ES3). We believe that this the the optimal order to ask the questions to ensure we capture the students' conclusion and reasoning regarding the task without guiding them through how to calculate the efficiency in the first place.

4.2.6 Creating Rubric

Throughout the design of the TaSPA tasks, we created rubrics for each of the evidence statements in the task itself instead of creating a rubric for the task as a whole. This allows us to identify what the students are getting correct in the tasks and where the students are struggling. This focus on grading different evidence statements rather than the task as a whole allows for clear feedback to instructors on where they need to improve their instruction in the classroom.

Before we identify how we created rubrics for the TaSPA, we need to introduce how we are grading the students' answers. We characterized the student answers as proficiencies: Met (M), Partially-Met (P),

or Not Met (N). The "Met" criterion indicates that we have evidence that the students met the KSA. In the "Partially-Met" criterion, we have some evidence the students Met the KSA. In the "Not Met" criterion, we do not have evidence that the students met the KSA.

Note that the student scores are not based off a point system (e.g: student score could be nine out of ten points), but instead the scores are based on how the student is correct in their answer/reasoning and did they provide enough evidence as highlighted in the KiU table. Grading of proficiencies instead of scores allows us to delve into the students' reasoning more and give more useful feedback for instructors.

While creating the CMR format of the task, we had in mind what the rubrics should look like as highlighted by the selection of the correct answers and reasoning when creating the CMR versions of the tasks. This allows us to quickly decide what should be considered "Met" when creating the rubric for each of the evidence statement.

To create a rubric for an evidence statement it is important to identify in the CMR versions what options students need to select, what options students need to not select, and which options it does not matter if the students select or not. For this example, we will use question 3 (evidence statement 1) from the Heat Engines task. In this question we need the students to select that they used both of the heat processes to calculate out the efficiency of the engine in order to fulfil the requirement for evidence statement 1. We highlighted in the CMR creation that the students could select the "Volumes from the processes" options if they used the volumes to calculate the efficiency of the engine, however we highlighted in the KiU table that the students must make the connection between efficiency and heat in order to have a correct and complete response. We decided that if they select or don't select the "Volumes" option but they still selected heats options then they would be considered "Met" for that evidence statement. The rubric for this evidence statement is shown in Table 4.3

In the rubric for Evidence Statement 1, we have the conditions for what options should be selected or not selected for met and what options should be selected or not selected for partially met for the student answers for this question. In the first column, we have the different proficiencies. The second row shows the logic for what the students need to do to be considered Met for evidence statement 1. In the columns after, we have the different options that the students can select in the question (e.g: 3a in the rubric corresponds to option a in question 3 in the task).

In this chart we have a series of "+", "-", and blank spaces regarding the options for what students

Table 4.3: *Rubric for evidence statement 1 for the Heat Engines Task, including the logic for students to be Met and Partially Met for this evidence statement. In this table the left column is the proficiency (Met, Partially-Met, Not Met). Every column after that is the options for each questions. For example, the second column is the possible options for students to select for question three option a) in order to get the proficiency Met or Partially-Met. Since for the Met proficiency the selections for options 3a and 3b are "+", that means that students have to select options a) and b) for question 3. Options c), d), and e) are "-" since this means the students do not select these options to be in the Met proficiency. For option f) the selection is blank which means that it does not matter if the student selects this option or not.*

Proficiency	**3a**	**3b**	**3c**	**3d**	**3e**	**3f**
M	+	+	-	-	-	
P	+	-	-	-	-	
P	-	+	-	-	-	

can select for each option. Throughout every rubric, these symbols represent every possible selection for each of the Met and Partially-Met proficiencies. The symbols are represented as the following: the "+" means that the student has to select that option, "-" means that the student must not select this option, and the blank space means that the student can select that option or not and it will not affect their proficiency rating.

Going through the different proficiencies, the Met proficiency has this logic: Students must select option a) *and* option b) (these are the two heat processes), not select options c), d), and e), and it does not matter if they selected option f) (the volumes option). If students pick these exact options, then they have demonstrated that they are able to correctly make the connection between efficiency and heat and are considered Met for this evidence statement. The Partially Met proficiency has the following logic: Students must select option a) *or* option b) (the two heat processes), not select options c), d), and e), and once again it does not matter if they select option f) (the volumes option). These options show that the student almost fully made the connection between heat and efficiency, but were not able to make the full connection so this means they only have a Partially Met as their proficiency for evidence statement 1. If the students select any other option combination for this specific question then they would be considered with a Not Met proficiency since they did not consider the connection between heat and efficiency.

Every rubric in the TaSPA follow the same logic of marking down which options the students should select to be in the Met or Partially Met proficiency. If the evidence statement spans multiple questions (such as evidence statement 3), there is one rubric for both of these questions.

4.2.7 Generating Feedback Reports

The final step of developing tasks for the TaSPA was creating the feedback the instructors would receive regarding the student performances on the tasks. My role in the feedback generation was making the connection between Nicol and Macfaralane's principles (clarify goals, deliver quality feedback, and provide opportunities to close the gap) to the aspects of the task development.

For clarifying the goals for instructors, we opted to show the instructors the KSAs for the task as these are what we defined as the goals the students need to know to fulfil the knowledge and skills outlined in the learning performance. In the feedback report, the instructors would see:

Students were asked to:

- Identify the relationship between efficiency (e) and heat (Q), i.e., $e = \frac{Q_{\mathrm{H}} - Q_{\mathrm{C}}}{Q_{\mathrm{H}}}$
- Calculate the efficiency from the given data
- Make a conclusion based on the calculated efficiency vs. efficiency given

For the statements regarding the student performance, we base these statements on how well all of the students performed on the evidence statements in the tasks. In the feedback report given to instructors, we would start the section with "The TaSPA:" followed by statements of what the evidence shows for how well the students performed. For example, if a majority of the students had a Not Met proficiency for all of the evidence statements in the Heat Engines task, the section the instructors would receive would look like:

The TaSPA:

- Did not provide evidence that your students correctly identified the relationship between efficiency and heat.
- Did not provide evidence that your students correctly calculated the efficiency from the given data.
- Did not provide evidence that your students made a correct conclusion based on the calculated efficiency vs. efficiency given.

For this section, we would also include a breakdown of the class percentages for each proficiency for each evidence statement that were based on the rubrics. Note that we presented this as "The TaSPA

provided evidence" as we want to present the evidence to instructors so that they can make the decisions to improve their class based on the evidence the TaSPA provided.

I also introduced the clarification of Partially Mets to the TaSPA task development. Depending on if any of the evidence statement rubrics contained Partially Mets, the next section of the developed feedback would be an explanation of what percentage of the class were marked as Partially Met and what that partially met means. This would be a drop down menu for instructors to provide instructors with a choice of viewing what the Partially Met case means for each task. As this only applies to evidence statement 1 for the Heat Engines task, this section would look something like "x% of students only considered one of the heat processes of the engine instead of both" where in this case the "x" would be the actual percentage of students who received Partially Met proficiency for this evidence statement.

In the final section of the feedback, we wanted to provide opportunities to close the gap between student performance and the goals set by the task. For this we gave instructors open-ended ways students could benefit from more opportunities to consider depending on the proficiencies of the students given from the rubrics. If the class was considered Met for all of the evidence statements, then this section would be blank as there are no opportunities to improve the students' knowledge and abilities. If the students are at a Partially Met proficiency or Not Met proficiency, then there would be some suggestions for improvement in the classroom. For example, if the students were considered Not Met as mentioned above, then the instructor would receive:

Students could benefit from more opportunities to:

- Explore how heat of the non-adiabatic processes of an engine contribute to the efficiency calculation.
- Calculate more efficiencies of an engine given a pressure vs. volume diagram.
- Make conclusions regarding the comparison of ideal efficiencies to real efficiencies.

Note that we intentionally give vague feedback for instructors. This is to allow the instructors to decide how they are going to make adjustments incorporate these recommendations in their classroom based on how they want to teach the class (e.g: Instructors could improve on lectures by adding on clicker questions based on these recommendation or adjust the homework to make students articulate their reasoning to go with their answers). To summarize the feedback professors receive, we want instructors to see specifically

what the students are struggling on in their conclusions and reasoning through clear evidence provided in the TaSPA and what are some of the recommendations that instructors could make to improve their classrooms.

Chapter 5

Validation of Tasks

5.1 Theory

5.1.1 Knowledge-in-Use

Evidence Centered Design is a framework that states that assessments are representatives of performances to make inferences about a wider set of a students' skills or knowledge.[3] This process of collecting evidence of student knowledge and skills is fundamental to all assessments as the set of knowledge collected affects all other aspects of the classroom highlighted in Figure 2.2.

The aspects of collecting evidence from an assessment are observation, interpretation, and cognition which are highlighted in Figure 5.1.[3;6] The cognition corner of the triangle refers to the theory, models, and set of assumptions that can be made about how students represent their knowledge on a subject matter domain.[6] This is not only what students should learn about the content domain in the classroom, but also *how* students should learn about the content domain.[3] The observation corner of the triangle refers to the specifications of the assessment task that shows a students' understanding of the course.[3;6] This should be done by identifying how students learn in the domain, what is important for students to learn, and what students would be expected to know based on their instructional history.[6] The interpretation corner of the triangle refers to making observations of students' knowledge and skills.[3;6] This includes any information that will improve instruction inside the classroom.

Harris et. al. created a Knowledge-in-Use framework that incorporates defining evidence of students'

knowledge and skills for a certain domain. The aspects of Knowledge-in-Use include identifying learning performances, knowledge skills and abilities, evidence statements, and task features that will elicit student understanding in a course.[36]

Learning performances refers to a set of proficiencies that are used to determine a performance expectation set in the classroom.[36] These are related to the learning goals of the classroom where both learning performances and learning goals identify the expected knowledge students are intended to learn in the classroom. Knowledge, skills, and abilities are the proficiencies that are targeted by the assessment task.[36] Evidence statements are the observable features of student performance that capture the proficiencies highlighted by the learning performance.[36] Task features are how the task is designed that will illicit the proficiencies of the learning performance.[36] When designing tasks through this Knowledge-in-Use framework it is easier to identify students' knowledge and skills that can then be used to improve course curriculum.

