Upper-Division Thermal Physics Assessment Development and the Impacts of Race & Gender on STEM Participation

by

Katherine Diane Rainey

B.S., Boise State University, 2014

M.S., University of Colorado Boulder, 2017

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Physics

2021

Committee Members:

Bethany R. Wilcox, Chair

Noah Finkelstein

Steven Pollock

Loren Hough

Jenny Knight

Rainey, Katherine Diane (Ph.D., Physics)

Upper-Division Thermal Physics Assessment Development and the Impacts of Race & Gender on STEM Participation

Thesis directed by Dr. Bethany R. Wilcox

Development and use of validated assessments in physics provide insight into student understanding of physics concepts and can be utilized to track student learning across time and instructional strategies. These tools can be used to gauge the efficacy of interventions designed to support students and persistence, and allow instructors to make data-based decisions about how they structure their classrooms to support student learning. Unfortunately, gaps in performance based on demographic groups can often appear in physics assessments. However, some statistical approaches can allow for identification of bias in assessment items, allowing for potential to reduce these "performance gaps." Additionally, studying how student identity influences their experiences can inform how these gaps are understood if they remain after bias within items has been addressed.

This dissertation reports two distinct but complementary studies. The first study discusses development of a novel upper-division thermal physics assessment composed of coupled, multiple-response items. This study includes assessment item development and refinement; assessment implementation; and assessment validation, with an explicit focus on differential item functioning and item response theory in addition to classical test theory approaches. The second, qualitative study focuses on how students' experiences in STEM vary based on race and gender. This includes investigations of perceptions of professor care and instruction styles; sense of belonging; and student perceptions of how race and gender impact those pursuing STEM. These studies in combination can inform practices intended to support students of all backgrounds in pursuing STEM and physics degrees.

Acknowledgements

The work presented in this dissertation was funded by NSF grants DUE-2013332, DUE-0969286, and DRL-1420363, and the Center for STEM Learning at the University of Colorado Boulder. This work would not have been possible without my collaborators and advisors. Bethany Wilcox and Michael Vignal were essential in completion of the work on thermal physics assessment. I am especially thankful to Dr. Wilcox, as she mentored me after I took a year off from graduate school, encouraged me to do more than I thought I was capable of, and supported me without judgement. To the Roots of STEM Success Project team—especially Melissa Dancy, Elizabeth Stearns, Roslyn Mickelson, and Stephanie Moller—I thank you for the opportunity to work on such an impactful project and your mentorship throughout the duration of my time working with you all. I am very thankful to Dr. Dancy, who supported and encouraged me when I was struggling to decide my career path, and whose guidance in research has led me to becoming the researcher I am today.

I want to thank the entire PER group at CU for their consistent, thoughtful, and effective feedback that has contributed to refinement and completion of this work. To Noah Finkelstein, thank you for your support throughout my time at CU; you provided many opportunities to explore when I could not decide my graduate school path, and I am extremely grateful. I would also like to thank my defense committee. Steven Pollock served on several of my committees throughout my time in graduate school, and his feedback and support did not go unnoticed. Loren Hough acted as a research advisor to me when I explored experimental physics prior to my deciding I belonged in PER; he was very supportive during and after my time in lab. Jenny Knight has provided helpful input and an outside perspective that has strengthened my work.

I am thankful to the organizations I have had the pleasure of engaging in throughout my time in graduate school, including CU-Prime, the Access Network, and the CU Teach Program. Engagement in these programs have helped shape the person, scholar, and educator I am today.

I am very grateful for all of my friends who supported me emotionally throughout my time in graduate school, including Gina, Katelyn, Amanda, Adam, Brianne, Simone, Jessica, Allie, and Holly. You have all been amazing cheerleaders and shoulders to lean on, and I couldn't have done this without you. I am also thankful to the mental health professionals who have helped me gain the confidence and skills I needed to press forward in pursuing my dreams. My cohort and roommates during my first year of graduate school and my study group members from Boise State deserve thanks as well. I would not be where I am academically without you all. I am also extremely grateful for the research opportunity Alex Punnoose gave me while I was an undergraduate; it was essential in getting me to this point.

I want to thank my family. To the Lupos, thank you for welcoming me into your family and being consistently supportive. To my dad, thank you for your support, love, and regularly checking in on me. To my mom, thank you for your continuous support and love over the many years I have been in school, and for leading by example as a strong, dedicated, and determined woman. To my niece, Kyleigh, thank you for serving as a reminder to not give up. And to my sister, Orianna, I thank you for always being there for me. Through every up and down over the course of my life, you have been there to support me. I love you and couldn't have made it without you.

Lastly, I want to thank my dear cat Rags and my loving partner, Brian. Rags has been a consistent source of comfort and love during my time in graduate school, and Brian has been my rock. Brian has been there through every high and low, every success and failure, and every difficult decision over the past four years. His never-ending love, support, and encouragement got me to the finish line. I am so very thankful to him.

Contents

Tables

Table

Figures

Figure

Chapter 1

Introduction

Validated assessments provide useful insight into student understanding of physics concepts. These assessments are often used to investigate the efficacy of instructional strategies, which can inform instructional changes, interventions, and curriculum development to improve students learning. Several introductory physics assessments exist, such as the *Force Concept Inventory* (FCI) [1] and the *Force and Motion Conceptual Evaluation* (FMCE) [2]. Upper-division assessments, such as the *Colorado Upper Division Electrostatics Diagnostic - Coupled Multiple Response* (CUE-CMR) [3], are becoming increasingly popular. Despite thorough validation processes for the majority of available physics assessments [4], performance differences (or "gaps") between men and women have been detected on several, including the FCI [5, 6]. Use of various techniques can allow for minimizing these gaps by identifying which items contribute most to their appearance. Complementary insight into understanding these gaps can be provided from investigating the influences of identity (e.g., racial and gender identity) on students' experiences while pursuing STEM degrees. This allows for better understanding of other, non-academic factors that can influence performance.

Despite the widespread presence of assessments in the domains of forces, motion, and electromagnetism at both introductory and upper-division levels, limited assessments are available for thermal physics at the upper-division level. In particular, to date there are no *upper-division* thermal physics assessments, nor any assessments that address both statistical mechanics *and* classical thermodynamics. Thermal physics, which includes both thermodynamics and statistical mechanics, is a core course required for attaining a physics bachelor's degree at most institutions. However, a

shortage of validated assessments in the realm of upper-division thermal physics presents challenges in measuring student understanding of this content to inform course transformations. In order to improve course instruction and student outcomes, researchers and instructors must first have some method of evaluating what students know and are able to do with that knowledge.

Developing an upper-division thermal physics assessment is a challenging task. One factor influencing this is the varying nature of content-foci across different thermal physics courses. Additionally, an ideal upper-division thermal physics assessment would capture student *reasoning* in addition to knowledge of the answer to multiple-choice questions. Though free-response (FR) assessments can capture rich reasoning patterns, they are harder to score in a streamlined way. One way to bypass the need for FR items while still soliciting student reasoning is to utilize coupled, multiple-response (CMR) items, which capture reasoning through mutliple-response questions [7]. To date, we know of one CMR-based, upper-division, content-based physics assessment—the CUE-CMR [3]. However, CMR-based assessments to address lab skills, such as the Physics Lab Inventory of Critical Thinking (PLIC) [8] and Modeling Assessment for Physics Laboratory Experiments (MAPLE) [9] are also available or under development.

Most validated physics assessments are analyzed using classical test theory (CTT) approaches. These analyses are sufficient for the format and uses of assessments thus far, but have limitations. To address these limitations, more robust techniques such as item response theory (IRT) can be employed. IRT approaches, particularly the one-parameter logistical (Rasch) model, offer advantages over traditional CTT analyses. In particular, the mathematics of the Rasch model allow for invariant scales for both student performance and item difficulty measures, making comparisons across different administrations more reliable. This is due to the reliance on sample population for CTT, which is less of a factor for Rasch analyses. This has important implications for analyses of differential performance based on demographic groups, as most assessments are piloted in large departments and physics is predominantly White and male [10].

Analysis of gender- and race-based performance gaps are rarely done *during* the development process of physics assessments. These types of analyses can aid in the identification of bias in assessment items. Biased items are questions characterized by a factor other than ability (such as gender or race) affecting how likely one is to answer a question correctly [5]. Appearance of bias in assessment items is problematic because it can mislead one in their conclusions or weaken inferences that can be made from their results. A common way to identify possible bias in items is differential item functioning (DIF). Some researchers have investigated the FCI using DIF or similar techniques [5, 6], which has resulted in identification of several items that disadvantage women. Some work has suggested revising prompts on the FCI to include more familiar contexts can address bias and its resulting performance differences (e.g., ref [11]). This implies bias can be addressed or minimized during item development if it is identified early.

Even when bias in test items is addressed, performance differences across race and gender may still appear. Analysis of course environments may be one lens to view these differences through, as active-learning classrooms have been shown to improve student learning in STEM courses as measured by similar assessments [12]. Improving instructional methods may be one way to address performance differences in STEM classrooms, as some have suggested [13].

Even if performance differences remain after accounting for instructional styles, instead of concluding some races or gender are more/less capable, one can view these differences through the lens of identity-based experiences. In addition to evidence of performance differences on assessment, there is also a large body of work evidencing people of different races and gender encounter different experiences in STEM due to their demographic identities. For example, women and students of color consistently report a lower sense of belonging in college and STEM compared to men and White students [14, 15, 16], which has been linked to overall course performance and persistence [16, 17, 18]. Despite the large body of literature documenting racialized and gendered experiences, little work has investigated whether students are aware of the impact of race and gender in STEM.

This dissertation presents two studies—one related to thermal physics assessment and one related to race- and gender-based experiences in STEM. The work in the latter part of the disser-

tation can inform the ways in which performance differences on validated assessments in physics are interpreted. Additionally, this work emphasizes the importance of considering race and gender during assessment development, such that bias can be addressed early on and minimize problematic takeaways from performance differences.

Part 1 of this dissertation (Ch. 2-Ch. 5) describes the development of an upper-division thermal physics assessment, and begins with a presentation of relevant background information (Ch. 2). Assessment item development is described in detail (Ch. 3), emphasizing relatively new techniques used, such as facilitating a wide range of input on content coverage and incorporation of race- and gender-based analyses to identify potential bias in items early on. Preliminary validation methods are discussed (Ch. 4), including CTT analyses along with novel Rasch analyses applied to CMR items, an avenue of validation that has yet to be deeply explored for CMR items. Part 1 concludes with a summary of presented work and discussion of future work (Ch. 5). Part 2 of this dissertation (Ch. 6-Ch. 9) focuses on various factors contributing to retention of underrepresented groups in STEM, particularly women and some people of color. After a presentation of background literature (Ch. 6), analyses are presented, which include perceptions of STEM course environments (Sec. 7.1); perceptions of professor care (Sec. 7.2); student sense of belonging in STEM (Sec. 7.3); and perceptions of race and gender impacts in STEM (Ch. 8). Part 2 concludes with a summary of presented work and discussion of implications (Ch. 9). The dissertation concludes with a discussion of the two presented projects, including their connections and implications (Ch. 10).

These studies in combination inform the ways in which assessments are developed and how their results can be interpreted, even if performance differences do appear between groups. These studies inform best practices for developing and validating upper-division CMR-based assessments, and add to the body of literature regarding influences of race and gender on STEM experiences. Work presented in Ch. 3 was published in Physical Review PER [19]. Work presented in Ch. 7 and Ch. 8 was published in the International Journal of STEM Education [20, 21, 22].

Chapter 2

Background: Upper-Division Thermal Physics Assessment

Part 1 of this dissertation describes the process of developing an upper-division thermal physics assessment, presenting the process of item development for the assessment (Ch. 3) as well as validation of the assessment (Ch. 4). In this chapter, relevant background from the literature is presented, including motivation for development of the assessment (Sec. 2.1); literature regarding thermal physics content understanding of students (Sec. 2.2.1); and the current status of thermal physics assessment (Sec. 2.2.2). Additionally, this chapter includes an introduction to coupled, multiple-response items (Sec. 2.3) and a summary of validation approaches used for this study (Sec. 2.4). A portion of the work presented in Part 1 was published in Physical Review PER [19].

2.1 Motivation

To date, there currently exists no *upper-division* thermal physics assessment, and filling this gap is no easy task. Creating an assessment that can provide data to guide instructors is a challenge for upper-division thermal physics because, anecdotally, the material covered in upper-division thermal physics courses often varies between instructors and across institutions. For example, some instructors teach thermodynamics, others teach statistical mechanics, and still others teach a combination of both (which will be referred to as "thermal physics" here). Additionally, some topical areas may not be prioritized by every instructor teaching thermal physics (e.g., diffusion or temperature-entropy diagrams). This content variability, in addition to notational differences across thermal physics texts, poses a significant challenge in the development of standardized thermal

physics assessments and teaching tools that can be utilized by a wide range of instructors. Therefore, an essential first step in creating a thermal physics assessment was to determine what thermal physics content is taught at most institutions, as discussed in Sec. 3.1. This process informed the development of items, as did current knowledge of student content understanding in thermal physics, which is discussed in the next section.

The process taken for development of the assessment discussed in Part 1 is unique in 3 ways. First, it has taken into account the diverse perspectives of *over seventy* instructors (at various types of institutions) with respect to content priorities in their upper-division thermodynamics and/or statistical mechanics courses. Second, it involves novel efforts to create coupled, multiple-response items [7] for upper-division thermal physics. Third, the assessment development process described will lay the foundation for an eventual flexible assessment that addresses both content and scientific practices [23], and also allows for customization of included content and practices. This flexible assessment will be developed as part of future work.

2.2 Literature Review

Assessments that address students' conceptual understanding of physics play an important role for both physics educators and physics education researchers. They can be used by educators to measure the impact of instructional approaches or curricular changes. For example, the *Force Concept Inventory* (FCI) [1] has been used for multiple introductory physics investigations since its initial development [24]. In physics education research (PER), assessments such as these are often an outcome of investigations into student reasoning and used to investigate student understanding of physics concepts.

Education research studies on thermal physics content, in the realms of biology, chemistry, engineering, and physics, are becoming increasingly present in the literature [25, 26]. These studies have utilized isolated thermal physics problems, small quizzes, and larger-scale assessments to investigate student conceptual understanding of primarily introductory thermal physics topics. Investigations of upper-division thermal physics content are comparatively less common than work

at the introductory-level. Additionally, there is currently no upper-division physics assessment that includes both thermodynamics and statistical mechanics. Existing thermal physics assessments instead focus on classical thermodynamics topics, such as heat, temperature, and thermodynamic laws [27, 28, 29, 30, 31]. At the time of this writing, searches have not resulted in evidence of any statistical mechanics assessments. A lack of a fully-encompassing, upper-division thermal physics assessment leaves a gap in the ability of professors and researchers to test upper-division students' understanding of higher-level thermal physics concepts (e.g., statistical mechanics) and facilitate further investigations of student understanding at the upper-division level. This section presents a brief review of research on student difficulties in thermal physics and the status of assessment in physics, highlighting thermal physics assessment.

2.2.1 Student Understanding of Thermal Physics Content

Many studies regarding thermal physics have focused on student understanding of thermodynamics content in physics, biology, and chemistry [25]. Most studies related to content understanding focus on concepts such as application of the ideal gas law; distinguishing between heat and temperature; heat, work, and the first law of thermodynamics; and entropy and the second law of thermodynamics.

Thermodynamics phenomena are often focused on gas systems. Studies revolving around student understanding and application of the ideal gas law in various contexts are common (e.g., refs. [32, 33]), while problem-solving with non-ideal gas systems is relatively less frequent in the literature.[1] On the other hand, student alternate conceptions[2] of heat and temperature is a very common theme throughout thermal physics literature and emerges at all levels, from K-12 through college-level courses [35, 36, 37]. One source of confusion in relation to heat and temperature is the colloquial use of the two terms, which are often used interchangeably [36, 38]. Students often

[1] A search for studies on student problem-solving with non-ideal gas systems yielded limited results in the field of physics education research.

[2] The term "alternate conceptions" is used in lieu of "misconceptions" or "misunderstandings" to align with recent literature and to avoid deficit language. Recent work has suggested use of the term "misconception" is at odds with the way students learn [34].

consider heat as a property of a system [37], like temperature, as opposed to a process quantity [38].

Much like heat, work is also commonly thought of by students as a state function as opposed to a process-dependent quantity [38, 39]. These conceptions can cause particular challenges when reasoning about net work and net heat for cycles; the view of work and heat as state functions leads to students reasoning that each of these quantities are zero for cyclic processes [40]. Additionally, it has also been found that many students tend to not recognize the utility of the first law of thermodynamics when considering heat, work, and changes in internal energy for processes [41].

Investigations of student conceptions revolving around heat and cycles have also addressed entropy [42], a core concept for thermal physics that is often unfamiliar to students entering thermal physics classes [43, 44]. Entropy can be viewed in terms of both classical thermodynamics and statistical mechanics, and it can sometimes be challenging for students to bridge these two frameworks. For example, students may make different predictions about entropy changes when applying microscopic and macroscopic views of entropy to similar phenomena [45]. Entropy has historically been a challenging topic for thermal physics students, often being considered as a conserved quantity [44, 46] or a measure of chaos or disorder[3] [47, 48]. Misapplication of the second law of thermodynamics is also very common, including assertions that entropy must increase in all contexts [44] or confusions about whether to apply the second law to the *system* or the *universe* [43].

Most studies in the realm of thermal physics content understanding have investigated student conceptual understanding in the realm of introductory thermodynamics content, mainly at the high school and introductory-college level [25]. Upper-division investigations are comparatively less common. One study by Meltzer compared student use of diagrammatic representation (i.e., PV-diagrams), notation, and mathematical equations, as well as verbal explanations, and found several alternate conceptions students leave introductory thermal physics with continue with them to when

[3] There are several well-founded critiques of presenting entropy as disorder to students, citing an unclear definition of what is meant by "disorder" (e.g., chaos, randomness, etc.) [47], as well as concerns about students' use of the term without provision of their own definition of what is meant by *disorder* [48]. Loverude *et al.* have also found that some students struggle to reconcile the idea of entropy as disorder when reasoning about phenomena such as approaching thermal equilibrium, a process by which entropy increases but the system reaches a more natural, and thus what they see as more *ordered*, state [46]. Several have suggested moving away from presentation of entropy as disorder [47, 49, 50].

they start more advanced thermal physics courses, and sometimes persist after instruction [51].

Compared to thermodynamics, statistical mechanics studies are relatively rare. Some studies include investigations about student reasoning surrounding the Boltzmann factor [52] and Taylor expansions in statistical mechanics [53], while some have looked into students' bridging of conceptions of the macroscopic and microscopic to study consistency between explanations of entropy changes based in statistical mechanics and classical thermodynamics [45]. Others have considered instructional strategies, such as using statistical approaches to teach entropy [54].

Content understanding and student reasoning investigations are often conducted through interviews, reviews of student coursework, or conceptual assessments, while others have utilized tutorials. In particular, Smith *et al.* created tutorials to investigate student reasoning using the Boltzmann factor to compare relative probabilities of states [52] and to address students' conceptions of entropy when approaching Carnot cycles [42]. A subset of questions from these tutorials was used to inform item development for the assessment, as described in Ch. 3.

2.2.2 Assessment in Physics

Though some lower-division thermal physics assessments exist, there is still a need for a standardized upper-division thermal physics assessment that includes both thermodynamics and statistical mechanics concepts. Currently, there are six assessments categorized as "thermal/statistical" assessments on PhysPort, an online repository of PER-based resources for physics faculty.[4] However, all of these assessments are categorized as being for "intro college" or "intermediate" levels (i.e., not upper-division) and, based on the four assessments that are readily available online [27, 29, 30, 31], none cover statistical mechanics concepts. It is of note that only five of these six thermal physics assessments were finalized and the sixth was recommended to not be used by the assessment developer. A request to the developer to view one of the five assessments did not recieve a response and thus it could not be accessed.

[4] A link to PhysPort is not presented, as it is susceptible to change. PhysPort can be found via an online search engine. The full webpage title is "PhysPort: Supporting physics teaching with research-based resources."

A literature search also resulted in another thermal physics assessment – the Thermodynamic Diagnostic Test (TDT) – which addresses student understanding of "three fundamental laws of thermodynamics" (the zeroth, first, and second laws of thermodynamics) [55]. The TDT is not available on PhysPort, but its questions are available in ref. [55]. The TDT is a two-tiered test (composed of coupled, multiple-choice questions) designed using the Thermal Concept Survey (TCS) as a baseline to start from.

All four assessments accessible through PhysPort[5] were developed based on research on student thinking, were studied and implemented at multiple institutions, and resulted in peer-reviewed journal articles and a dissertation [27, 28, 29, 30, 31]. Development of the TDT also utilized student interviews to construct distractors [55].

The development cycle of the assessments available on PhysPort is not discussed in most publications associated with them. However, Brown's dissertation [31], as well as ref. [56], outline the development of the most recent thermal physics assessment on PhysPort – the Survey of Thermodynamic Processes and First and Second Laws (STPFaSL), an assessment that was referenced during the item development process of the assessment presented in this dissertation. The process the STPFaSL developers took aligns with the historical approach taken for developing other assessments, included identifying focus topics (via course learning goals) and iterative question development and refinement based upon expert feedback, student interviews, and classroom testing [57]. This method closely matches the approach taken in developing the assessment focused on in this dissertation.

Most PER-based assessments, including existing thermal physics assessments, have intentionally narrow scopes and hone in on a very specific subset of topics within a particular sub-discipline of physics. Some of the very first conceptual physics assessments, such as the FCI [1], have done this, and other assessments have followed suit. The narrow scopes of these assessments makes them more useful for a broad range of practitioners who need to identify specific conceptual difficulties to

[5] The accessible assessments are the Heat and Temperature Conceptual Evaluation (HTCE), the Survey of Thermodynamic Processes and First and Second Laws (STPFaSL), the Thermal Concept Evaluation (TCE), and the Thermodynamic Concept Survey (TCS).

inspire pedagogical changes. Additionally, narrow scopes can be motivated by assessment validation restrictions, as discussed in Sec. 2.4. The scope of existing thermal physics assessments includes concepts such as specific thermodynamic laws and processes [31, 55], as well as basic thermodynamic concepts such as heat and temperature [27, 28, 29], phase transformations [29], and thermal properties of materials [29, 30]. The scopes of currently accessible thermal physics assessments do not include statistical mechanics concepts.

2.3 Coupled, Multiple-Response Items

Most assessments utilized in PER are multiple-choice (MC) or free-response formats (see PhysPort). One multiple-choice thermal physics assessment, the TDT [55], is a two-tiered test with each item being composed of two coupled MC questions—one prompting a response to a question and the other prompting reasoning used to achieve the first answer. The final version of the assessment discussed in Part 1 takes on a different format than these more traditional assessments, though it does have some commonalities with the two-tiered nature of the TDT.

The assessment discussed here is largely composed of coupled, multiple-response (CMR) items in addition to (relatively few) MC items. CMR items are composed of an MC question followed with a *multiple-response* prompt asking students to select one or more reasoning elements utilized to find the first answer [7]. Unlike two-tier and traditional MC questions, CMR formats allow for scoring based on both consistency and accuracy of responses, and allow for partial credit as opposed to being solely dichotomous in their scoring.

CMR formats are ideal for this assessment because they provide insight into student reasoning (much like free-response items) while also allowing for online administration, more streamlined scoring that can be automated, and partial credit. Some of the items in the assessment discussed in Part 1 are composed of a series of CMR questions. For brevity, Part 1 refers to MC and CMR items collectively as multiple-response (MR) items when appropriate.

2.4 Validation Methods

From the beginning of the assessment development process, multiple possible validation approaches for the assessment were considered. The two considered approaches—classical test theory and item response theory (IRT)—are discussed here and in more depth in Ch. 4. One of these (IRT) can provide more rigorous analyses to identify bias within assessment items, which results in differential performance between individuals of the same ability but different backgrounds (e.g., gender or race) [58].

With an eye towards reducing bias in the assessment, the development process began with taking diverse perspectives into account early on. For example, faculty perspectives were solicited from a diverse range of institutions and departments, many of which may often be excluded from large-scale assessment studies (see Sec. 3.1). This is due to the fact that these studies typically require large student populations and thus target large research institutions, which tend to be predominately White. Inclusion of multiple faculty perspectives allows expansion of the scope of the assessment while still making it broadly applicable for many institutions and instructors.

2.4.1 Classical Test Theory

As mentioned in Sec. 2.2.2, narrow content scopes within assessments can provide more pinpointed insights into student difficulties in a specific content area. However, the need for specific content areas to target in pedagogical reform is not the only motivation for these limited scopes. The narrow scopes within assessments can also be attributed to content variability across institutions (especially at the upper-division level) and the restrictions of many classical test theory (CTT) validation requirements, such as unidimensionality—the idea that an assessment's scope must focus on one construct (e.g., focusing on "forces" or "motion," as opposed to "introductory physics").