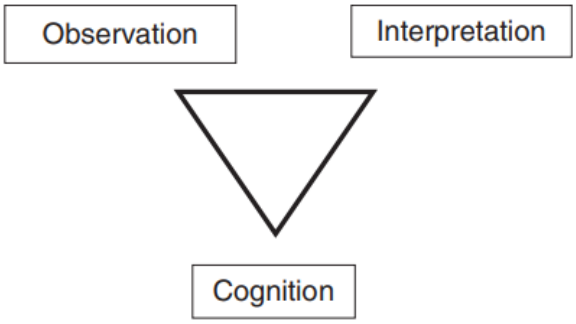

Figure 5.1: *Triangle identifying elements involved in conceptualizing assessment as a process of reasoning from evidence.*[3;4]

5.1.2 Validation focusing on instruction

Pellegrino et al. highlighted different components of validation that incorporate the original components of validation (face validation, content validation, construct validation, and criterion validation). These components should focus more on improving instruction in the classroom and further emphasize the Education System Triangle in Figure 2.2.[6] These components of validation are cognitive validation, instructional validation, and inferential validation.

Cognitive validation is the extent to which the assessment assesses student knowledge and skills in the classroom.[6] This relates to the cognition element from Figure 5.1 as cognitive validation should focus on what students should know in a course but also how students should learn.

Instructional validation is the extent to which the assessment is aligned with the instruction in the class and the learning goals of the curriculum.[6] This includes assessment providing useful and timely instructional related information designed to improve the classroom. The instructional information given should be aligned with the knowledge and skills derived from a class' curriculum.

Inferential validation is the extent to which the assessment provides accurate and reliable information about the students' performance in the classroom.[6] This should be done through some statistical model or analysis (e.g: Item Response Theory) that determines whether a task reliably aligns with the intended learning goal. These components of validation also incorporate the components from Figure 5.1 as the inferential validation and interpretation corner both involve taking assessment scores and interpreting student knowledge and skills from these scores.

Cognitive validation, instructional validation, and inferential validation are critical for evaluating the validity of assessments that focus on supporting instruction.[6] Despite these three components focusing more on instruction and alignment with the Education System components, these three validation components still contain the original four components of validation (face validity, content validity, construct validity, and criterion validity). For example, cognitive validation contains aspects of content validation as both are components that gather evidence towards the assessment measuring the knowledge for a topic. Cognitive validation expands on this and ensures gathering evidence towards knowledge and skills for a topic.

Table 5.1: *Three components of validation designed to support learning and instruction.*[6]

Validation Method	**Description**
Cognitive Validation	Assessment measures knowledge *and skills* of students
Instructional Validation	Assessment provides timely and useful instructional information
Inferential Validation	Assessment scores are reliable and provides accurate information about student performance

5.1.3 Validating the TaSPA

Previous assessment design and validation in PER has focused on students' conceptual understanding of a particular physics topic.[14;31;96] Despite these assessments like the Force Concept Inventory being an accepted way to assess a student's understanding in the course, there are concerns about how to interpret assessment scores to improve instruction for the classroom.[14]

We are designing and validating the TaSPA to make the assessment results easy to interpret to improve instruction and curriculum in the classroom. We do this by designing task through the Knowledge-in-Use framework where we are collecting evidence of student knowledge and skills and interpreting this evidence as useful feedback for instructors.

We also need to validate the TaSPA so that the feedback reports are accurately showing student performance in the course *and* the TaSPA provides useful and timely feedback to make instruction and course learning goals better. For validating the TaSPA, we aligned our methods with the three components of validation (cognitive validation, instructional validation, and inferential validation). For cognitive validation, we conducted student interviews and expert interviews to establish if the TaSPA tasks are eliciting the expected knowledge and skills of a thermal and statistical physics class while also ensuring that the TaSPA contains relevant tasks to the course's curriculum. For instructional validation, we interviewed faculty regarding their thoughts on how useful the feedback will be for improving their classroom. For this chapter we will be going over exclusively cognitive validation and instructional validation of TaSPA while inferential validation will be in future publications as data collection is ongoing. Once we complete all three components of validation for the TaSPA, we will have a validity argument for the TaSPA being an assessment designed to improve instruction in the classroom.[6]

We applied these theories about designing and validating assessments for improving instruction for the classroom to answer the following research questions:

1. How can we use student interviews to establish evidence for cognitive validation?

2. How can we use student interviews to establish evidence for instructional validation?

5.2 Methods

5.2.1 Cognitive Validation - Student Interviews

We collected evidence from student interviews where we conducted think aloud interviews with one moderator and one student in a room.[97] These students were from one university who had just taken a thermal and statistical physics course the semester before the interviews.

The students were asked to work out the tasks on their own but also encouraged to say what they are thinking out loud. Once the students were done working on a task they would then move on to the next task. Each interview lasted about an hour. The students were encouraged to talk about what they were thinking about to make sure that we could more fully capture their reasoning in their answers.

We collected interview data for ten students who answered the free-response versions of four of the tasks and two students who answered seven CMR tasks. While we only want to gather evidence for cognitive validation through student interviews for the CMR versions as the finished TaSPA will be CMR only, we were only able to analyze one student interview for the CMR versions of the tasks and ten student interviews for the free-response version of the tasks before the project ended. This allowed us to focus more on developing the methodology for free-respones interviews, but in the next section we will explain how the two types of interviews are similar.

5.2.2 Cognitive Validation - Analysis

We then looked at students' answers to tasks on the TaSPA to see if they aligned with the predetermined evidence statements since the evidence statements are what we are looking for in student answers to see if they have the necessary skills for solving thermal and statistical physics problems.

We analyzed student responses along two categories: "Demonstrated Evidence Statement?" (Student answer lined up with evidence statement) and "Correct?" (Student answer/evidence is correct in both the written/verbal response and the CMR response). We used these two categories for breaking down each individual evidence statement in a task (e.g: if a task had three evidence statements of "showing relationships", "solving for a variable", and "making a conclusion", we would check the students' response for *demonstrated evidence statement* and *is it correct* for "showing relationships", *demonstrated*

Code	Meaning
W	In written anwser
V	In verbal answer
M	In CMR answer
C	Correct answer
X	Not in answer
I	Answer is incorrect
G	Student guessed answer

Table 5.2: *Table showing the individual codes for validation*

evidence statement and *is it correct* for ”solving for a variable”, and *demonstrated evidence statement* and *is it correct* for ”making a conclusion” individually). We answered these categories on a simple ”yes” or ”no” depending if they demonstrated the skill outlined by the evidence statement or got the correct evidence/answer.

To put the students in these categories, we marked the students' responses based on the following coding: Student wrote down their answer on paper (W), student said the answer out loud (V), students put answer in the CMR format (M), student's answer is correct (C), student did not use skill in their answer (X), answer is there but incorrect (I), student guessed their answer (G). The student would be marked as “yes” for both “skill used” and “correct” if they were marked W, V, or M and C for that evidence statement since this means we see that they used the correct skill and it is assumed that if they were not marked with an I then they got the evidence/answer correct. The student would be marked as “yes” for “skill used” and “no” for “correct” if they were marked W and or V and M, but also I. The student would be marked as “yes” for “correct” and “no” for “skill used” if they were marked as M, C and G or I and G for their answer. The student would be marked as “no” for both categories if they were marked X. A chart for these is found in Table 5.2. Note that for a student to be fully “correct” for CMR validation they *have* to mark the correct answer on the CMR version regardless of if they wrote out or said the correct answer.

Once we marked all of the students' responses, we then added up their marked responses together per evidence statement to validate each task. The goal is for students to be under the “yes” category for both or the “no” category for both as this shows that the task is able to portray that if they use the correct skill they will get the correct answer or reason or if they do not know how to use the correct skill they will not get the correct answer or reason which accurately shows student performance. Students in the “yes” category for “skill used” and “no” for “correct” means that they can use the skill, but they might have

made a mistake getting the correct answer or reason. Students falling in these categories are fine, but too many fall under these categories then we need to consider there to be issues with the task itself. We also want very few students marked as “yes” for “correct” but “no” for “skill used” as this means that they can guess the correct answer or reason without using any of the needed skills to answer the question.

This analysis can be used for both the free-response interviews and CMR interviews. In both cases the students have to work out and explain their thoughts for the in order come up with their answer. This is due to the free-response version only containing the task and the CMR version the students can only select their answer at first before they can move on to selecting the answer for the reasoning. The only difference between the two is for the CMR interviews we want to make sure the answer they write down or say out loud matches what they select on the CMR version. Since the free-response interviews and CMR interviews have the same methodology, we can analyze the free-response interviews more in depth since we have more data for the free-response interviews and then connect this analysis to the analysis for the CMR analysis.

5.2.3 Instructional Validation - Expert Interviews

Collecting evidence from expert interviews is where we asked experts in the field to read over our tasks and learning performances and see if the tasks are requiring the students to demonstrate the right knowledge/skills and, if they taught the class, do the learning performances align with the learning goals from their class. In our case, we would define ”experts in the field” as people who have taught a thermal/statistical physics class or if they conduct research in a thermal/statistical physics related field.

We conducted 1-on-1 interviews with 2 experts (1 who teaches a statistical mechanics class and whose research relies heavily on statistical mechanics, and 1 who teaches a statistical mechanics class) to read all sixteen tasks which included the learning performance and the task with multiple choice to see if the tasks are requiring the students to demonstrate the right knowledge/skills and if they taught the class do the learning performances align with the learning goals from their class. The instructors were sent a document with all 16 tasks and instructions on how to read the document. The experts were asked to read all sixteen tasks and learning performances before the meeting so that we could make sure that we received comments on their feedback for the tasks in the one hour meeting. In the task document we asked them the following

questions:

- Does the task elicit the proficiencies outlined in the learning performance?
- Do the multiple response questions elicit the proficiencies outlined in the learning performance?
- (If you have taught upper division thermal/statistical physics) Do the learning performances align with the learning performances/goals in your class?
- Are there any general changes we need to make to the problem (e.g: gave students irrelevant equation for a task)

These interviews were recorded and they were allowed to write their own comments and thoughts on the tasks themselves. During the interview the interviewer took notes on the experts' comments to the tasks.