CTT is a commonly used approach for validation of conceptual assessments in education, and was utilized for validation of the assessment presented here to align with the methods of similar assessments in physics. One advantage of CTT that has made it so commonly used is its theoretical

assumptions, which make it easily applicable to different testing situations. Despite its advantages, CTT also has some drawbacks. The CTT requirement of unidimensionality limits the scope of content that can be included on an assessment. Thermal physics spans a large space ranging from classical thermodynamics to statistical mechanics, and thus a thermal physics assessment that captures both of these areas may not be unidimensional enough, as it would test more than one core topic or idea. In addition to unidimensionality, there is a more fundamental issue at the heart of CTT that has led to problematic outcomes in scores between different groups of students. This issue comes from fundamental assumptions of CTT: the validation statistics, and therefore the reliability and validity, are dependent on the group of students whose results were used in the initial validation [59]. This population-dependence of validation statistics makes it difficult to compare assessment performance between different groups of students.

Results of some CTT-validated assessments have found "achievement gaps," particularly between women and men [13, 60, 61, 62]. In these studies, men tend to perform better on these assessments than women. This could lead an unwary reader to conclude that men are more capable of performing well in physics than women. However, due to the demographic composition of physics majors, which is predominately male (and White) [10], it is likely that these assessments were implemented in predominately White and male departments. This means that the group-dependent statistics resulting from CTT use are normalized to a group of mostly White men. It may well be that the explanation for these "achievement gaps" is *not* the result of inherent differences between genders (such as one gender being more capable of doing physics than another), but instead are a result of the population-dependent nature of the approach used for validation or bias in the items.

2.4.2 Item Response Theory

CTT-based differential item functioning (DIF) can help address differential performance and bias. However, to address potential bias and better address variations of scores across different demographic groups, researchers can also invoke different validation methods, such as item response theory (IRT). Note a historically used term in IRT literature is the term "student ability,"

which refers to the underlying latent trait the statistical models are attempting to quantify. Fundamentally, however, it is a measure of performance as opposed to innate ability of individuals. This historically utilized term is used throughout Part 1 for consistency with the existing literature. However, it worth acknowledging the potentially problematic nature of this term, particularly when it comes to the appearance and intepretation of performance gaps between subgroups of students (e.g., men and women).

The primary IRT model used in this study is the one-parameter logistical model, also known as the Rasch model. The mathematics of the Rasch model separate parameters describing student ability from the difficulty of items. The Rasch model assumes student performance on an assessment item is based solely on their ability and the item difficulty [63]. This is a mathematical outcome of assuming equal discriminations for each item (i.e., each item is equally good at distinguishing between high- and low-achievers). A result of this model is that these parameters, to an extent,[6] are *not* dependent on the population used to validate the assessment, so long as the validation pool is reasonably broad within the targeted group (i.e., upper-division physics students). This makes comparisons across institutions susceptible to less ambiguity.

The Rasch model was initially created for analysis of dichotomously scored items (i.e., items scored as either correct or incorrect). Thus, this model cannot be directly applied to items in which partial credit is possible, such as for CMR items. To analyze items such as these, an additional IRT method is considered here: the partial credit model (PCM). Equal discriminations can also be assumed with the PCM, thus resulting in similar benefits to Rasch in reducing potential bias in items. The PCM produces similar, but also distinctly different, outputs as the Rasch model; the main difference being the difficulty parameters, as discussed in more detail in Sec. 4.7.4.

Ch. 4 presents initial exploration of the Rasch model and PCM for validation statistics. More robust statistical analyses with the PCM method will require a larger set of students than was available for this study (N$\approx$100 is not sufficient). Analyses will also require a diverse (and larger) set

[6] Note the parameters are in some ways still dependent on population (e.g., only upper-division thermal physics students are included in the sample).

of students in order to make claims about performance differences across different populations and try to reduce test bias. Neither of these requirements were met during the initial validation process for the assessment.[7] However, preliminary PCM analyses are presented to propose a potential method for applying the PCM to CMR items, something that has yet to be done.

2.5 Developing the Assessment: An Overview

The remainder of Part 1 of this dissertation describes the process of developing (Ch. 3) and validating (Ch. 4) an upper-division thermal physics assessment, concluding with a summary and discussion of future work (Ch. 5). Ch. 3 begins by describing the construction and distribution of a survey to physics faculty, at both large and small departments as well as minority-serving institutions and women's colleges, to probe content priorities in upper-division thermal physics courses. Results of this survey were used for developing assessment items, as described in Sec. 3.2. Ch. 4 describes validation of the assessment, including CTT and preliminary IRT analyses and more in-depth discussions about these methods and their outputs.

[7] PCM analyses were not initially a goal of the validation process. However, if future pilot tests allow for larger datasets to be collected, more robust analyses could be possible.

Chapter 3

Developing an Upper-Division Thermal Physics Assessment: The U-STEP

With potential validation methods in mind, the process of developing the assessment began by first probing the physics community about content priorities in upper-division thermal physics courses, followed by developing items informed by collected information. This chapter presents the development process of the Upper-level Statistical Mechanics and Thermodynamics Evaluation for Physics (U-STEP), including development and distribution of a faculty content survey (Sec. 3.1); development of assessment objectives (Sec. 3.2.1); and development of items (Sec. 3.2.2). The chapter concludes with presentation of the cycle for developing two assessment items—one focusing on classical thermodynamics (Sec. 3.3.1) and the other on statistical mechanics (Sec. 3.3.2). A majority of the work presented in in this chapter was published in a first-author paper in Physical Review PER [19].

3.1 The Faculty Content Survey

It is commonly suggested that content coverage in upper-division courses can vary significantly between different physics departments and even different instructors within the same department. This poses a significant challenge in developing an assessment that will be useful for a broad range of instructors and courses. For example, if an assessment contains even a single question related to content not covered in an instructor's course, the assessment will likely be less useful for that instructor or not be used by that instructor at all. There are two strategies for addressing this issue. One is to create an assessment that can be customized to different courses, allowing topics

present on the assessment to be tailored for specific courses. The other is to create an assessment that only touches on topics covered by the majority of instructors. Development of the U-STEP began with attempting the latter, with the aim of working toward a customizable assessment in the future.

In an effort to inform content foci for assessment development, a content survey to be completed by physics faculty familiar with teaching upper-division thermal physics was developed. The survey was distributed to a diverse range of institutions and designed to solicit key information about thermal physics courses, such as content covered, general course structure and emphasis (thermodynamics, statistical mechanics, or both), and needs or interest in an upper-division thermal physics assessment. This section describes methods for developing and distributing the survey with an emphasis on creating a format that was accessible and relatively short in duration, while still soliciting sufficient information (Sec. 3.1.1), and presents results from the survey, including texts utilized and content foci in upper-division thermal physics courses (Sec. 3.1.3).

3.1.1 Survey Development

Prior to constructing the faculty survey, a focus group was conducted with four experts, all with experience teaching thermal physics and/or researching student difficulties in thermal physics. The focus group solicited expert perspectives surrounding upper-division thermal physics, including textbooks, content coverage, learning goals, and existing thermodynamics assessments. Outcomes from the focus group informed several questions included on the faculty survey. For example, participants discussed notational conventions as one major challenge for a thermal physics assessment (e.g., the sign convention of work). To address this concern, one question on the survey solicited specific notational issues worth considering in development of a thermal physics assessment. Additionally, textbooks brought up during the focus group comprised the list of textbook options provided on the survey.

To facilitate ease of responses, the faculty survey was a primarily multiple-response format with only a select set of questions being free-response. Thus, one of the first steps in survey

development was determining which options to provide for various multiple-response questions. This began with an investigation of the scope of thermal physics in texts; six thermal physics texts brought up during the focus group [64, 65, 66, 67, 68, 69] were analyzed for key content coverage. This process involved reviewing each text and identifying topical areas for each based on chapter titles, section headings, and emphasized key terms. Based on the frequency of topics appearing across the different texts, topical areas were classified into *core topics* and *supporting topics*. To put these into an accessible form for use in the survey, topics were sorted and condensed into 29 core topics, most with roughly 4 supporting topics (see Table 3.1 and Appendix A). For example, the core topic of "thermodynamic laws" had four supporting topics: 0th law, 1st law, 2nd law, 3rd law. Some core topics had no supporting topics (e.g., semiconductors) while some had as many as seven (e.g., energy and thermodynamic potentials); the one exception to this was statistical mechanics, which had 14 supporting topics.

In addition to focusing on concepts, the survey also solicited information on the scientific practices valued by respondents in their thermal physics courses.[1] The list of scientific practices provided on the survey was pulled from the Next Generation Science Standards (NGSS) list of science and engineering practices [23]. This list was developed in collaboration with a team of practicing scientists in a variety of STEM disciplines [70]. Though the list of practices was developed for K-12 purposes, one can argue they are still applicable at the upper-division level because they are meant to encapsulate practices that scientists engage in while doing authentic science. In the NGSS list of scientific practices, similar practices are combined together (e.g., developing and using models); however, in upper-division courses, it is less clear that all paired practices would be targeted together. Thus, to collect more specific data about individual practices, paired NGSS practices were split into separate categories. For example, "developing and using models" was split into "developing models" and "using models" for the survey.

[1] Scientific practices are gaining more focus in PER college-level assessment and will be integrated into a future iteration of the assessment presented here.

The survey was administered online[2] and divided into 4 major sections: (1) general course information; (2) content coverage; (3) scientific practices; and (4) interest in, and concerns about, an upper-division thermal physics assessment. In addition to providing general course information in Section 1, respondents also had the option to identify their institution and submit their course syllabus. Finally, gender and racial identity information were collected at the end of the survey, after Section 4.

After initial construction of the survey, feedback was solicited from physics faculty at the University of Colorado Boulder who were familiar with teaching upper-division thermal physics. Based on these discussions, and informed by the frequency of topical areas appearing across the six different analyzed texts, we grouped the core topics into two categories: assumed core topics and other core topics. Assumed core topics are topics that one might expect are covered in every thermal physics course. In alphabetical order, the assumed core topics were:

- Energy and Thermodynamic Potentials
- Engines and Refrigerators
- Entropy
- Equilibrium
- Monatomic Gases
- Heat
- Temperature
- Thermodynamic Laws
- Work

The survey presented these assumed core topics at the beginning of Section 2 of the survey, with their supporting topics shown on the same page. A free-response textbox followed these assumptions to allow respondents to indicate disagreement with the assumptions made. All other core topics were provided on the following page of the survey without their supporting topics

[2] The survey was administered through the survey platform Qualtrics and hosted by the University of Colorado Boulder (CU).

displayed. After selecting from the list of other core topics, associated supporting topics for each of the selected core topics were displayed on the following page (while all other supporting topics associated with core topics not selected were not displayed). This conditional formatting was motivated by the desire to reduce respondent fatigue due to survey length.

3.1.2 Survey Distribution

To ensure the course data was reflective of a broad range of institutions, contact information collected for survey distribution included a large variety of physics degree-granting institutions, inclusive of minority serving institutions (MSIs) and women's colleges. Institutions were identified using the American Physical Society's "Top Educators" lists [71], each of which identifies 16-20 institutions with the highest average number of physics bachelors' degrees awarded by the institution per year, sorted by highest physics degree offered at each institution. We utilized the overall and underrepresented minority (URM) lists for PhD-granting, MS-granting, and BS-granting institutions. Beyond that, the American Physical Society's MSIs list [72] was used to identify all other physics-degree-granting MSIs not on the "Top Educators" lists; the MSI list included institutions with both large and small physics departments. This list identifies Historically Black Colleges and Universities, Black-serving institutions, Hispanic-serving institutions, and institutions serving Asian Americans and Native American Pacific Islanders that offer degrees in physics. Women's colleges were identified with the "Women in Physics" report produced by the American Institute of Physics [73].

After identifying institutions, contact information of department chairs was obtained from physics department websites. The survey solicitation was then sent to the department chairs, with a specific request for the email to be forwarded to all faculty within their department who were currently teaching or had previously taught upper-division thermal physics. In addition to department chairs, the research team solicited the help of their professional contacts at different institutions to take the survey or forward it to faculty in their department.

After initial solicitations were sent and data were collected, a large geographic region had not

been captured by the data. This prompted a process of identifying institutions in geographic areas within the United States that were not captured by the Top Educators, MSI, or women's colleges lists using the PhysNet directory of US Physics Institutions [74]. The survey was re-opened and a second round of emailed solicitations were sent to newly identified departments. In this second round of solicitations, 18 new responses were collected, largely from states that were not represented initially.

3.1.3 Survey Results

The survey was open for response collection for three and a half months, then opened again a few months later for an additional month and a half. During these time windows, 73 respondents fully completed the survey while 2 completed all of the survey except questions regarding scientific practices (Section 3 of the survey) and interest in assessment (Section 4). Only responses that completed the sections with core topics and supporting topics (Section 2) or beyond were used for analysis. Response rate for the survey is not reported, as it is unclear how many people received the solicitation forwarded from their department chairs. An estimated 200-250 emailed solicitations were sent.