Once the interviews were done, we altered the tasks depending on the comments from the experts. These alterations could include changing the learning performance to be more relevant to the course, adjusting the wording in the questions to make it easier to figure out what to do for the problem, and including equations for students that the experts believed were necessary for the problem.

5.2.4 Limitations

We have a relatively small number of students from a single university in this study. This allows us to look at a small section of the student population and we therefore hesitate to make broad claims. We also have identified these methods for specifically an upper division thermal and statistical physics class, other classes might need adjustments to the methodology.

5.3 Free Response Analysis

For an example of validating the free-response tasks we analyzed a student's written and spoken responses to the Gasses in Atmosphere task. The claim we want to make from this task is that the student is able

to take the given speed distribution graphs and determine the composition of gasses from the speed distribution graphs. From this claim, we decided the knowledge, skills, and abilities (KSAs) we want students to have when working on this task and also the evidence statements (ES) required. These aspects are highlighted in Table 5.3. From these aspects of the KiU table we are able to create the following task:

> You come across an article in a news bulletin which discusses the speeds of hydrogen (H_2) and nitrogen (N_2) molecules. The article also provides two plots (Figures I and II) describing the relationship between v - the speed of the gas molecules, and D(v) - the probability density of the gas molecules, at 1000K (the temperature of the earth's upper atmosphere). However, you observe that the plots do not specify which gases they correspond to. You know that the atmosphere of earth consists of 78% of N_2 and less than 1% of H_2 gases. Given the fact that the earth's escape velocity is 11.2 km/s, discuss whether or not the two graphs can be used to explain the composition of N2 and H2 in earth's atmosphere. Explain your reasoning using appropriate principles.

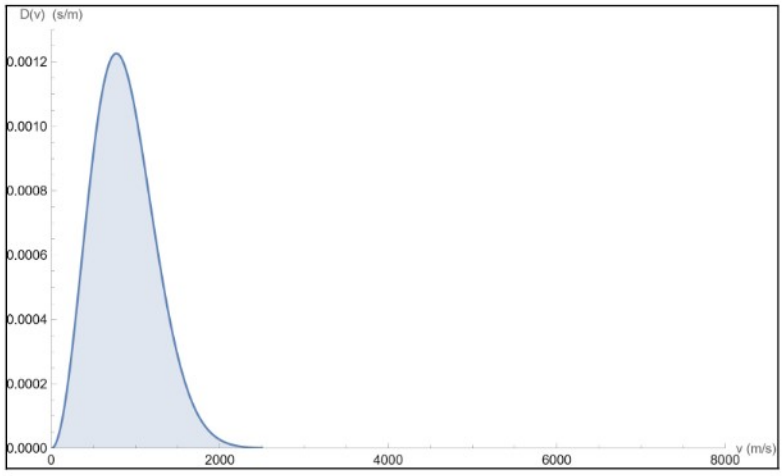

Figure 5.2: *First speed distribution graph given in the problem.*

The following is the student's verbal response to the Gasses in Atmosphere task:

> [Looking at speed distribution graphs] I'm not sure what either of these tells me about composition since this is probability density per particle... (Writes on paper)... I mean at this point, since we're not asking for a numeric answer, I'm going to go ahead and say Figure 2 is more

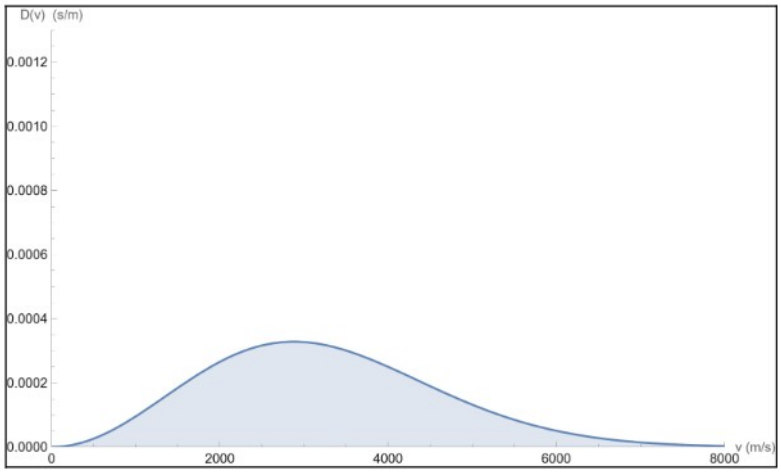

Figure 5.3: *Second speed distribution graph given in the problem.*

than likely hydrogen because of its higher probability of reaching higher velocities, so it's going to escape more frequently. [Writes on paper]... Yeah I think that answers that question

After the student responds with "Yeah I think that answers that question" they start to look at the next task. What the student wrote down can be found in Figure 5.4.

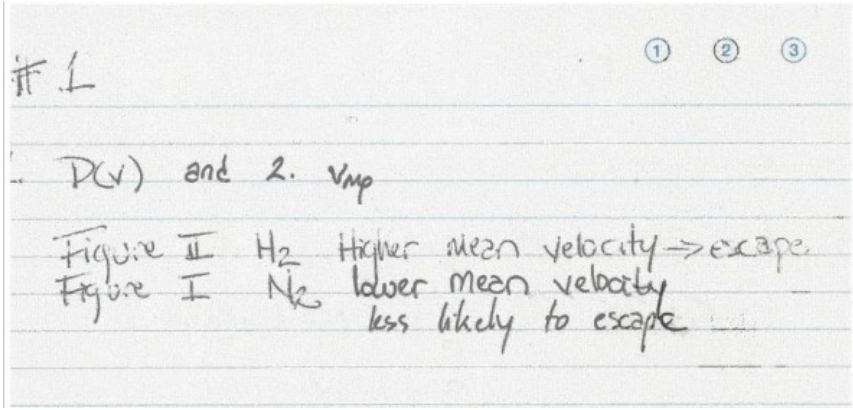

Figure 5.4: *Student's written conclusions to the Gasses in an Atmosphere task.*

5.3.1 Evidence Statement 1

We need to see if this task is valid by identifying what parts of the answers line up with the three evidence statements identified in the KiU table. The first evidence statement we are looking for is did the student identify the relation between Figure 5.2 and Figure 5.3 and the range of speeds from these figures. Looking

Table 5.3: *KiU table for the Gasses in an Atmosphere task where we explicitly show the claim, KSAs, and ES*

Aspects of Knowledge in Use	Description
Claim	Use representation(s) of the speed distribution of a gas to reason about its composition or properties.
Focal knowledge, skills, and abilities (KSAs)	**KSA 1**: Unpack the relation between the given representation and the fraction of molecules with a particular range of speeds. **KSA 2**: Extract the relevant information from the representation to reason about the composition or properties of the gas. **KSA 3**: Make a conclusion regarding the composition of the gas based on the extracted information from the representation.
Evidence Statements (ES)	**ES 1**: Statements expressing the relation between given representation and the fraction of molecules with a particular range of speeds. **ES 2**: Statements concerning the extracted information from the representation to reason about the composition or properties of the gas. **ES 3**: Concluding statements regarding the composition of the gas based on the extracted information from the representation

at the student's written response they do not mention anything about the range of speeds from the figures, but in their response they mention "Figure 2 is more than likely hydrogen because of its higher probability of reaching higher velocities", showing that the student was able to identify the relation. This student's response for ES 1 would be coded as "VC" and would be marked as such in Table 5.4.

5.3.2 Evidence Statement 2

Next we looked to see if the student's answer is able to follow the second evidence statement of relating the extracted information from the figures to the composition of the gasses. From their written answer, we see that they mention "Figure II H_2 higher mean velocity", showing that the student was able to identify

		Skill Used?	
		Yes	No
Correct	Yes	X	
	No		

Table 5.4: *Validation table for student's free-response answer for ES 1*

that the figure with the higher velocity distribution corresponded to Hydrogen. The written "Figure I N_2 lower mean velocity" shows that they were also able to identify the graph that corresponds to Nitrogen. Looking at their verbal answer they were also able to identify determine that the composition of the second figure is Hydrogen due to the higher speed distribution. This student's response for ES 1 would be coded as "X" and would be marked as such in Table 5.5.

5.3.3 Evidence Statement 3

For the final evidence statement we are looking to see if the student is able to make a conclusion about which graph corresponds to which gas. In their written answer we see "Figure II H_2" and "Figure I N_2" which shows that they are able to determine which graph corresponds with which gas. We also see it in their verbal answer when they said "Figure 2 is more than likely hydrogen". This student's response for ES 1 would be coded as "VC" and would be marked as such in Table 5.6.

5.4 CMR Analysis

For an example of validating the CMR version of the tasks we analyzed a student's written and spoken responses to the Dry Ice task. The claim we want to make from this task is that the student is able to determine macroscopic features of an Einstein solid from given information. From this claim we decided the knowledge, skills, and abilities (KSAs) we want students to have when working on this task and also the evidence statements (ES)required for this knowledge. These aspects are highlighted in Table 5.7 From these aspects of KiU we are able to create the following task:

> You wish to use cubes of dry ice to cool a cube of ceramic (with a crystal structure). To do so, we will model dry ice (solid CO2) as an Einstein solid, such that each CO2 molecule is

		Skill Used?	
		Yes	No
Correct	Yes		
	No		X

Table 5.5: *Validation table for student's free-response answer for ES 2*

		Skill Used?	
		Yes	No
Correct	Yes	X	
	No		

Table 5.6: *Validation table for student's free-response answer for ES 3*

treated as having three independent 1D simple harmonic oscillators in orthogonal directions. The high temperature limit relationship between entropy and internal energy for an Einstein solid of N 1D oscillators is $S = Nk(ln(U) - ln(\epsilon N) + 1)$, where S is the total entropy, k is the Boltzmann constant, U is the internal energy, and ϵ is the oscillator energy unit.