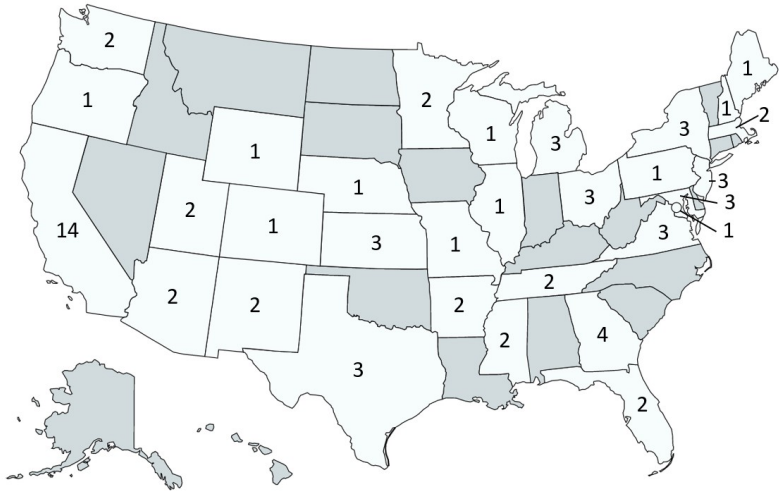


Figure 3.1: Number of survey responses per state. Darker gray states had no respondents. One additional response was received from an international location not displayed on the map.

Figure 3.1 shows locations of responding institutions; many geographic regions across the United States are represented in the sample, with the exception of the north central region. Overall, 63 unique institutions were identified from the survey, 32 of which were MSIs and/or women's colleges; 2 institutions could not be identified. Figure 3.2 presents institution type by highest physics degree offered and MSI/women's college classification. In a few cases, (N=7) institutions were represented by 2-3 responses; it was evident from submitted syllabi and individual item responses that these were submitted by different people.

Racial demographics of respondents included Asian (14%, N=10), Black/African American (1%, N=1), Caucasian (76%, N=55), and Hispanic (4%, N=3); no other racial identities were indicated and 6% (N=4) preferred not to answer. Additionally, 85% (N=61) of respondents were men and 13% (N=9) were women (no other gender identities were indicated); 3% (N=2) preferred not to provide their gender. Three additional respondents did not provide any demographic information.

Institutional information, including selectivity, research activity, student population, and highest physics degree offered was collected via the Carnegie Classifications [75] and institutions' physics department websites. From the Carnegie Classifications, 73% (N=45) of identifiable insti-

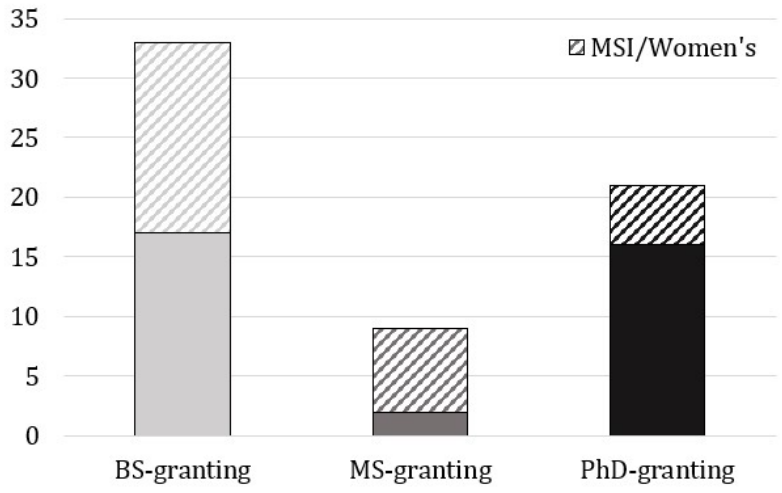


Figure 3.2: Highest physics degree offered by institution classification. Minority-Serving Institutions (MSIs) or Women's Colleges are indicated in striped boxes and other institutions are presented in solid colors. Institutions granting Bachelor's degrees (BS), Master's degrees (MS), and PhDs are indicated.

tutions are considered being selective or more selective with regards to admissions practices, while 27% (N=17) are considered "inclusive" institutions. Additionally, 54 schools are classified as having high or very high research activity.[3]

3.1.3.1 Course Information

The survey asked respondents if their course focused on thermodynamics, statistical mechanics, or both (thermal physics); 95% (N=71) selected thermal physics and the remaining 5% (N=4) of responses were split evenly between thermodynamics and statistical mechanics. Most institutions reported offering one semester of thermal physics (80%, N=59); some reported two quarters (7%, N=5) or two semesters (8%, N=6), while a small minority reported one quarter (4%, N=3). Respondents reported that their courses were composed of mostly juniors (N=56) and seniors (N=49), though some (N=12) reported sophomores in the course as well. Respondents could choose multiple student populations for this portion of the survey.

The majority of respondents (69%, N=52) reported using *An Introduction to Thermal Physics* by Daniel V. Schroeder [68]. *Thermal Physics* by Charles Kittel and Herbert Kroemer [66] was the second most frequently cited text (15%, N=11). All other texts appeared at a frequency of 7% or below. Figure 3.3 shows frequencies of most common textbooks; 5% reported using no textbook and 27% reported using other texts not provided on the survey, though none of these other texts appeared at frequencies over 5%. Many of the instructors (40%, N=30) indicated they teach with the assumption that their students have no prior exposure to thermal physics content. Some expected familiarity with topics such as energy, heat, the first and second laws of thermodynamics, and the ideal gas law. A few (N=11) said they expect thermal physics exposure from the introductory physics sequence, though several noted that thermal physics is only covered for a few weeks, and sometimes not at all, in that sequence.

Overall, these data show most institutions require one semester of thermal physics, most instructors use Schroeder's text [68], and many instructors assume their students have no prior

[3] Note one institution was not in the Carnegie Classification database, as it was not an institution in the US.

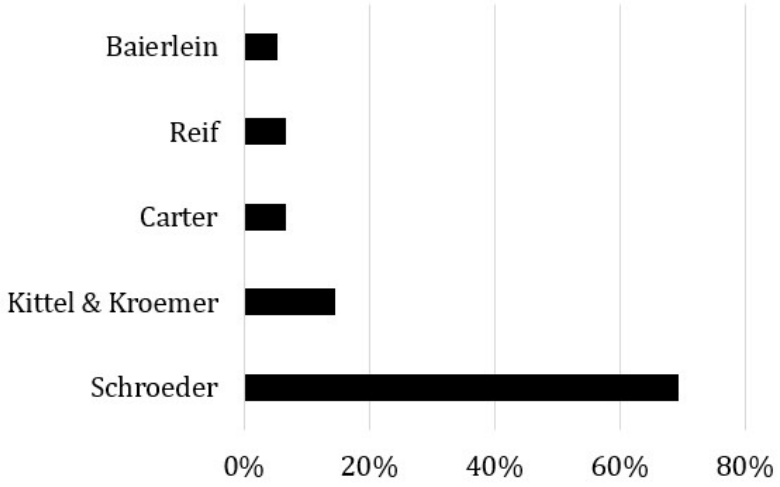


Figure 3.3: Most common textbooks reported being used in upper-division thermal physics courses [68, 66, 65, 76, 64]. Respondents could select multiple textbooks on the survey or indicate using another textbook not on the provided list or none at all. Thus the percentages do not sum to 100%.

exposure to thermal physics content. This assumption about lack of exposure to thermal physics may be particularly relevant for concepts such as entropy and statistical mechanics, which are concepts unique to thermal physics and rarely included in other physics courses. These results suggest two implications for physics education research: (1) development of Schroeder-based thermal physics assessments and materials could serve many instructors and institutions, though would still exclude the sizable population of instructors and institutions who do not use that text; and (2) pretest administration of upper-division thermal physics assessments may not produce meaningful measurements of student understanding of thermal physics content prior to taking the course due to unfamiliarity with content and/or jargon. Based on implication (2), we chose to create the U-STEP with the intention of it not being administered in a pretest format.

3.1.3.2 Key Topical Areas

Table 3.1 shows frequency of assumed core topics, assumed core topics' supporting topics and, other core topics. Most assumed core topics (see Sec. 3.1.1) appeared at a frequency of 100%. In two responses, no supporting topics were selected under one of the nine assumed core topics.

Table 3.1: Topic frequency from faculty content survey. The two left columns show data for assumed topics. All assumed core topics appeared at a frequency of 100% with the exception of engines and refrigerators. Frequencies for engine and refrigerator supporting topics are calculated using the total number who selected that core topic (N=71). The right-most column shows all other core topics. Frequencies above 95% appear in **bold**. See Appendix A for other core topics frequencies. Topics that appeared on syllabi but not the survey (e.g. ensembles and thermodynamic identities) are also not presented.

Assumed Core (& *Supporting*) Topics	%	Assumed Core (& *Supporting*) Topics	%	Other Core Topics	%
Energy & Thermodynamic Potentials	**100**	Equilibrium	**100**	Statistical Mechanics	**97**
Chemical Potential	**95**	*Thermal Equilibrium*	**99**	Processes	91
Energy Sources	47	*Stable & Unstable Equilibrium*	44	Fermions	84
Enthalpy	88	Heat	**100**	Bosons	83
Equipartition	**96**	*Heat Capacity*	**100**	Blackbody Radiation	83
Free Energy (Gibbs & Helmholtz)	**96**	*Heat Transfer*	72	Phases	80
Internal Energy	**97**	*Latent Heat*	91	Diatomic Gases	79
Maxwell's Relations	80	Temperature	**100**	Quantum Phenomenon	75
Engines & Refrigerators	**95**	*Absolute Zero*	**97**	Kinetic Theory	72
Heat Engines	**97**	*Negative Temperature*	65	Pressure Diagrams	72
Refrigerators	87	*Thermodynamic Temperature*	88	Scaling	69
Entropy	**100**	*Temperature Measurement*	57	Magnetism	65
Boltzmann's Law	92	Thermodynamic Laws	**100**	Chemical Reactions	55
dS=dQ/T	**95**	*0th*	89	Solids	55
Entropy & Information	56	*1st*	**100**	Conduction, Convection, Radiation	51
TS Diagrams	71	*2nd*	**100**	Pure Substances	47
Gases	**100**	*3rd*	88	Diffusion	44
Ideal Gas Law	**100**	Work	**100**	Cooling Techniques	29
Mixtures of Gases	60	*Mechanical*	**99**	Fluids	20
van der Waals Interactions	71	*Path dependence*	80	Semiconductors	11

However, in neither of these instances was it indicated in the provided textbox that the respondent disagreed with the assumption that they cover that topic. The one assumed core topic that did not appear at 100% frequency (engines and refrigerators) was selected by only 95% of respondents.

Table 3.1 shows the frequency of supporting topics relative to the number of times the corresponding core topic was selected, and the frequency of core topics relative to the total number of valid responses (i.e., responses completed through Section 2 of the survey). Frequencies of all other core topics is also presented. See Appendix A for presentation of the 56 other supporting topics and frequencies.

All 9 of the assumed core topics, as well as statistical mechanics, were found to be covered by at least 95% of respondents in their upper-division thermal physics courses. These results are relevant for all researchers interested in materials and assessment development in upper-division thermal physics. For the purpose of developing the U-STEP, these results were used to prioritize content-foci for assessment item development. Using these results as a baseline helps in creating an assessment that can serve a wide range of instructors and institutions.

3.1.3.3 Scientific Practices

Of the 16 practices presented on the survey, three appeared at a frequency of over 85%: using mathematical thinking (99%, N=72), asking questions (97%, N=71), and using models (89%, N=65). Review of syllabi indicates the practice of "asking questions" may have been misinterpreted by survey respondents; the NGSS practice refers to asking scientific questions (namely for scientific investigations), but the research team suspects many respondents may have interpreted this practice as referring to asking questions about content during class or office hours. The next most frequently appearing practice was constructing explanations (71%, N=52), while the remaining 12 practices appeared at a frequency of 64% or less, as can be seen in Table 3.2.

These results highlight four scientific practices that stand out as valued by a large majority thermal physics instructors in this sample: using mathematical thinking, asking questions, using models, and constructing explanations. The data demonstrate many other scientific practices are

Table 3.2: Frequencies of scientific practices valued by physics faculty in upper-division thermal physics. Frequencies were determined using the number of respondents who completed through Section 3 of the faculty content survey (N=73). The practices were taken from the NGSS list of science and engineering practices [70].

Scientific Practice	%	Scientific Practice	%	Scientific Practice	%
Using Mathematical Thinking	99	Interpreting Data	59	Developing Models	45
Asking Questions	97	Defining Problems	58	Obtaining Information	40
Using Models	89	Using Computational Thinking	58	Designing Solutions	38
Constructing Explanations	71	Analyzing Data	51	Carrying Out Investigations	12
Communicating Information	64	Engaging in Argument	47	Planning Investigations	10
Evaluating Information	62	from Evidence			

less of a universal focus for thermal physics courses at the upper-division level. This has implications for others who may want to integrate scientific practices into development of materials (e.g., assessments or tutorials) for upper-division thermal physics courses.