Using the relationships between the total entropy, as expressed above, and internal energy for an Einstein solid, calculate the temperature as a function of internal energy. Represent your answer in terms of the internal energy (U), the oscillator energy unit (ϵ), the boltzmann constant (k), and the number of 1D oscillators (N).

Now, assume the ceramic can also be treated as an Einstein solid and thus has the same T(U) function that you solved for previously. For simplicity, assume the dry ice cubes have been cut such that they have the same number of oscillators (N) as the ceramic cube. The ceramic was precooled to 190 degrees Kelvin by other means, which is below the sublimation point of dry ice. The dry ice cubes start at 150 degrees Kelvin. How many cubes of ice do you need to cool the ceramic to 160 degrees Kelvin by letting the temperature equilibrate while thermally isolating the system of cubes?

For the next subsections we go in order of when the evidence statements appear in the CMR version. This means that we give the students the question for ES 3 first and then give the students the questions for ES 2 and then ending the task with the questions for ES 1. This is to ensure that we get the students' answer first and then have them explain their reasoning instead of possibly having the students' final answer influenced by the reasoning questions leading up to their final answer.

Table 5.7: *KiU table for the Dry Ice task where we explicitly show the claim, KSAs, and ES*

Aspects of Knowledge in Use	**Description**
Claim	Use mathematics to calculate the temperature and internal energy and use them to predict macroscopic features of an Einstein solid given information about the entropy.
Focal knowledge, skills, and abilities (KSAs)	**KSA 1**: Unpack the relationship between the temperature, entropy, and internal energy of an Einstein solid.
	KSA 2: Solve for the equation of temperature in terms of internal energy.
	KSA 3: Use the obtained relations to make a prediction about the macroscopic features of the Einstein solid.
Evidence Statements (ES)	**ES 1**:Statement of the relationship between temperature, entropy, and internal energy.
	ES 2: Statement of the relationship between temperature and internal energy. .
	ES 3: Statement of conclusion drawn from using the obtained relationships about the macroscopic features of the Einstein solid.

5.4.1 Evidence Statement 3

For ES 3 students are asked to give a statement the microscopic features in the scenario (in this case it is the number amount of ice cubes needed to cool). Here is how the CMR version is structured for this evidence statement:

1. Determine the number of dry ice cubes needed to cool the ceramic cube:

    A. 1/3

    B. 1/2

    C. 1

    D. 2

    E. 3

Once the student thought about and worked through the this given question, the student would then select their answer based on the responses above. Here is the transcript for the student working through ES 3:

Temperature in terms of... I see. I think there should be a differential that gives you what T is, I might be wrong. I think that should be somewhere. [Starts circling equation S = -T dF/dT with mouse] I feel like this exists in a different form, but I think it might be dS/dU maybe is equal to T. I forget the exact formula but I think it's something like that. [Starts to highlight the last paragraph of the task]. Um... So my formula of dS/dU = T, at least T of U in a sense, you want entropy within the system to be... So you want temperature to be the same between the two, right so, if you want NK... This is the [writes on paper] I get NK/U = T(U). Again this is assuming dS/dU = T which I'm not super sure that it is. I might be misremembering that equation which in that case what I'm doing isn't right. If it is correct then you want delta T to be equal to 190 - 160. This is 30 for ceramic and, or -30, and 10 for the cubes. So NK... delta U is equal to -30. [Repeatedly looks at work then task for a minute]. So we're not given any information about the mass of the ceramics? [Interviewer: That's Correct]. That's kind of confusing because my first inclination to solve something might also be to use Q, but Q depends on m, and you don't know m. I think another way of doing this is to say delta S is equal to Nk ln(Tf/Ti) because that's your final temperature, initial temperature. You just do, using Log rules. And then dU = T dS.. Find the number of ice cubes? [Looks at the phrase "the ceramic can also be treated as an Einstein function that you solved for previously"] Ok... Oh it has the same function ok I see. Oh, then if it has the same function then I think you can just, well you want to take the temperature/internal energy you know three times as much for the ceramic than increasing it for ice cubes then I think 3 ice cubes is how you would do that. (Student selects 3 ice cubes then moves on to the next question).

Before the student moves on to the next question, they fully answer the task including their equations and reasoning in order to select their answer. In the beginning of their work they start to use the equation dS/dU = T to solve out the T(U) asked in the task. The student then starts to question how do they get the number of ice cubes using this method, so they go back to reading the task to determine another way to solve how many ice cubes are needed to cool down the system. The student then says that the ice cubes and ceramic "have the same function" and decides that there needs to be three ice cubes for one ceramic. The student then selects the "3" answer and moves on to the next question. The student would be coded as

		Skill Used?	
		Yes	No
Correct	Yes	X	
	No		

Table 5.8: *Validation table for student's CMR answer for ES 3*

"VMC" as the student said out loud their answer with their reasoning and kept this consistent with what is on the CMR version. This answer is also the correct answer, so their validation chart is reflected in Table 5.8.

5.4.2 Evidence Statement 2

The next question asks them how did they solve for T(U) in the first place since that is what the task asks them to do first:

2. Determine the relationship of temperature and internal energy of an Einstein solid from the given relationship between entropy and internal energy:

    A. T = Nk/U

    B. T = 2Nk/U

    C. T = 3Nk/U

    D. T=U/Nk

    E. T=U/2Nk

    F. T = U/3Nk

The student then says the following:

> I want to say it's this [selects option T = Nk/U] because D = TdS. Oh wait... My bad so then 1/T, or rather T = dU/dS, so then dS/dU = 1/T. I mean I don't like saying what's differentials like that but, ya that would mean that it is U/Nk [Switches option to T = U/Nk and moves on to next question]

		Skill Used?	
		Yes	No
Correct	Yes	X	
	No		

Table 5.9: *Validation table for student's CMR answer for ES 2*

The student recalls from their work that when they solved for T using the equation dS/dU = T, they got the answer T = Nk/U and this was reflected in their selected answer in the CMR format. They then realized that the original equation they used to solve for T was wrong and used the correct relation dS/dU = 1/T to get the relation T = U/NK for their answer. The student then changes their answer to this in the CMR format and then moves on to the next question. This response would be coded as "VMC" as they said their answer out loud and marked down the same answer in the CMR version. This response would be the correct so the student's validation table would be reflected in Table 5.9.

5.4.3 Evidence Statement 1

The final question for the task asks them the relation the students ended up solving for the number of ice cubes:

3. How did you determine the ratio of dry ice cubes to ceramic cubes?

    A. Starting from the Thermodynamic Identity i.e. $dU = TdS - PdV + \mu dN$

    B. Considering the heat transfer with the environment

    C. Utilizing conservation of energy

    D. Looking at the equipartition change in the system

    E. Considering the heat capacity

The student then says the following:

> [Student selects option "Looking at the equipartition change in the system"] So ya looking at the net um, or temperature/internal energy difference [Student finishes this task].

		Skill Used?	
		Yes	No
Correct	Yes		
	No		X

Table 5.10: *Validation table for student's CMR answer for ES 1*

The student equates the original statement where they look at some temperature/internal energy difference and equate this to the equipartition function, selects this answer on the CMR version, then moves on to the next task. This student's response would be coded "XI" as they did make a statement the required relation for their reasoning and got it wrong so the validation table would be reflected in Table 5.10. In total it took about 9:02 for the student to complete this question.

5.5 Discussion

5.5.1 Evidence for Cognitive Validation

From the analyzed interview, we can easily put the student's responses into different categories based on what they used for their evidence and what they put as their answer. After going through all ten of the student responses, we added up all of the students' responses and determined from the responses that this task is valid for eliciting students' knowledge/skills required to answer the task. This evidence is found in Table 5.11, Table 5.12, and Table 5.13. Note that when students answer "I don't know" as their answer to one of the tasks we would mark this as "incorrect" and no for "skill used" as we believe that the answer "I don't know" means the students were not able to come up with a conclusion to the task

From the free response interviews we saw that we can use our way of collecting evidence for cognitive validation by seeing if a student response has the correct skill used and if the student response has the correct answer or reasoning. While think-aloud interviews are not novel, our methodology of analyzing the student response by determining if the student both used the correct skill and also if the student got the answer/reasoning correct is novel. Due to this methodology being novel we recommend that to determine if the task is evidence for the cognitive validity argument by analyzing if the students are using the correct pre-determined knowledge/skills and are not making any misinterpretations on the task.

Similar to the free-response interview, we can easily classify the student's responses into different

		Skill Used?	
		Yes	No
Correct	Yes	8	0
	No	1	1

Table 5.11: *Validation table for all ten student responses answer for ES 1 of extracting velocity from the graphs*

		Skill Used?	
		Yes	No
Correct	Yes	8	0
	No	0	2

Table 5.12: *Validation table for all ten student responses answer for ES 2 of using reasoning regarding the gasses' properties*

categories based on what they used for their evidence and what they put as their answer. From the CMR interview, we see the validation charts show that the student was able to make the correct conclusion regarding the number of ice cubes needed to cool down the ceramic and they found the correct relation for temperature and internal energy, but their reasoning regarding how they found the number of ice cubes was incorrect. All of this information will be highlighted in the feedback that the TaSPA gives to instructors. Regardless if the task is free-response or CMR format, we can use the same methodology of identifying for either format of identifying the student's' use of the correct skill and also if the student got the answer/reasoning correct.

Collecting evidence through CMR interviews is a key aspect for TaSPA as it highlights any disconnect between what the students say and write and what they select on the CMR format. This makes sure that the knowledge and skills the students exhibit are correct when giving feedback to instructors. This allows for instructors to improve on their instruction in the classroom.