3.1.3.4 Assessment Concerns & Interest

Beyond collecting information about upper-division thermal physics courses, the survey also included questions directly related to the assessment. In particular, we asked about potential notational issues they could see as worthy of consideration in the assessment design and faculty interest in the assessment, including how they would use it and if they would be interested in piloting the multiple-response assessment in their upper-division thermal physics courses.

Notational Concerns. Of all respondents, 29 faculty provided possible notational concerns on the survey. The survey question about notational concerns gave the sign convention of work as an example, and 13 respondents indicated in their response to this open-ended question that the sign of work would be a concern for assessment development; this was the most common concern cited. Others mentioned commonly used symbols (e.g., k vs. k_B); the sign of work's impact on the form of the first law of thermodynamics (e.g., $\Delta U=Q+W$ vs. $\Delta U=Q-W$); the sign convention of heat for certain applications (e.g., engines); and units (e.g., some texts use unitless quantities

whereas others do not). Informed by these notational concerns, and knowing which texts are used the most frequently, notation used in the U-STEP was largely based on notation used in Schroeder's text [68]; when possible, use of notation was avoided completely when developing assessment items. Additionally, all symbols are defined directly within assessment items or at the beginning of the assessment. Potential confusion about the sign of work and heat was avoided by asking which direction work made energy flow or which direction heat flowed (i.e., into or out of the system). Additionally, when applicable, the magnitude of work was solicited when looking for symbolic or numeric responses related to work.

Faculty Interest in Assessment. The survey included one question regarding faculty interest in the assessment, specifically asking if and how faculty would use the assessment once completed. The most common use cited by respondents was using the assessment to measure student learning in their course (N=13), often with pre-post administration (N=11). Others said they'd be interested in using it to inform their teaching strategies (N=12) and compare student learning in their course with other courses (N=11). A few (N=4), said they would use the assessment to inform content or tracking student progress. Of the 43 respondents who answered this question, none indicated they did not see obvious value in the assessment. However, in interpreting respondents' sense of value, it is important to note this group was composed of a self-selected set of faculty that chose to complete the survey.

3.1.3.5 Response Consistency

As a verification of the survey data, survey responses and submitted syllabi were checked for consistency for the 51 responses that provided a syllabus. Key topics on each syllabi were compared with the associated survey response to ensure topics appearing on the syllabus also appeared on the survey response. Here, "discrepancies" refers to a mismatch between topics on the syllabus and those selected on the survey (e.g., topics appear on syllabus but were not selected on the survey). The topic of diffusion was listed by five respondents in their syllabus and not their survey responses. No other core or supporting topics had more than 3 discrepancies when comparing between survey

responses and the syllabus.

Discrepancies could be due to the amount of focus placed on those topics in the course. For example, Bose-Einstein condensates may appear on the syllabus but may not be seen as a major content focus for the instructor when completing the survey, resulting in a discrepancy between their syllabus and response. Some topics, such as interacting systems (N=12), ensembles (N=13), large systems (N=16), and Boltzmann and/or quantum statistics (N=17), appeared in syllabi but did not appear as explicitly named core or supporting topics on the survey. However, those who included topics such as these on their syllabus selected other topics on the survey that encompass or require the same idea, such as multiplicity, thermal equilibrium, and specific statistical distributions. This comparison shows that the survey reliably captured the scope of content coverage for most survey responses without large discrepancies.

3.1.3.6 Content Variability

To investigate the claim of content variation across upper-division thermal physics courses, survey responses were examined to see how many topics were selected by all instructors. In particular, the focus was on the three groups of topics laid out in Table 3.1: assumed core topics, assumed core topics' supporting topics, and other core topics.

Analysis showed that 9/9 (100%) of assumed core topics, 11/32 (34%) of assumed supporting topics, and 2/20 (10%) of other core topics were selected by at least 90% of respondents. When looking at the 100% threshold, we found that 8/9 (89%) of assumed core topics, 4/32 (13%) of assumed supporting topics, and 0/20 (0%) of other core topics were selected by all respondents. These results suggests that there is little alignment (outside of the assumed core topics) across instructors and institutions. This lack of alignment holds true even within an institution, though to a lesser extent: at institutions with multiple survey respondents, an average of 73% of assumed supporting topics and 54% of other core topics were chosen by all respondents at that institution.

These results support the anecdotal claim that upper-division thermal physics content coverage varies both across institutions and between instructors at the same institution. It also makes

the case, however, that there are some topics, namely our assumed core topics, that all or most instructors prioritize in their upper-division thermal physics courses.

3.2 Item Development: The Process

Evidence of content variability motivates the need for a flexible-format assessment in which instructors are able to customize assessment content for their particular class, a focus area for future work. A first step in this process is to create a core set of items that could act as a baseline assessment for every upper-division thermal physics course. After analysis of survey results, items were developed to target key topics identified on the survey, focusing on 10 topical areas chosen by most instructors; this allows for the items to serve a wide range of upper-division thermal physics courses.

Initial item development took place in two broad stages: (1) developing assessment objectives for each of the 10 identified topical areas (as described in Sec. 3.2.1);and (2) writing and refining assessment items based on those objectives (Sec. 3.2.2). The full development process is summarized in Figure 3.4 and aligns with processes used for developing similar upper-division assessments [57]. The free-response items were piloted once and the multiple-response items were piloted twice, with the intention of a third pilot before finalization.

3.2.1 Assessment Objectives

To guide the writing of our assessment items, assessment objectives were developed. These are intentionally called *assessment objectives* (AOs), as opposed to learning objectives, because they were specifically designed to guide development of assessment items. It is not intended for the written objectives to span the full scope of objectives that may appear in upper-division thermodynamics, statistical mechanics, or thermal physics courses. Instead, they span the space of feasible, testable outcomes for the specific assessment developed.

Two criteria were set for these AOs: (1) they must collectively span the space of content areas identified as important based on faculty responses on the survey; and (2) they must be directly

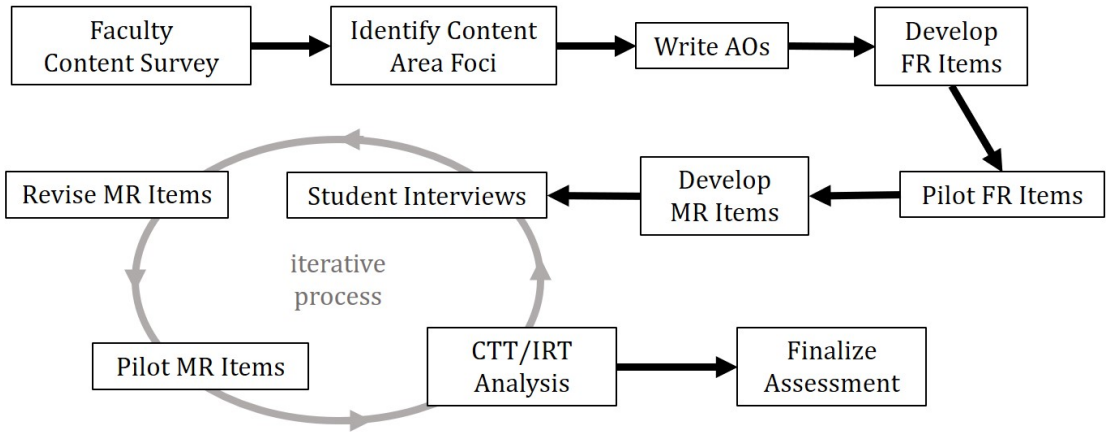


Figure 3.4: The full development process of the U-STEP. After the content survey, key content areas are identified. Free-response (FR) items are then written, revised iteratively, then piloted. Using results of the FR pilot administration, multiple-response (MR), composed of both multiple-choice (MC) and coupled, multiple-response (CMR), items are developed. MR items are then tested for face validity in student interviews. After interviews, MR items are revised, and then piloted. After initial MR piloting, preliminary validation statistics are completed, including classical test theory (CTT) and item response theory (IRT). Then, MR item development is revisited, involving student interviews and revision. Items are then piloted again. This iterative process concinutes until the assessment is finalized.

assessable in a MR format. When developing the objectives, we began by identifying concepts and turned these into actionable statements that could be assessed. For example, a concept would be something like "probability" or "Boltzmann factor," whereas an AO would be "students can determine probabilities for thermodynamic systems using the Boltzmann factor." Here, the AO meets both criteria by expressing an assessable action students should be able to do with their content knowledge.

First, AOs were developed for core topics and supporting topics appearing with frequencies above 90% on the faculty content survey. Some topics that were similar but appeared in different areas on the survey, such as the 2nd law and entropy, were combined into a single topical area. In the end, we focused on developing AOs for the following topical areas (in alphabetical order): energy, engines and refrigerators, entropy and the 2nd law, equilibrium, the 1st law, gases, heat, statistical mechanics, temperature, and work.

We identified high-frequency supporting topics for each of the topical areas from the survey. We then identified concepts within each category, based upon these supporting topics. For example, a concept for the supporting topic of *equipartition* for the core topic of *energy* was *degrees of freedom*. From these topical areas, we created initial AOs. These initial drafts of objectives were iterated and refined. After satisfaction with the AO drafts was reached, the AOs were distributed to external reviewers. We received feedback on the AOs from by two of these external reviewers and then finalized the AOs based on reviewer comments. For example, a finalized AO that targeted equipartition and degrees of freedom was: "Students can articulate that the internal energy of an ideal gas is determined by the number of degrees of freedom available to atoms/molecules comprising the gas system."

Overall, 57 AOs were drafted across the 10 identified topical areas, some of which had supporting objectives as well. Statistical mechanics had the most AOs (eleven). Some AOs overlapped across topical areas, but remained in the topical area that it more closely addressed. For example, some 1st law AOs contained both work and heat but stayed within the "1st law" topical area because they addressed heat and work only within the context of 1st law, as opposed to in isolation. Crossover such as this was documented for future reference. All finalized AOs can be found in Appendix B.

3.2.2 Assessment Items

AOs were used to directly inform assessment item drafting. All AOs were addressed by at least one, and often more than one, drafted assessment item. In total, 86 potential assessment items were drafted. Many items were inspired by problems from Schroeder's text [68], research-based tutorials (e.g., ref. [52]), and common questions asked of student in the upper-division thermal physics course taught at the University of Colorado Boulder. All initial items were free-response or multiple-choice with a prompt to explain reasoning. This was motivated by the need for attractive distractors to be included on the finalized MR assessment.

After drafting potential assessment items, the focus was narrowed on a subset of items that

addressed the widest range of AOs while resulting in a feasible number of items that we could commit to revising. The most essential AOs from the pool of 57 written objectives were identified, ensuring the shortlist spanned the 10 key topical areas. We then identified 24 items that addressed the majority of these objectives. The next step was to pilot these items so they could be turned into MR items for the final assessment.

In developing pilot assessments for these 24 items, a balance was found between collecting enough information about individual items to adequately inform developing them into MR items and having the piloted assessment be short enough to complete in the given time period (50 minutes). The assessment items were split into four groups of six items; two "anchor sets" and two "secondary sets." Anchor questions were questions that were likely to be on the finalized assessment, while questions composing the secondary sets were questions with more uncertainty related to their inclusion in the finalized assessment (e.g., they may prove to be too easy or too difficult to include in the final version, or be difficult to transform into a MR format).

Four beta-assessment versions were developed, each containing one anchor set followed by one secondary set, making each beta-assessment 12 items long. For two of the beta-assessment versions, the order of questions in the secondary sets were reversed so that, if students answered questions in order but ran out of time, a sufficient number of responses to every item would still be collected. Each group of six questions was included on two versions of the beta-assessment (outlined in Table 3.3).

The beta-assessments were piloted as free response (FR), ungraded[4] post-tests in an upper-

[4] Students received participation credit for completion of the beta-assessment.

Table 3.3: Number of items for each version of the free response thermal physics beta-assessment.