		Skill Used?	
		Yes	No
Correct	Yes	8	0
	No	0	2

Table 5.13: *Validation table for all ten student responses answer for ES 3 of making a conclusion*

5.5.2 Evidence for Instructional Validation

Based on the expert interviews, we found that the learning performances outlined in the TaSPA are relevant to the learning goals found in their statistical physics courses. One of the biggest criticisms came from how we worded one of the tasks regarding people talking about a thermodynamic principle. The expert said that the correct answer was obvious because the student who was right "talked like a robot" while the other two students talked normally. This was changed as to make the conversation flow more naturally.

These expert interviews are important for the instructional validation component of the TaSPA as it allows us to see if the TaSPA is requiring the students to use the knowledge and skills to solve thermal and statistical physics problems found in the classroom. This allows for better feedback towards professors as these interviews show that the feedback will be relevant towards improving their course.

Chapter 6

Conclusions and Future Work

Throughout my dissertation, I worked on projects with the intent of better supporting faculty in responsible conduct of research and instruction. Throughout this dissertation, I highlighted two projects that focused on the "Developing: Reflective Teachers" by introducing methods for evaluating the effectiveness of RCR training, and designing and validating an assessment that focuses on assessing student knowledge and skills in a thermal or statistical physics course while also providing useful feedback to instructors such as recommended course modifications based on the evidence provided by the TaSPA.

6.1 Goals and Values in Science

In the end, we found that guided deliberations on the values implicit in real examples of scientific practice shifted scientists' invoked values with respect to research ethics. After scientists participated in the fellowship meetings, we found that scientists' moral sensitivity improved: they not only noticed more values in the RCR Vignettes, but also invoked more unique values. We did not see a change in their moral reasoning initially, however we did identify a new way of looking at moral reasoning. This led us to identify that value focused discussions improved scientists' moral sensitivity, but there was more to look at for scientists' moral reasoning.

Since we suspected that the scientists' moral reasoning changed due to the larger number of reasonings after the fellowship, we introduced a more fine grained measure to identify moral reasoning. On top of

using Rest's moral schema to identify the level of reasoning, we also identified the stakeholders they invoked and the principles they used. Through this we were able to break down the scientists' moral reasoning even further to determine how scientists were reasoning in the scenario.

In the end, we want scientists to be more attentive towards ethical issues throughout their work and to ultimately make better ethical judgements about both research conduct and the broader consequences of their work. Our findings suggest that they actually were more attentive to ethical values after the fellowship and we believe that the scientists' moral reasoning changed. This suggests there is promise in approaches to ethics training that focus on the general role of values in scientific practice. We hope that other people in PER use the Four Component Model when trying to improve scientists' ethical decision making.

6.2 Task Development

We developed a new methodology to generate faculty feedback through evidence acquired from assessments. More specifically, we explored how to develop a thermal and statistical physics assessment that focused on providing timely and useful feedback for instructors. We do this through collecting evidence from students, guided by knowledge-in-use tables that identify a relevant learning performance, proficiencies that go along with the learning performance, evidence from the students based on the required proficiencies, and task features on how to illicit the evidence from the students. We created feedback for instructors that focused on the specific knowledge and skills students were struggling with and potential ways to improve on their courses.

The TaSPA consists of sixteen tasks with a variety of scientific practices, crosscutting concepts, and core ideas. In the process of developing the assessment, we collected data from 286 students from 6 unique universities across the United States to help us construct the CMR versions of our tasks and to make sure that the tasks were relevant and at the right level for the students. We also interviewed 16 faculty members about the relevancy of the learning performances presented in the TaSPA. The faculty members also gave us valuable information on how to make the feedback more informational and easier to read like how we needed to add on the feedback reports the breakdown of the partially met breakdowns.

My part in this project was to finalize the details for developing tasks using the knowledge-in-use framework. As the framework for developing the tasks was written before I joined the project, Chapter IV

reflects the finalized version of how to develop tasks for the TaSPA. I led development five tasks for the assessment as well as adding in the step to the development process of writing out what do the different partially-mets mean as to provide more clarity for instructors when they receive feedback.

6.3 Task Validation

With the tasks completed for the TaSPA, the next step for this project was to develop a methodology to validate the assessment to make sure that the feedback provided for instructors is useful. This includes making sure that the tasks are relevant to thermal and statistical physics, the feedback is useful and can be easily understood for instructors, and the assessment scores are representative of student knowledge and skills.

We created the TaSPA to support instructors on how assessment scores can be used to benefit course instruction. Before we can provide useful and clear feedback for professors, we had to validate the assessment so that the tasks are eliciting the knowledge and skills expected of them in an undergraduate thermal or statistical physics class while providing useful feedback to instructors. Since we used a relatively new method creating the TaSPA using Evidence-Centered Design, we needed to identify a new way to validate the tasks by looking to see if the tasks successfully invoke the evidence statements from student interviews.

We adopted new components for validating an assessment. In Chapter V, I introduced new components of validation which were cognitive validation, instructional validation, and inferential validation. After introducing these components to PER, I then introduced a methodology that focused on receiving evidence for cognitive validation through student interviews and receiving evidence for instructional validation through expert interviews.

From the student interviews, we looked at both the free response formats and CMR formats and saw if the students are using both the required skills and knowledge to solve for tasks given in a thermal and statistical physics class. If the students were not using the right knowledge in skills in their answers but still getting all of the questions correct then we could pinpoint down the issues with the validation charts and fix these issues. Through this new methodology in PER we can now receive evidence towards cognitive validation.

For the expert interviews, we wanted to see comments from experts in the fields of statistical and

thermal physics on the relevance of the learning performances and tasks to the learning goals in their class. The experts agreed that the learning performances were relevant to the learning goals in their classroom when applicable. In summary, we adapted a new way to use expert interviews to determine if the tasks are at the right level for the students while also making sure that the proficiencies outlined in the learning performance line up with the tasks and that the tasks are relevant to the experts' courses.

6.4 Future Work

For the GVS project, in order to determine how values effect the full process of ethical decision making, we need to determine a way to see how the value focused discussions effected the scientists' moral motivation. We would also like to see how their moral implementation was effected by the value focused discussions, but in order to accurately see how the scientists' moral implementation changed, we would need to see the scientists react in an actual scenario with no outside intervention regarding the scenario which would be hard to gauge. You and Bebeau used scores from a Professional Problem Solving course but they stated that currently the relationship between the scores in the class and long-term competence has not been established.[28] Instead of focusing on all of the components, future work should consist of identifying and improving scientists' moral sensitivity, moral reasoning, and moral motivation as all of the components are effected by each other.[29] We have identified how to look at moral sensitivity and moral reasoning, but we have not identified how to identify participants' moral motivation yet. Also while we have identified a new way to look at moral reasoning, we still need to identify how does someone improve using the new methodology. One way to identify how the scientists' moral motivation was effected by the fellowship is to apply the components of the Professional Role Orientation Inventory which are authority, responsibility, agency, and autonomy as this is what You and Bebeau did to identify dentists' moral motivation.[28;50]

For TaSPA Task Development, we currently need to finish up the remaining tasks to get a final total of sixteen tasks for the assessment. Once we finish the remaining tasks we need to put the assessment on LASSO to make the TaSPA more available for instructors. Future work for the TaSPA is validating the assessment, which is described in the next paragraph.

For TaSPA validation, my work work a method to collect evidence for cognitive validation for the assessment. Now that I have identified a method for collecting evidence for cognitive validation, future

work will consist of applying the validation tables for *all sixteen* of the CMR tasks. While we identified how to collect evidence for cognitive validation, in order for the TaSPA to be valid we also need to consider instructional validation and inferential validation. Future work regarding instructional validation will consist of interviewing faculty to determine if the feedback is useful and makes sense. We will also need to determine a way to see, once instructors receive feedback, how do instructors apply the feedback to improve their class. We also need to see how instructors apply the feedback does the feedback improve their class. We would also need new instructors to look at the tasks as we created new tasks since our last instructor interviews. For inferential validation future work will consist of identifying and applying a statistical measure to make sure that the test scores accurately reflect student performance.

Bibliography

[1] Amogh Sirnoorkar, Amali Priyanka Jambuge, Katherine D. Rainey, Alexander Adamson, Bethany R. Wilcox, and James T. Laverty. Theoretical Approach for Providing Feedback for Instructors Through a Standardized Assessment for Undergraduate Physics. 2023. URL https://repository.isls.org//handle/1/9912. Publisher: International Society of the Learning Sciences.

[2] James W Pellegrino. The Design of an Assessment System for the Race to the Top: A Learning Sciences Perspective on Issues of Growth and Measurement. *Performance Management*.

[3] *Developing Assessments for the Next Generation Science Standards*. National Academies Press, Washington, D.C., May 2014. ISBN 978-0-309-28951-1. doi: 10.17226/18409. URL http://www.nap.edu/catalog/18409.

[4] James W. Pellegrino. Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *PsicologÃa Educativa*, 20(2):65–77, December 2014. ISSN 1135-755X. doi: 10.1016/j.pse.2014.11.002. URL https://www.sciencedirect.com/science/article/pii/S1135755X14000128.

[5] Caleb L. Linville, Aidan C. Cairns, Tyler Garcia, Bill Bridges, Jonathan Herington, James T. Laverty, and Scott Tanona. How Do Scientists Perceive the Relationship Between Ethics and Science? A Pilot Study of Scientists' Appeals to Values. *Science and Engineering Ethics*, 29(3):15, April 2023. ISSN 1471-5546. doi: 10.1007/s11948-023-00429-1. URL https://doi.org/10.1007/s11948-023-00429-1.

[6] James W. Pellegrino, Louis V. DiBello, and Susan R. Goldman. A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments. *Educational Psychologist*, 51(1):59–81, January 2016. ISSN 0046-1520. doi: 10.1080/00461520.2016.1145550. URL

https://doi.org/10.1080/00461520.2016.1145550. Publisher: Routledge _eprint: https://doi.org/10.1080/00461520.2016.1145550.

[7] Dennis W. Sunal, Jeanelle Hodges, Cynthia S. Sunal, Kevin W. Whitaker, L. Michael Freeman, Leo Edwards, Ronald A. Johnston, and Michael Odell. Teaching Science in Higher Education: Faculty Professional Development and Barriers to Change. *School Science and Mathematics*, 101 (5):246–257, 2001. ISSN 1949-8594. doi: 10.1111/j.1949-8594.2001.tb18027.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1949-8594.2001.tb18027.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1949-8594.2001.tb18027.x.

[8] Sheila Tobias. They're Not Dumb, They're Different–Stalking the Second Tier. Technical report, Science News Books, 1719, 1990. ERIC Number: ED331702.

[9] Andrea Venezia and Laura Jaeger. Transitions from High School to College. *The Future of Children*, 23(1):117–136, March 2013. ISSN 1550-1558. doi: 10.1353/foc.2013.0004. URL https://muse.jhu.edu/article/508223.

[10] Frances A. Houle, Kate P. Kirby, and Michael P. Marder. Ethics in physics: The need for culture change. *Physics Today*, 76(1):28–35, January 2023. ISSN 0031-9228. doi: 10.1063/PT.3.5156. URL https://physicstoday.scitation.org/doi/full/10.1063/PT.3.5156. Publisher: American Institute of Physics.

[11] Thomas R Guskey. Does It Make a Difference? Evaluating Professional Development, March 2002. URL https://www.ascd.org/el/articles/does-it-make-a-difference-evaluating-professional-development.

[12] Kate Kirby and Frances A. Houle. Ethics and the Welfare of the Physics Profession. *Physics Today*, 57(11):42–46, November 2004. ISSN 0031-9228. doi: 10.1063/1.1839376. URL https://physicstoday.scitation.org/doi/full/10.1063/1.1839376. Publisher: American Institute of Physics.

[13] Michael Mumford. *Read "Fostering Integrity in Research" at NAP.edu*. doi: 10.17226/21896. URL https://www.nap.edu/read/21896/chapter/20.

[14] Charles Henderson. Common Concerns About the Force Concept Inventory. *The Physics Teacher*, 40:542–547, December 2002. doi: 10.1119/1.1534822.

[15] 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. ISSN 1098-2736. doi: 10.1002/tea.20439. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/tea.20439. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/tea.20439.

[16] 19.1 Guidelines on Ethics (Full Statement). URL http://www.aps.org/policy/statements/guidlinesethics.cfm.

[17] *Read "Fostering Integrity in Research" at NAP.edu*. doi: 10.17226/21896. URL https://www.nap.edu/read/21896/chapter/2.

[18] Michael J. Reiss. Teaching Ethics in Science. *Studies in Science Education*, 34(1):115–140, January 1999. ISSN 0305-7267, 1940-8412. doi: 10.1080/03057269908560151. URL http://www.tandfonline.com/doi/abs/10.1080/03057269908560151.

[19] Responsible Conduct of Research (RCR) Basic | CITI Program. URL https://about.citiprogram.org/en/course/responsible-conduct-of-research-basic/.

[20] Sean Powell, Matthew Allison, and Michael Kalichman. Effectiveness of a Responsible Conduct of Research Course: A Preliminary Study. *Science and engineering ethics*, 13:249–64, July 2007. doi: 10.1007/s11948-007-9012-y.

[21] Alison L. Antes, Stephen T. Murphy, Ethan P. Waples, Michael D. Mumford, Ryan P. Brown, Shane Connelly, and Lynn D. Devenport. A Meta-Analysis of Ethics Instruction Effectiveness in the Sciences. *Ethics & Behavior*, 19(5):379–402, September 2009. ISSN 1050-8422. doi: 10.1080/10508420903035380. URL https://doi.org/10.1080/10508420903035380. Publisher: Routledge, eprint: https://doi.org/10.1080/10508420903035380.

[22] Matthew J. Brown. Values in Science beyond Underdetermination and Inductive Risk. *Philosophy*

of Science, 80(5):829–839, 2013. ISSN 0031-8248. doi: 10.1086/673720. URL https://www.jstor.org/stable/10.1086/673720.

[23] Heather Douglas. Inductive Risk and Values in Science. *Philosophy of Science*, 67(4):559–579, 2000. ISSN 0031-8248. URL https://www.jstor.org/stable/188707. Publisher: [The University of Chicago Press, Philosophy of Science Association].

[24] Greg Lusk and Kevin C. Elliott. Non-epistemic values and scientific assessment: an adequacy-for-purpose view. *European Journal for Philosophy of Science*, 12(2):35, June 2022. ISSN 1879-4920. doi: 10.1007/s13194-022-00458-w. URL https://doi.org/10.1007/s13194-022-00458-w.

[25] M. J. B. Stokhof. Ethics and morality, principles and practice. *Zeitschrift fÃŒr Ethik und Moralphilosophie*, 1(2):291–304, October 2018. ISSN 2522-0071. doi: 10.1007/s42048-018-0016-x. URL https://doi.org/10.1007/s42048-018-0016-x.

[26] James R. Rest and Darcia Narvaez, editors. *Moral development in the professions: Psychology and applied ethics*. Moral development in the professions: Psychology and applied ethics. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1994. ISBN 978-0-8058-1538-2 978-0-8058-1539-9. Pages: xii, 233.

[27] Darcia Narvaez and J.R. Rest. The four components of acting morally. Moral behavior and moral development: An introduction. *Handbook of moral and character education*, pages 385–400, January 1995.

[28] Di You and Muriel J. Bebeau. The independence of James Rest's components of morality: evidence from a professional ethics curriculum study. *Ethics and Education*, 8(3):202–216, November 2013. ISSN 1744-9642, 1744-9650. doi: 10.1080/17449642.2013.846059. URL http://www.tandfonline.com/doi/abs/10.1080/17449642.2013.846059.

[29] James R. Rest. Research on Moral Development: Implications for Training Counseling Psychologists. *The Counseling Psychologist*, 12(3):19–29, September 1984. ISSN 0011-0000. doi: 10.1177/

0011000084123003. URL https://doi.org/10.1177/0011000084123003. Publisher: SAGE Publications Inc.

[30] Laura S. Hamilton, Brian M. Stecher, Stephen P. Klein, and National Science Foundation (U.S.), editors. *Making sense of test-based accountability in education*. Rand, Santa Monica, CA, 2002. ISBN 978-0-8330-3161-7.

[31] David Hestenes, Malcolm Wells, and Gregg Swackhamer. Force Concept Inventory. *The Physics Teacher*, 30:141–158, March 1992. doi: 10.1119/1.2343497.

[32] PhysPort Assessments, . URL https://www.physport.org/assessments/.

[33] Carl Kaestle. Testing Policy in the United States: A Historical Perspective.

[34] 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, February 2016. ISSN 2469-9896. doi: 10.1103/PhysRevPhysEducRes.12.010115. URL https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.12.010115.

[35] Kristen Huff, Linda Steinberg, and Thomas Matts. The Promises and Challenges of Implementing Evidence-Centered Design in Large-Scale Assessment. *Applied Measurement in Education*, 23(4):310–324, September 2010. ISSN 0895-7347. doi: 10.1080/08957347.2010.510956. URL https://doi.org/10.1080/08957347.2010.510956. Publisher: Routledge _eprint: https://doi.org/10.1080/08957347.2010.510956.

[36] 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. ISSN 1745-3992. doi: 10.1111/emip.12253. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/emip.12253. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/emip.12253.

[37] Katherine D. Rainey, Michael Vignal, and Bethany R. Wilcox. Designing upper-division thermal physics assessment items informed by faculty perspectives of key content coverage. *Phys-*

ical Review Physics Education Research, 16(2):020113, August 2020. ISSN 2469-9896. doi: 10.1103/PhysRevPhysEducRes.16.020113. URL https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.16.020113.

[38] American Educational Research Association, editor. *Report and recommendations for the reauthorization of the institute of education sciences*. American Educational Research Association, Washington, D.C, 2011. ISBN 978-0-935302-35-6. OCLC: ocn826867074.

[39] Paula V Engelhardt. An Introduction to Classical Test Theory as Applied to Conceptual Multiple-choice Tests, April 2009. URL https://www.per-central.org/items/detail.cfm?ID=8807.

[40] Nonye Alozie, Phyllis Haugabook Pennock, Krystal Madden, Sania Zaidi, Christopher J. Harris, and Joseph S. Krajcik. Designing and Developing NGSS-Aligned Formative Assessment Tasks to Promote Equity. In *Paper presented at the annual conference of National Association for Research in Science Teaching*, March 2018. URL http://nextgenscienceassessment.org/wp-content/uploads/2018/06/4-Alozie_etal_NARST-2018-Equity-in-Assessment-Design_2018.pdf.

[41] Kevin C. Elliott. Values in Science. *Elements in the Philosophy of Science*, June 2022. doi: 10.1017/9781009052597. URL https://www.cambridge.org/core/elements/values-in-science/8C9899A25764AA9A791287325A511C3C. ISBN: 9781009052597 9781009055635 Publisher: Cambridge University Press.

[42] Liam Kofi Bright. Du Bois' democratic defence of the value free ideal. *Synthese*, 195(5):2227–2245, May 2018. ISSN 1573-0964. doi: 10.1007/s11229-017-1333-z. URL https://doi.org/10.1007/s11229-017-1333-z.

[43] James Rest, Darcia Narvaez, Stephen Thoma, and Muriel Bebeau. DIT2: Devising and Testing A Revised Instrument of Moral Judgment. *Journal of Educational Psychology*, 91:644–659, December 1999. doi: 10.1037/0022-0663.91.4.644.

[44] Muriel Bebeau, J Rest, and C Yamoor. Measuring dental students' ethical sensitivity. *Journal of dental education*, 49:225–35, May 1985. doi: 10.1002/j.0022-0337.1985.49.4.tb01874.x.

[45] P. C. HÃ©bert, E. M. Meslin, and E. V. Dunn. Measuring the ethical sensitivity of medical students: a study at the University of Toronto. *Journal of Medical Ethics*, 18(3):142–147, September 1992. ISSN 0306-6800. doi: 10.1136/jme.18.3.142.

[46] About the DIT. URL https://ethicaldevelopment.ua.edu/about-the-dit.html.