	Anchor Questions	Secondary Set 1	Secondary Set 2
Version A	6 (set i)	6	-
Version B	6 (set ii)	-	6
Version C	6 (set i)	6 (reversed)	-
Version D	6 (set ii)	-	6 (reversed)

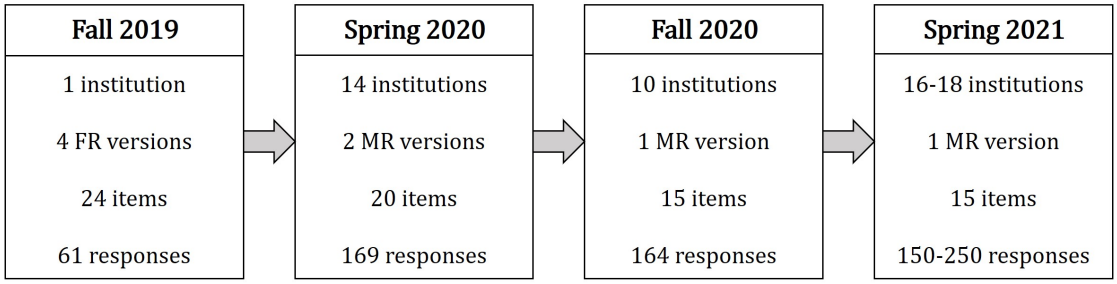


Figure 3.5: Information about each pilot administration of the U-STEP. The free-response (FR) items were piloted in 4 version in Fall 2019. Two mutliple-response (MR) versions were piloted in Spring 2020, and 1 MR version was piloted in Fall 2020. An additional MR administration is planned for Spring 2021, with an estimated 150-250 responses, dependent on response rate and number of participating institutions. Note all MR pilot semesters were during the COVID-19 pandemic and many institutions across North America were requiring online or remote instruction instead of allowing in-person teaching.

division thermal physics course in a large physics department in Fall 2019. Figure 3.5 presents a timeline for all pilot administrations of the assessment. The FR beta-assessments were administered in-class during a 50 minute period. Responses from 61 students were received (out of 67), the majority of whom self-identified as White (77%), men (70%), and seniors (87%). The completion times for students ranged between 18 and 50 minutes, taking an average of 31 minutes overall.

After piloting, student responses were analyzed and coded based on what directly emerged from what students wrote [77]. For example, in response to a question regarding the difference between heat and temperature, responses such as "heat is a measure of how much energy is transferred to or from a system" and "heat only exists as a transfer of energy" were both coded as "heat as energy transfer." This example code resulted in MR selections such as "heat is a flow of thermal energy" and "heat is a quantity exchanged between systems" for CMR items related to heat. For all FR items, responses and corresponding codes were used to inform the rewriting and revising of items to convert them into MC or CMR formats.

Once we reached MC and CMR formats, we conducted five think-aloud interviews with students to verify the items were being interpreted in the way they were intended to [78, 79, 80]. Small revisions were made based on the outcomes of interviews. For example, one question related to the equipartition theorem asked: *When does the equipartition theorem hold?* For the question,

students would select *any* situation that the equipartition theorem holds for (e.g., selecting "when the system is a solid" because the equipartition theorem *does* apply for Einstein solids) instead of *requirements* that must hold for the equipartition theorem to be applied, as the question originally intended. Thus, the prompt was changed to: *Which of the following* ***must*** *be true in order for the equipartition theorem to be used for a given system?*[5]

These MR items were piloted in two versions for the first time at 14 physics departments in an online format in Spring 2020, with approximately 248 students total being asked to complete the assessment by their instructor. After piloting, classical test theory (CTT) validation statistics were run, including estimations of item difficulty and discrimination, and we conducted detailed analyses of response patterns. Response patterns were analyzed for each item to see if any correct selections were not selected frequently, if any distractors were selected very frequently, or if strange response patterns appeared.

Additionally, CTT differential item functioning (DIF) analyses were conducted to identify any performance differences on items between different genders and races. This is an essential aspect of item and assessment *development*, and has been relatively absent in the development of other physics assessments. The importance of DIF analyses come from their facilitation of identifying performance differences on items between students of different populations with similar abilities, which could be the result of bias within items. Identification of these performance differences allow the potential to curb bias in items and the assessment while it is being constructed. Any items in our pool that displayed problematic performance differences were investigated further when developing the assessment.

Using the results from the various analyses discussed above, some items prompts and MR options were revised while others were dropped. An additional four think-aloud interviews were conducted with students with the newly revised set of items. From here, the MR assessment was piloted in one version at 10 physics departments, with approximately 227 students being asked to complete the assessment by their instructor. After the Fall 2020 pilot administration,

[5] After response analyses and CTT investigations, it was decided to not include this item in the final U-STEP.

similar analyses as done for the Spring 2020 pilot were conducted, resulting in a small number of revisions. Revised items were tested in 3 think-aloud interviews. No revisions were made after those interviews.

3.3 Item Development: Two Examples

This section outlines the full item development process, from AOs to the finalized CMR version, for two assessment items—one addressing classical thermodynamics (Sec. 3.3.1) and one addressing statistical mechanics (Sec. 3.3.2).

3.3.1 An Example in Classical Thermodynamics

As discussed in Sec. 2.2.1, confusion surrounding heat and temperature is quite common for thermal physics students at all levels [35, 36, 37]. The final item presented in this section was designed to address common alternate conceptions about heat and temperature. The item was constructed from two different items addressing similar themes, and was eventually combined into a single item after analysis.

3.3.1.1 Developing the Assessment Objectives

The items described here addressed two key content areas identified from the faculty content survey—*heat* and *temperature*. Despite common connections between these content areas, heat and temperature were kept as distinct categories due to unique AOs applicable to only one of the two. For example, the AO *"students can articulate that objects in thermal equilibrium have the same temperature"* was unique to temperature and not applicable to heat. The heat AO addressed in the items discussed here is:

(1) *Students can articulate that heat is a flow of energy, caused by a temperature difference.*

   (a) *Students can articulate why heat has no meaning at a single state.*

      (i) *Students can articulate that since heat flow between two states depends on the*

path between those states, it is not a state variable.

This AO and its supporting AOs address the definition of heat as well as the important characteristic of heat as *not* being a state function, a common confusion for students [38]. The temperature AO addressed in the items discussed here is:

(1) *Students can articulate differences between heat and temperature.*

(a) *Students can articulate that temperature is a property of a system and heat is not.*

This AO and its supporting AO address the common alternate conception of heat being a property of a system [37]. One can see commonalities, but also distinctions, between these heat and temperature AOs. These similarities contributed to the combination of the two distinct items into the single finalized CMR item after initial piloting and analyses. To begin to address these AOs separately, two distinct items—one for heat and one for temperature—were developed, as described in the following section.

3.3.1.2 Developing the Beta-Assessment Items

After developing the AOs for heat and temperature presented above, two free-response (FR) items were developed. The heat prompt was: *"Heat is a property of a system, like temperature, pressure, and volume."* The temperature prompt was: *"Any object at a given temperature contains a certain amount of heat."* Both items asked students to decide if the statement was true or false, which was then followed by an open-response prompt asking for reasoning.

3.3.1.3 Piloting the Beta-Assessment Items

One of the items discussed in Sec. 3.3.1.2 was included in each beta-assessment version (heat: Versions A and D; temperature: Versions B and C), both within secondary item sets (see Table 3.3). Table 3.4 summarizes responses to the items, including percent responding "false" (the correct choice for each) and the percent answering "false" with appropriate reasoning. Reasoning justifying "false" for each item was similar, and most responses were along the lines of heat being a form of

Table 3.4: Percent of correct responses to the FR items adressing heat and temperature included in the Fall 2019 beta-assessment. A "correct" response refers to an answer of "false" for each item, and the "appropriate reasoning" refers to fully correct responses.

Target Concept	% Correct	% Appropriate Reasoning
Heat	93%	72%
Temperature	86%	52%

energy that flows, transfers, or is exchanged between systems.

3.3.1.4 Refining the Items: Coupled, Multiple-Response

After analyzing student responses to the FR items presented in Sec. 3.3.1.2, the items were transformed into CMR formats [7]. Student explanations for their true/false choice for each FR item were used to create *reasoning elements* for the CMR versions. Reasoning elements are provided to students in the CMR item as choices to support their answer to the MC prompt. The expectation for the CMR items is for students to select one response for the MC question (*true* or *false*), then select as many reasoning elements as needed in the follow-up question to support their choice. The research team also created additional reasoning elements to capture possible reasoning not expressed in students FR responses. This was done to ensure every possible MC response had associated reasoning elements, in part to decrease the possibility of students concluding the correct response through test-taking strategies and process of elimination. An "other" option was provided on these versions as well to solicit other students reasoning elements that do not appear in the provided list of options inspired by FR explanations during the initial beta-assessment pilot. This was done in part to offset the limitation of piloting the FR version at only one institution. The selected reasoning elements give insight into particular student alternate conceptions or understandings. Figure 3.6 presents the CMR versions of these items, which were both piloted in Spring 2020. The temperature item was piloted in both assessment versions, while the heat item appeared in only one. Both items underwent 5 students interviews in the Spring 2020 semester.

Heat

Consider the following statement:

Heat is a property of a system,
like temperature, pressure, and volume.

This statement is...

A. true
B. false

because... (*select all that support your response above*)

a) heat is a property that is determined by a system's temperature, pressure, and volume
b) heat is a property that can be changed by changing temperature, pressure, and volume
c) a system contains a finite amount of heat, like pressure or volume
d) heat is a quantity exchanged between systems
e) heat is a flow of thermal energy
f) heat is a scalar, like temperature, pressure, and volume
g) heat is not a state function
h) other: _____

Temperature

Consider the following statement:

Any object at a given temperature
contains a certain amount of heat.

This statement is...

A. true
B. false

because... (*select all that support your response above*)

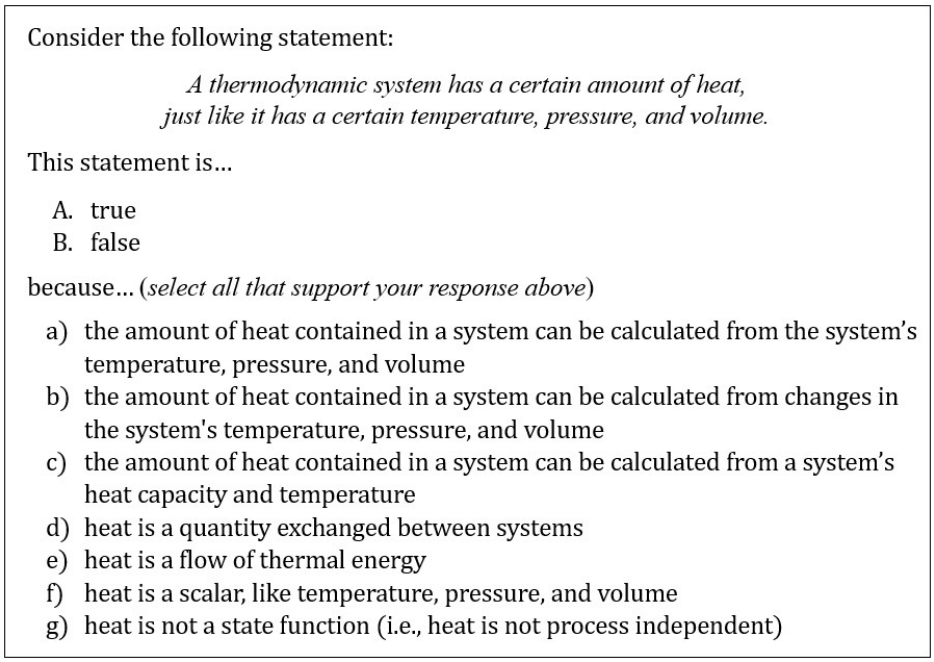
a) heat is a quantity exchanged between systems
b) it is impossible to know exactly how much heat an object contains
c) heat is a flow of thermal energy
d) heat and temperature are the same thing
e) heat is not a state function
f) the amount of heat in an object determines its temperature
g) it does not make sense to talk about heat as a quantity that can be contained
h) other: _____

Figure 3.6: CMR versions of the items addressing heat (left) and temperature (right), developed from responses to the FR versions.

3.3.1.5 Finalizing the Item: Coupled, Multiple-Response

DIF analyses based on gender and race yielded no significant results for the heat-targeted item, but some performance differences did appear for the temperature-related item. Due to this, in addition to similar reasoning elements in each item and similar response patterns across the two, the decision was made to combine the two items into a single CMR item. This was also motivated by the desire to avoid redundancy in items within the finalized assessment and because the AOs were similar enough that they could both be addressed within the same item.

The finalized CMR item addressing both heat and temperature is presented in Figure 3.7. The prompts from both initial CMR items were combined, such that the new prompt addressed both the idea of heat as a property of a system and distinguishing between heat and temperature (i.e., temperature is a property of a system and heat is not). A few changes were made to reasoning elements when initial items were combined. For example, both initial CMR items had a reasoning element of *"heat is not a state function."* During several interviews (out of 7), students aritculated


Consider the following statement:

A thermodynamic system has a certain amount of heat, just like it has a certain temperature, pressure, and volume.

This statement is...

A. true
B. false

because... (*select all that support your response above*)

a) the amount of heat contained in a system can be calculated from the system's temperature, pressure, and volume
b) the amount of heat contained in a system can be calculated from changes in the system's temperature, pressure, and volume
c) the amount of heat contained in a system can be calculated from a system's heat capacity and temperature
d) heat is a quantity exchanged between systems
e) heat is a flow of thermal energy
f) heat is a scalar, like temperature, pressure, and volume
g) heat is not a state function (i.e., heat is not process independent)

Figure 3.7: Finalized CMR version of a single item addressing heat and temperature, developed after piltoing individual CMR items addressing each content area in Spring 2020.

that they did not remember what the term *state function* meant, causing them to not select that option. This was determined to be problematic because (1) it is a correct response, and (2) the goal of the assessment is not to assess recall of terminology. Thus, in the final CMR version, this selection was changed to *"heat is not a state function (i.e., heat is not process independent)."* Presenting a definition of "state function" within the reasoning element eliminates the reliance on recall in order to answer the question correctly. The choice was made to add the definition of state function to each item in which it appeared for the finalized U-STEP.

3.3.2 An Example in Statistical Mechanics

A total of seven statistical mechanics questions were piloted across the four beta-assessment versions, with 3-4 statistical mechanics items appearing on each version. This section focuses on one of these items to demonstrate the full cycle of item development, from AOs to its finalized CMR version. It focuses on student understanding of Boltzmann factors and the partition function, and

was informed by a tutorial developed by Smith *et al.* [52].

3.3.2.1 Developing the Assessment Objectives

From the content survey, we identified several content goals for statistical mechanics. The example highlighted here covers the following content goals, which were prioritized based on responses to the faculty content survey: *probability*, *counting*, the *Boltzmann factor*, and the *partition function*. From these content goals, the following initial AOs were developed:

(1) *Students can determine probabilities.*

   (a) *Students can determine probabilities for simple systems using counting.*

   (b) *Students can determine probabilities for thermodynamic systems using the Boltzmann or Gibbs factor.*

(2) *Students can deduce state probability from a partition function.*

(3) *Students can compare probabilities of states.*

   (a) *Students can compare probabilities for simple systems using counting.*

   (b) *Students can determine and compare probabilities for a system after an energy shift using the partition function.*

These AOs all have action words (e.g., *determine*, *deduce*, and *compare*), much like those for heat and temperature (i.e., *articulate*), and could be directly assessed with a well-written assessment item. After these were finalized, they underwent review from two experts familiar with teaching and researching upper-division thermal physics. With reviewer feedback in mind, the AOs from above were modified and combined into the following singular AO:

(1) *Students can determine and compare probabilities of states.*

   (a) *Students can determine and compare probabilities of simple systems using counting.*

(b) *Students can determine and compare probabilities for a thermodynamic system using the Boltzmann factor.*

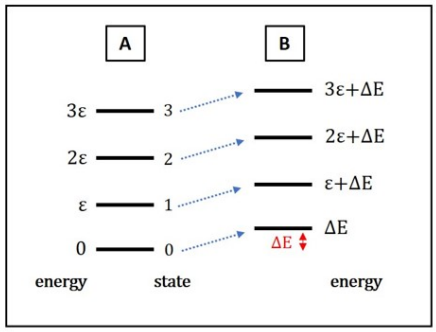
(c) *Students can determine and compare probabilities for a thermodynamic system after an energy shift using the Boltzmann factor.*

These assessment objectives still address the specific ideas they were intended to (i.e., the content goals) and contain actions the assessment could actually measure (i.e., *determine* and *compare*), but in a more straightforward and less redundant way.

3.3.2.2 Developing the Beta-Assessment Item

After articulating this AO, an item was drafted inspired by a problem developed for a Boltzmann factor tutorial [52], as shown in Figure 3.8. The item asks students about changes in the partition function and state probabilities after an atomic system undergoes an energy shift of ΔE

A particle is in system A, with the energy of each available state indicated on the left. The system undergoes a process such that the energy of the states is shifted by $\Delta E > 0$ (referred to as system B).



(i) How does the partition function change due to the energy shift?
- A. increase
- B. decrease
- C. remains the same

Explain your reasoning.

(ii) How does the probability of the particle being in state 2 change due to the energy shift?
- A. increase
- B. decrease
- C. remains the same

Explain your reasoning.

(iii) How does the ratio of probabilities of being in states 1 and 2 compare between A and B? That is, how does $P_A(1)/P_A(2)$ compare to $P_B(1)/P_B(2)$?
- A. less than
- B. greater than
- C. equal

Explain your reasoning.

Figure 3.8: A statistical mechanics free-response (FR) item piloted in the Fall 2019 beta-assessment. The item is considered to be FR despite its inclusion of multiple-choice options because of the prompt to explain reasoning.

that takes the system from *system A* to *system B*.[6] This item was designed to directly address parts (b) and (c) of the finalized AO discussed in the previous section.

3.3.2.3 Piloting the Beta-Assessment Item

The item shown in Figure 3.8 was included in Versions A and D of the beta-assessment (see Table 3.3) and 29 student responses were collected. Table 3.5 summarizes responses to the MC portions of the item presented in Figure 3.8. For (i), incorrect responses were split somewhat evenly between *increase*, *remains the same*, and a written *"I don't know"* (indicated in Table 3.5 as "unsure/no answer"). For (ii), *decrease* was a popular incorrect answer, provided by 31% of respondents. Justifications for *decrease* were often related the the negative exponent in the Boltzmann factor that appears in the numerator of the probability function. Here, students did not identify the identical change in the partition function due to the energy shift, which appears in the denominator. Most students answered (iii) correctly, but there was a variety of reasoning used to justify their responses. Some utilized similar reasoning from (i) and (ii), saying that everything changed by the same factor, whereas others relied on physical reasoning without invoking mathematics (e.g., there is the same spacing between the energy levels in both systems).

[6] As described in a later section, the wording of the problem was changed to ask "what happens when the system goes from configuration A to configuration B" to address potential confusion about the number of systems discussed in the problem (i.e., there is only one system, but it could be interpreted as two separate systems).

Table 3.5: Response frequency for the multiple-choice portion of the assessment item presented in Fig. 3.8. Bolded responses indicate the correct response.

	increase	decrease	remains the same	unsure/no answer
(i)	21%	**45%**	17%	17%
(ii)	14%	31%	**38%**	17%
	less than	greater than	equal to	unsure/no answer
(iii)	0%	17%	**72%**	10%

3.3.2.4 Refining the Item: Coupled, Multiple-Response

After analyzing student responses to the free-response (FR) item presented in Figure 3.8, the item was transformed into a CMR format [7]. One change was made to the prompt after the analysis: "system A" and "system B" were changed to "configuration A" and "configuration B." This was done to mitigate any confusion that may arise for students. Configuration B is the same *system* as configuration A, just after an energy shift; with the original wording, the two diagrams could potentially be interpreted by students as two independent systems.

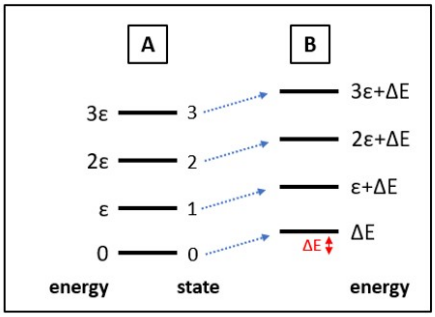
Much like the thermodynamics item discussed in Sec. 3.3.1, student responses to the *"explain your reasoning"* prompts for each portion of the FR item were used to create reasoning elements for the CMR version of the statistical mechanics item. The research team also created additional reasoning elements to capture possible reasoning not expressed in students FR responses.

3.3.2.5 Finalizing the Item: Coupled, Multiple-Response

Figure 3.9 shows the finalized CMR version of the item presented in Figure 3.8. The expectation for the CMR item would be for students to select one response for the MC question (*increase*, *decrease*, or *remains the same*), then select reasoning elements in the follow-up question that reflect reasoning they used to determine their answer. It is worth noting that the lists of reasoning elements provided in the item presented in Fig. 3.9 are relatively long and one may hypothesize the length of these lists exhaust students or influence response patterns. However, research shows that long lists of reasoning elements in CMR items do not significantly change response patterns or validation statistics when comparing CMR items to the FR versions used for development of those items [7, 81].

As an example of possible insight provided from reasoning elements, if a student incorrectly said the probability decreased and selected "the Boltzmann factor decreased due to the energy shift" but did not select anything related to the partition function or the relationship between the partition function and Boltzmann factor in determining probability, this would indicate the

A particle is in configuration A, with the energy of each available state indicated on the left. The system undergoes a process such that the energy of the states is shifted by $\Delta E > 0$ (referred to as configuration B). (See figure.)



How does the partition function change due to the energy shift?

(i) The partition function...
- A. increases
- B. decreases
- C. remains the same

because... *(select all that support your response above)*
- a) there are the same number of states before and after the energy shift
- b) the partition function does not depend on the value of the energy level
- c) each term in the partition function gains an $e^{\Delta E/k_B T}$ factor
- d) each term in the partition function gains an $e^{-\Delta E/k_B T}$ factor
- e) the states in configuration B would have more energy than the states in configuration A
- f) a particle in configuration B would have more energy than a particle in configuration A

(ii) How does the probability of the particle being in state 2, $\mathcal{P}(2)$, change due to the energy shift?

$\mathcal{P}(2)$...
- A. increases
- B. decreases
- C. remains the same

because... *(select all that support your response above)*
- a) the particle is more excited
- b) the Boltzmann factor increased due to the energy shift
- c) the Boltzmann factor decreased due to the energy shift
- d) the Boltzmann factor doesn't depend on energy
- e) the partition function increased due to the energy shift
- f) the partition function decreased due to the energy shift
- g) the partition function doesn't depend on energy
- h) the Boltzmann factor and partition function changed by the same factor
- i) the Boltzmann factor changed by a greater factor than the partition function
- j) the partition function changed by a greater factor than the Boltzmann factor
- k) there are the same number of possible states before and after the energy shift

How does the ratio of probabilities of being in states 1 and 2 compare between A and B? That is, how does $\mathcal{P}_A(1)/\mathcal{P}_A(2)$ compare to $\mathcal{P}_B(1)/\mathcal{P}_B(2)$?

(iii) The ratio $\mathcal{P}_A(1)/\mathcal{P}_A(2)$ is...
- A. less than
- B. greater than
- C. equal to

$\mathcal{P}_B(1)/\mathcal{P}_B(2)$ because... *(select all that support your response above)*
- a) probability increases with increasing energy
- b) probability decreases with increasing energy
- c) state 2 is more probable in configuration A than configuration B
- d) state 2 is more probable in configuration B than configuration A
- e) each probability changes by the same factor going from configuration A to B
- f) the differences between energy levels are the same in both systems
- g) the ground state energy changes the same amount as other states' energies

Figure 3.9: A finalized statistical mechanics coupled, multiple-response item developed based on responses to the FR version of the same item piloted in the Fall 2019 beta-assessment (see Figure 3.8). Initial CMR versions were piloted twice (Spring and Fall 2020) before finalization.

student likely *does* know that probability is directly proportional to the Boltzmann factor, but didn't consider the partition function's contribution to the probability as well. Each possible MC option has some combination of reasoning elements provided in the follow-up that could support the MC selection, and students are able to form their own reasoning from the provided reasoning elements. An "other" option was provided on initial CMR versions as well, but is not included on the final version. After developing the CMR item, it was piloted in 13 think-aloud interviews before finalization to verify it was being interpreted as intended. This particular item did not change after interviews. No students indicated in the interviews that the lists of reasoning elements were too long for them to comprehend effectively. The CMR version was piloted in the Spring and Fall 2020 semesters. No changes were made to the item based on the CTT and DIF analyses, as no issues appeared in these analyses.

3.4 Item Development: Summary

The previous section presented illustrations of the item development process for two items included in the finalized U-STEP, which consisted of developing assessment objects from results of the faculty content survey; drafting and piloting of free-response items; and creating and finalizing coupled, multiple-response (CMR) items. In between each CMR pilot administration, validation analyses using classical test theory and interviews were performed to inform selection of items to be included in the final U-STEP and item revisions. These analyses, as well as preliminary item response theory analysis, is presented in detail in the following chapter.

Chapter 4

Preliminary Validation Strategies for the U-STEP

Item development and assessment validation were not separate processes. Instead, item development was interspersed with interviews and statistical tests to establish both qualitative and statistical validity. Additionally, approaches for establishing face validity began early in the development process. These aspects of item development are lightly touched on in Ch. 3, and are discussed in more depth in this chapter.