[47] James Rest, Darcia Narvaez, Muriel Bebeau, and Stephen Thoma. A neo-Kohlbergian approach: The DIT and schema theory. *Educational Psychology Review*, 11(4):291–324, 1999. ISSN 1573-336X(Electronic),1040-726X(Print). doi: 10.1023/A:1022053215271. Place: Germany Publisher: Springer.

[48] Michael D. Mumford, Lynn D. Devenport, Ryan P. Brown, Shane Connelly, Stephen T. Murphy, Jason H. Hill, and Alison L. Antes. ARTICLES: Validation of Ethical Decision Making Measures: Evidence for a New Set of Measures. *Ethics & Behavior*, 16 (4):319–345, October 2006. ISSN 1050-8422. doi: 10.1207/s15327019eb1604_4. URL https://doi.org/10.1207/s15327019eb1604_4. Publisher: Routledge _eprint: https://doi.org/10.1207/s15327019eb1604_4.

[49] Joann Franklin Klinker and Donald G. Hackmann. An Analysis of Principals' Ethical Decision Making Using Rest's Four Component Model of Moral Behavior. *Journal of School Leadership*, 14(4):434–456, July 2004. ISSN 1052-6846. doi: 10.1177/105268460401400404. URL https://doi.org/10.1177/105268460401400404. Publisher: SAGE Publications Inc.

[50] M. J. Bebeau, D. O. Born, and D. T. Ozar. The development of a professional role orientation inventory. *The Journal of the American College of Dentists*, 60(2):27–33, 1993. ISSN 0002-7979.

[51] Trisha Phillips, Franchesca Nestor, Gillian Beach, and Elizabeth Heitman. America Competes at 5 Years: An Analysis of Research-Intensive Universities? Rcr Training Plans. *Science and Engineering Ethics*, 24(1):227–249, 2018. doi: 10.1007/s11948-017-9883-5. Publisher: Springer Verlag.

[52] James M. Dubois and Jeffrey M. Dueker. Teaching and Assessing the Responsible Conduct of Research: A Delphi Consensus Panel Report. *The Journal of Research Administration*, 40(1):49–70, 2009. ISSN 1539-1590.

[53] Michael W. Kalichman. Responding to Challenges in Educating for the Responsible Conduct of Research. *Academic Medicine*, 82(9):870, September 2007. ISSN 1040-2446. doi: 10.1097/ACM.0b013e31812f77fe. URL https://journals.lww.com/academicmedicine/fulltext/2007/09000/responding_to_challenges_in_educating_for_the.10.aspx.

[54] Robert Pennock and Michael O'Rourke. Developing a Scientific Virtue-Based Approach to Science Ethics Training. *Science and engineering ethics*, 23, February 2017. doi: 10.1007/s11948-016-9757-2.

[55] Aidan C Cairns, Caleb Linville, Tyler Garcia, Bill Bridges, Scott Tanona, Jonathan Herington, and James T Laverty. A phenomenographic study of scientists' beliefs about the causes of scientists' research misconduct. *Research Ethics*, page 17470161211042658, September 2021. ISSN 1747-0161. doi: 10.1177/17470161211042658. URL https://doi.org/10.1177/17470161211042658. Publisher: SAGE Publications Ltd.

[56] Chris Argyris. Learning and Teaching: A Theory of Action Perspective. *Journal of Management Education*, 21(1):9–26, February 1997. ISSN 1052-5629. doi: 10.1177/105256299702100102. URL https://doi.org/10.1177/105256299702100102. Publisher: SAGE Publications Inc.

[57] Randy Elliot Bennett. Cognitively Based Assessment of, for, and as Learning (CBAL): A Preliminary Theory of Action for Summative and Formative Assessment. *Measurement: Interdisciplinary Research and Perspectives*, 8(2-3):70–91, August 2010. ISSN 1536-6367. doi: 10.1080/15366367.2010.508686. URL https://doi.org/10.1080/15366367.2010.508686. Publisher: Routledge _eprint: https://doi.org/10.1080/15366367.2010.508686.

[58] Gail M. Sullivan. A Primer on the Validity of Assessment Instruments. *Journal of Graduate Medical*

Education, 3(2):119–120, June 2011. ISSN 1949-8349. doi: 10.4300/JGME-D-11-00075.1. URL https://doi.org/10.4300/JGME-D-11-00075.1.

[59] Julius Sim and Peggy Arnell. Measurement Validity in Physical Therapy Research. *Physical Therapy*, 73(2):102–110, February 1993. ISSN 0031-9023, 1538-6724. doi: 10.1093/ptj/73.2.102. URL https://academic.oup.com/ptj/article/2729073/Measurement.

[60] S Maulita, Sukarmin Sukarmin, and A Marzuki. The Content Validity: Two-Tier Multiple Choices Instrument to Measure Higher-Order Thinking Skills. *Journal of Physics: Conference Series*, 1155: 012042, February 2019. doi: 10.1088/1742-6596/1155/1/012042.

[61] C.J. Harris, E. Wiebe, S. Grover, and J. W. Pellegrino. Classroom-based STEM assessment: Contemporary issues and perspectives. Technical report, Community for Advancing Discovery Research in Education (CADRE). Education Development Center, Inc, 2023.

[62] Daniel M. Hausman. *Preference, Value, Choice, and Welfare*. Cambridge University Press, 2011. doi: 10.1017/CBO9781139058537.

[63] Shelly Kagan. *Normative Ethics*. Routledge, February 2018. ISBN 978-0-429-96720-7. Google-Books-ID: 7f_EDwAAQBAJ.

[64] Daniel Steel. Epistemic Values and the Argument from Inductive Risk. *Philosophy of Science*, 77(1):14–34, January 2010. ISSN 0031-8248, 1539-767X. doi: 10.1086/650206. URL https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/epistemic-values-and-the-argument-from-inductive-risk/CAC4C1C1D4C5CB37EE50D5C3B2D99F9E.

[65] Stephen J. Thoma. The Defining Issues Test of moral judgment development. *Behavioral Development Bulletin*, 19(3):55, 2014. ISSN 1942-0722. doi: 10.1037/h0100590. URL /fulltext/2014-55726-010.html. Publisher: US: Joseph D. Cautilli.

[66] Stephen J. Thoma. Measuring moral thinking from a neo-Kohlbergian perspective. *Theory and Research in Education*, 12(3):347–365, November 2014. ISSN 1477-8785, 1741-

3192. doi: 10.1177/1477878514545208. URL http://journals.sagepub.com/doi/10.1177/1477878514545208.

[67] Liisa Myyry and Klaus Helkama. The Role of Value Priorities and Professional Ethics Training in Moral Sensitivity. *Journal of Moral Education*, 31(1):35–50, March 2002. ISSN 0305-7240. doi: 10.1080/03057240120111427. URL https://doi.org/10.1080/03057240120111427. Publisher: Routledge _eprint: https://doi.org/10.1080/03057240120111427.

[68] Henriikka Clarkeburn, J. Roger Downie, and Bob Matthew. Impact of an Ethics Programme in a Life Sciences Curriculum. *Teaching in Higher Education*, 7(1):65–79, January 2002. ISSN 1356-2517. doi: 10.1080/13562510120100391. URL https://doi.org/10.1080/13562510120100391. Publisher: Routledge _eprint: https://doi.org/10.1080/13562510120100391.

[69] Henry Hsu and Peter A. Lachenbruch. Paired t Test. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, 2014. ISBN 978-1-118-44511-2. doi: 10.1002/9781118445112.stat05929. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat05929. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat05929.

[70] Viv Bewick, Liz Cheek, and Jonathan Ball. Statistics review 8: Qualitative data - tests of association. *Critical Care*, 8(1):46–53, 2004. ISSN 1364-8535. doi: 10.1186/cc2428. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC420070/.

[71] Hae-Young Kim. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative Dentistry & Endodontics*, 42(2):152–155, March 2017. doi: 10.5395/rde.2017.42.2.152. URL https://synapse.koreamed.org/articles/1090217. Publisher: The Korean Academy of Conservative Dentistry.

[72] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971. ISSN 1939-1455. doi: 10.1037/h0031619. Place: US Publisher: American Psychological Association.

[73] Kevin A. Hallgren. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23–34, 2012. ISSN 1913-4126. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/.

[74] Julius Sim and Chris C Wright. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, 85(3):257–268, March 2005. ISSN 0031-9023. doi: 10.1093/ptj/85.3.257. URL https://doi.org/10.1093/ptj/85.3.257.

[75] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159, March 1977. ISSN 0006341X. doi: 10.2307/2529310. URL https://www.jstor.org/stable/2529310?origin=crossref.

[76] Tom L Beauchamp, James F Childress, and Oxford University Press. *Principles of biomedical ethics*. Oxford University Press, New York; Oxford, 2013. ISBN 978-0-19-992458-5. OCLC: 827736605.

[77] R Edward Freeman. *The Stakeholder Concept and Stragegic Management*. Number 2. Pitman Publishing Inc., 1984.

[78] Marjolein C. Achterkamp and Janita F. J. Vos. Critically identifying stakeholders. *Systems Research and Behavioral Science*, 24(1):3–14, 2007. ISSN 1099-1743. doi: 10.1002/sres.760. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sres.760. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sres.760.

[79] Ulrich Werner. Critical heuristics of social systems design. *European Journal of Operational Research*, 31(3):276–283, 1987.

[80] James W. Pellegrino and Margaret L. Hilton. *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. National Academies Press, Washington, D.C., December 2012. ISBN 978-0-309-25649-0. doi: 10.17226/13398. URL http://www.nap.edu/catalog/13398.

[81] 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, April

2018. ISSN 2469-9896. doi: 10.1103/PhysRevPhysEducRes.14.010123. URL https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.14.010123.

[82] Three Dimensional Learning | Next Generation Science Standards, . URL https://www.nextgenscience.org/three-dimensional-learning.

[83] *Read "A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas" at NAP.edu.* . doi: 10.17226/13165. URL https://nap.nationalacademies.org/read/13165/chapter/7.

[84] Henry Petroski. *Invention by Design: How Engineers Get from Thought to Thing*. Harvard University Press, 1996. ISBN 978-0-674-46367-7. doi: 10.2307/j.ctv1pncnzm. URL https://www.jstor.org/stable/j.ctv1pncnzm.

[85] Samuel C. Florman. *The Existential Pleasures of Engineering*. Macmillan, 1994.

[86] NGSS States, . URL https://www.twigscience.com/ngss-states/.

[87] Deborah L. Butler and Philip H. Winne. Feedback and Self-Regulated Learning: A Theoretical Synthesis, 1995. URL https://journals.sagepub.com/doi/10.3102/00346543065003245.

[88] Monique Boekaerts. Self-regulated learning: where we are today. *International Journal of Educational Research*, 31(6):445–457, January 1999. ISSN 0883-0355. doi: 10.1016/S0883-0355(99)00014-2. URL https://www.sciencedirect.com/science/article/pii/S0883035599000142.

[89] Barry J. Zimmerman and Dale H. Schunk, editors. *Self-regulated learning and academic achievement: Theory, research, and practice*. Self-regulated learning and academic achievement: Theory, research, and practice. Springer-Verlag Publishing, New York, NY, US, 1989. ISBN 978-0-387-96934-3 978-3-540-96934-1. doi: 10.1007/978-1-4612-3618-4. Pages: xiv, 212.

[90] D. Royce Sadler. Formative assessment and the design of instructional systems. *Instructional Science*, 18(2):119–144, June 1989. ISSN 1573-1952. doi: 10.1007/BF00117714. URL https://doi.org/10.1007/BF00117714.

[91] 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, April 2006. ISSN 0307-5079, 1470-174X. doi: 10.1080/03075070600572090. URL http://www.tandfonline.com/doi/abs/10.1080/03075070600572090.

[92] Jeremy Williams. Creating authentic assessments: A method for the authoring of open book open web examinations. January 2004.

[93] Kim H. Koh. Authentic Assessment. In *Oxford Research Encyclopedia of Education*. February 2017. ISBN 978-0-19-026409-3. doi: 10.1093/acrefore/9780190264093.013.22. URL https://oxfordre.com/education/display/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-22?source=post_page----------------------------.

[94] Otto cycle - Energy Education, . URL https://energyeducation.ca/encyclopedia/Otto_cycle.

[95] 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, October 2014. ISSN 1554-9178. doi: 10.1103/PhysRevSTPER.10.020124. URL https://link.aps.org/doi/10.1103/PhysRevSTPER.10.020124.

[96] Adrian Madsen, Sarah B. McKagan, and Eleanor C. Sayre. Best Practices for Administering Concept Inventories. *The Physics Teacher*, 55(9):530–536, December 2017. ISSN 0031-921X. doi: 10.1119/1.5011826. URL https://doi.org/10.1119/1.5011826.

[97] Michael D. Wolcott and Nikki G. Lobczowski. Using cognitive interviews and think-aloud protocols to understand thought processes. *Currents in Pharmacy Teaching and Learning*, 13(2):

181–188, February 2021. ISSN 1877-1297. doi: 10.1016/j.cptl.2020.09.005. URL https://www.sciencedirect.com/science/article/pii/S1877129720303026.

Appendix A

KiU Tables for created Tasks

Table A.1: *Table listing all tasks I worked on with their learning performances as well as their associated scientific practice, core idea, and crosscutting concept*

Task Name	Learning Performance	Scientific Practices	Core Ideas	Crosscutting Concepts
MRI	Use models to determine the number of microstates for a given macrostate and find the probability of a system being in that particular macrostate.	Developing and Using Models	States/Boltzman Factor	Probability
Gasses in Atmosphere	Use representation(s) of the speed distribution of a gas to reason about its composition or properties.	Developing and Using Models	Energy	Scale and Proportion
Alloy Properties	Use mathematics to determine the heat capacity of argon and helium from the monatomic gas partition function and see if either gasses are usable in an experiment.	Using Math	Partition Function	Scale
Heat Engines	Analyze and interpret data about the variation of pressure and volume inside an engine to determine the efficiency for a given amount of heat supplied over one cycle.	Analyze and interpret data	Energy	Heat Flow
Equations of State	Construct an explanation on how the physical dynamics of gasses determines if a scenario can be described with the ideal gas law	Constructing Explanations	Properties of Matter	System and System Models

Table A.2: *KiU table for the MRI task where we explicitly show the Learning Performance, KSAs, ES, and Task Teatures*

Aspects of Knowledge in Use	**Description**
Learning Performance	Use models to determine the number of microstates for a given macrostate and find the probability of a system being in that particular macrostate.
Focal knowledge, skills, and abilities (KSAs)	KSA1: Unpack the relation between the probability of a given system to be in the given macrostate (P(Ω)) and the number of accessible microstates (Ω). KSA2: Determine the probability of a system to be in the given macrostate through the unpacked relation.
Evidence Statements (ES)	ES1: Relation between the probability of a given system to be in the given macrostate (P(Ω)) and the number of accessible microstates (Ω). ES2: Statements or mathematical expressions conveying the probability of a given system to be in the given macrostate (P(Ω)).
Task Features	Question gives an event, observation, or phenomenon for the student to explain or make a prediction about. Question gives a representation or asks student to construct a representation. Question asks student to explain or make a prediction about the event, observation, or phenomenon. Question asks student to provide the reasoning that links the representation to their explanation or prediction.

Table A.3: *KiU table for the Gasses in Atmosphere task where we explicitly show the Learning Performance, KSAs, ES, and Task Teatures*

Aspects of Knowledge in Use	**Description**
Learning Performance	Use representation(s) of the speed distribution of a gas to reason about its composition or properties.
Focal knowledge, skills, and abilities (KSAs)	KSA1: Unpack the relation between the given representation and the fraction of molecules with a particular range of speeds. KSA2: Extract the relevant information from the representation to reason about the composition or properties of the gas. KSA3: Make a conclusion regarding the composition of the gas based on the extracted information from the representation.
Evidence Statements (ES)	ES1: Statements expressing the relation between the given representation and the fraction of molecules with a particular range of speeds. ES2: Statements concerning the extracted information from the representation to reason about the composition or properties of the gas. ES3: Concluding statements regarding the composition of the gas based on the extracted information from the representation.
Task Features	Question gives an event, observation, or phenomenon for the student to explain or make a prediction about. Question gives a representation or asks student to construct a representation. Question asks student to explain or make a prediction about the event, observation, or phenomenon. Question asks student to provide the reasoning that links the representation to their explanation or prediction.

Table A.4: *KiU table for the Alloy Properties task where we explicitly show the Learning Performance, KSAs, ES, and Task Teatures*

Aspects of Knowledge in Use	**Description**
Learning Performance	Use representation(s) of the speed distribution of a gas to reason about its composition or properties.
Focal knowledge, skills, and abilities (KSAs)	KSA1: Unpack the relation between the given representation and the fraction of molecules with a particular range of speeds. KSA2: Extract the relevant information from the representation to reason about the composition or properties of the gas. KSA3: Make a conclusion regarding the composition of the gas based on the extracted information from the representation.
Evidence Statements (ES)	ES1: Statements expressing the relation between the given representation and the fraction of molecules with a particular range of speeds. ES2: Statements concerning the extracted information from the representation to reason about the composition or properties of the gas. ES3: Concluding statements regarding the composition of the gas based on the extracted information from the representation.
Task Features	Question gives an event, observation, or phenomenon Question asks student to perform a calculation or statistical test, generate a mathematical representation, or demonstrate a relationship between parameters. Question 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.

Table A.5: *KiU table for the Heat Engine task where we explicitly show the Learning Performance, KSAs, ES, and Task Teatures*

Aspects of Knowledge in Use	**Description**
Learning Performance	Analyze and interpret data about the variation of pressure and volume inside an engine to determine the efficiency for a given amount of heat supplied over one cycle.
Focal knowledge, skills, and abilities (KSAs)	KSA1: Identify the relationship between efficiency (e) and heat (Q), i.e., $e = \frac{Q_{\mathrm{H}} - Q_{\mathrm{C}}}{Q_{\mathrm{H}}}$ KSA2: Calculate the efficiency from the given data KSA3: Make a conclusion based on the calculated efficiency vs efficiency given
Evidence Statements (ES)	ES1: Correctly identify the relationship between efficiency and heat ES2: Calculation of the efficiency of the engine ES3: Identify if the calculated efficiency is reasonable given the context of the task
Task Features	Question gives a scientific question, claim, or hypothesis to be investigated. Question gives a representation of the data (e.g., table or graph, or list of observations) provided to answer the question or test the claim or hypothesis. Question gives an analysis of the data or asks student to analyze the data. Question asks student to interpret the results or assess the validity of the conclusions in the context of the scientific question, claim, or hypothesis.

Table A.6: *KiU table for the Equations of State task where we explicitly show the Learning Performance, KSAs, ES, and Task Teatures*

Aspects of Knowledge in Use	**Description**
Learning Performance	Construct an explanation on how the physical dynamics of gasses determines if a scenario can be described with the ideal gas law
Focal knowledge, skills, and abilities (KSAs)	KSA1: Describe the mechanisms, or properties of a gas in the scenario KSA2: Make a claim based on the properties of gasses KSA3: Make a claim relating the ideal gas law to the properties of gasses
Evidence Statements (ES)	ES1: Statements about the processes/information describing the properties of ideal gasses ES2: Statements unpacking the relationship of the ideal gas law to their answer ES3: Claim about which scenario is correct regarding the ideal gas law
Task Features	Question gives an event, observation, or phenomenon. Question gives or asks student to make a claim based on the given event, observation, or phenomenon. Question asks student to provide scientific principles or evidence in the form of data or observations to support the claim. Question asks student to provide reasoning about why the scientific principles or evidence support the claim. Question provides at most two of the following: 1) a cause, 2) an effect, and 3) the mechanism that links the cause and effect, and the student is asked to provide the other(s)