Items were piloted in several iterations throughout the development process, both at many institutions and through interviews, as described in Sec. 4.1 and Sec. 4.3. Between each CMR pilot, validation statistics via classical test theory were performed (Sec. 4.6), including an explicit focus on differential item functioning (Sec. 4.6.2). Results from these analyses informed revisions, refinement, and elimination of items for inclusion in the finalized U-STEP. Additionally, exploratory item response theory analyses were conducted (Sec. 4.7), to set a baseline for analysis of future iterations of the U-STEP and analyses of similarly structured CMR assessments to be developed in the future. The U-STEP is not yet finalized. An additional pilot administration of the assessment will be done in Spring 2021, which will allow for a higher-N sample for more robust validation analyses. This chapter presents preliminary work regarding validation analyses, particularly using the statistical approaches discussed in Sec. 4.6 and Sec. 4.7, and has yet to be published. This work will inform which validation approaches will be most appropriate when finalizing the U-STEP.

4.1 Pilot Administrations of the U-STEP

As mentioned in Sec. 3.2.2, the U-STEP was piloted in a free-response (FR) beta-assessment version in Fall 2019 and twice in a coupled, multiple-response (CMR) format in Spring 2020 and Fall 2020. The FR pilot was conducted in person at one institution. Piloting at only a single institution is a limitation, but we did not have access to more institutions for that administration and asking instructors to dedicate a full class period to the assessment is a lot more to ask than administering a CMR assessment outside of class time. Each CMR version was piloted in an online format. It is worth noting that both the Spring 2020 and Fall 2020 semesters were taught during the COVID-19 pandemic and thus many courses were taught remotely or online during those semesters. This had little impact on piloting the assessment, as the intention for the U-STEP was always for it to be administered in an online format. The remote-nature of courses used for piloting may have impacts on the use and interpretation of the U-STEP for in-person classes, but these types of impacts can be revisited and investigated once in-person instruction resumes.

Key piloting information is summarized in Table 4.1. The Spring 2020 assessment was piloted in two versions, each composed of 13 items. The first 6 items were identical on both versions, with the last 7 items differing for each. Due to the varying class sizes at our piloting sites, the number of institutions receiving each version is not equal. Instead, we distributed each version to a different number of insitutions such that the number of students receiving each version was roughly equal.

Table 4.1: Information about the Fall 2019, Spring 2020, and Fall 2020 pilot administrations of the U-STEP. The Fall 2019 assessment versions were free-response, while the Spring 2020 and Fall 2020 versions were multiple-response. The most popular text was *An Introduction to Thermal Physics* by Daniel V. Schroeder [68]. Note the average response rate does not include classes with 0% response rate (N=1 for Spring and Fall 2020).

	$N_{institutions}$	$N_{students}$ per class			$N_{students}$	Response Rate		Most
		avg.	min.	max.	total	overall	avg.	Common Text
Fall 2019	1	N/A			67	91%	N/A	Schroeder (N=1)
Spring 2020	14	18	2	90	248	73%	78%	Schroeder (N=12)
Fall 2020	10	23	8	86	227	75%	73%	Schroeder (N=7)

Thus, 5 distributions of one version were made (with N=125 students) and 9 distributions of the other version were made (with N=123 students). Unequal response rates for each institution lead to the first version receiving more responses than the second.[1]

The Fall 2020 assessment was piloted in one version and composed of 15 items; it was the first composition of items that will be included in the final U-STEP. Of the 10 piloting institutions, 9 received responses, contributing to a total of 170 responses on the piloted U-STEP. At the end of each piloted assessment and student interview, students were asked to provide demographic information.[2] This was used primarily for the analyses discussed in Sec. 4.6.2. Demographics for each pilot administration can be found in Appendix C.

4.2 Content Validity

Validity refers to a measure of whether a test or assessment measures what is says it does, in addition to the interpretations that can be inferred from produced scores [59]. *Content validity* refers to how well the assessment covers the targeted content domain [59]. Content validity is typically established early in the development process, while other types (see Sec. 4.3 and Sec. 4.5), are established after the assessment has been administered in some way. Content validity can be addressed by several routes, including expert input and review. The first route taken for establishing content validity for the U-STEP was to solicit faculty input from the beginning of item development, as established by the faculty content survey (see Sec. 3.1). This process ensured that the topics within the U-STEP would address key topics within the domain of thermal physics. Results from the survey were used to inform writing of assessment objectives (AOs), which were then used to guide item development. After completing a finalized draft of preliminary AOs, we provided our list to 7 independent reviewers with experience teaching and studying upper-division thermal physics prior to item development. We received responses from 2 reviewers, who provided feedback that

[1] Instructors are encouraged to give credit for participation; however sometimes the incentive isn't enough to get students to participate. COVID-19 and remote learning likely caused some burnout for students, contributing to low response rates. (Note Spring 2020 was the semester when the COVID-19 pandemic hit and many institutions had to switch to remote/online instruction styles mid-semester.)

[2] Students were given a "prefer not to answer" option for demographics questions.

aided in revisions and finalization of the AOs used to develop items.

4.3 Construct Validity

Construct validity is associated with the characteristics being measured by the assessments as well as interpretations of the results [59]. To establish construct validity, throughout item development and revision, student interviews were conducted (N=13). After CMR items were drafted, prior to being piloted in Spring 2020, 5 interviews were conducted in a pencil-and-paper format, 2 and 3 for each version.[3] Additionally, prior to the Fall 2020 pilot, 5 remote interviews were conducted via Zoom using a single CMR version. To facilitate these interviews, we utilized the online platform used for distributing the pilot assessment (i.e., Qualtrics) and Zoom screen-sharing. After the Fall 2020 pilot, 3 additional remote validation interviews were conducted. Thus, each item underwent 12-13 validation interviews before finalization.[4] Most interviews were conducted with men (N=10) and White students (N=10).[5] Interviewed students came from four different institutions, with 10 students from one institution and 1 student from each of the remaining three.

Through these interviews, it was verified that most item prompts and response options were interpreted as intended; when this was not the case, the items were revised to address students' comments made during the interviews. For example, one reasoning element in a question addressing engines (item 14) read: "the first law of thermodynamics." The intended meaning of this option was unclear for several students, prompting it to be changed to "the first law of thermodynamics relates W and Q to ΔU" (where all variables were defined in the preceeding prompt). Additionally, interviews were analyzed to ensure students' selections aligned with their ariticulated reasoning. In only a few instances[6] did students' choices not align with reasoning. Often this was due to reading

[3] Interviews happened the week before CU closed campus due to COVID-19. Thus, fewer interviews were conducted than originally intended.

[4] Two items (both focused on entropy) underwent an additional 4 interviews in a FR format, and are not included in this interview count. These items were initially developed as part of a class assignment for a graduate-level assessment development course.

[5] All interviewed students provided demographic information. Note that though 10 men and 10 White students completed interviews, these groups didn't fully overlap, and thus fewer than 10 White men were interviewed.

[6] "A few instances" refers to 1-2 errors per interview, with the number of these occurences only appearing in a small number of interviews.

the response options too quickly or not remembering the definition of a term (e.g., state function) or entity (e.g., the Boltzmann factor). A small number of revisions were made to items after the interviews; these revisions happened between each set of interviews (as opposed to after all 13 were conducted).

4.4 Scoring

Careful consideration was taken when developing the scoring scheme for the U-STEP. A key consideration was how many points to assign the multiple-choice (one correct response) questions and mutliple response (partial-credit possible, related to reasoning) questions. The research team collectively decided to assign up to 3 points for multiple-response (MR) prompts, and investigated the impacts of changing the weighting of the multiple-choice (MC) prompt (i.e., 2 points vs. 3 points). This ultimately was meant to address the question of whether MC response and reasoning should be of equal worth, or if reasoning should be weighted more heavily. Since reasoning is a key aspect of CMR items, we did not consider weighting the MC response more than that of the MR reasoning questions.

Weighting for MC responses was explored using data from the two versions of the assessment piloted in Spring 2020. Overall, weighting the MC portion more heavily did increase overall averages and item difficulties, namely because students tended to do better on the MC question than responding to reasoning prompts. However, we also observed the magnitude of performance gaps based on race and gender increased for some items when MC answers were weighted more heavily. Due to this reason, and the desire to value reasoning more heavily, the choice was made to weight MC questions as worth 2 points and reasoning worth 3 points.

With weighting decided, the scoring schemes for items could be finalized. Scoring was based on both correctness and consistency. It was possible on some items to answer the MC prompt incorrectly and still get some credit if reasoning was consistent with their incorrect response. For example, item 11 on the U-STEP relied on recall of the Boltzmann factor; students could receive partial credit if their reasoning was consistent with a response corresponding to assumption of an

incorrect sign in the exponent of the Boltzmann factor. Most MC prompts, with the exception of two,[7] were scored dichotomously (i.e., either fully correct or fully incorrect, with students receiving either 0 or 2 points).

Scoring for MR prompts were more complex, as they had to account both for accuracy and consistency with the MC selection. Each MR selection was assigned a certain number of points, which could either add or subtract from their score for reasoning. For example, some reasoning elements were worth +0.5 points, +1 points, +2 points, etc. while others were worth -0.5 points, -1 points, etc. (all scoring allowed for only half-integer and integer point values). If a reasoning element was a correct statement, but irrelevant in determining the correct response, that reasoning element was assigned a score of 0 points. Some reasoning selections were assigned a score such that if that reasoning element was selected, the entire reasoning score would be cancelled to zero. In some instances, reasoning elements had to be selected in conjunction for either to count towards the score. For reasoning, a code was created in a spreadsheet such that total scores would remain within the bounds of 0 points and 3 points.[8] An example scoring scheme and its application is presented in Figure 4.1. Note Student 2 in Figure 4.1 received partial credit for consistency despite selecting an incorrect answer for the MC prompt. The scoring schemes used for each item is presented in depth in Appendix E.

Scoring schemes were discussed and finalized with a team of three researchers, and thus there is some level of subjectivity in the judgements made regarding how many points each selection was worth. It is worth acknowledging the idea that, if others were conducting this work, they may have potentially developed different scoring schemes than the researcher-generated ones presented in this dissertation. However, it is unlikely any large differences would appear in deciding which selections are definitively correct or incorrect.

After analysis of both interviews and validation statistics, some scoring schemes were modified

[7] These two prompts were part of the same item and had the same response options (see item 14 in Appendix D). A partially-correct response would receive 0.5 points out of 2 points.

[8] It is possible for students to select a series of reasoning elements that summed to less than zero. The minimum score was set to zero such that reasoning selections could not subtract from their MC or overall score.

Scoring Scheme:

	A	B*	C	D
a	0	0	0	0
b	-3	-3	0	0
c	0	0	0	0
d	-3	-3	0	0
e	-3	0	0	0
f	-3	-3	0	0
g	0	0	0	0
h	0.5	1.5	0	0
i	-3	-3	0	0
	A	B*	C	D
a, c	0.5	1.5	0	0

Example Responses:

Student 1:
Selected **B** for MC prompt; **a, c, h** for MR
MC score: 2 (correct answer)
MR score: 0 + 0 + 1.5 + 1.5 = 3 pts.
Total score: 2 + 3 = **5 pts.**

Student 2:
Selected **A** for MC prompt; **a, g, h** for MR
MC score: 0 (incorrect answer)
MR score: 0 + 0 + 0.5 = 0.5 pts.
Total score: 0 + 0.5 = **0.5 pts.**

Student 3:
Selected **B** for MC prompt; **a, c, d, f, h** for MR
MC score: 2 (correct answer)
MR score: 0 + 0 + (-3) + 1.5 + (-3)+ 1.5 = -3 pts.
-3 → 0pts.
Total score: 2 + 0 = 2 pts.

Figure 4.1: Scoring scheme for Item 1 of the U-STEP (left) and example responses with corresponding scores (right). Multiple choice (MC) options (A-D) are in the top row, while the left-most column lists multiple-response (MR, reasoning) options. Students select one MC answer and as many MR options as they desire in order to support their response. All other entries within the table are scores assigned to each response. Note for this example, a & c must be selected together with B to receive credit for either. The * indicates the correct MC response. See Appendix D to see this item and Appendix E to see scoring schemes for all items.

before finalization. This was often done because the initial scoring seemed too harsh causing a decrease in scores for many students or students from particular demographic groups. Sometimes it was realized by the research team that some elements docked points because it was not a line of reasoning that would directly support the MC response despite being a true statement, and thus scoring for these statements were changed such that the selection would neither add nor subtract from the score (i.e., assigned 0 pts.).

Due to these scoring schemes, some items were worth up to 15 pts. (if composed of multiple CMR pairs, such as the item in Figure 3.9) while others were worth only 2 pts. (i.e., pure MC). To avoid some items dominating the scoring of the assessment due to their allotment of more possible points, the research team decided to weight each item equally for overall test scoring. After assigning

point values for each item, each maximum item score was normalized to 1 using the total number of possible points for each item. For example, for the item presented in Figure 4.1, each score would be divided by 5; this would mean Student 1 received a score of 1, Student 2 received a score of 0.1, and Sudent 3 received a score of 0.4. Normalized scores were all added together then divided by 15 (the maximum score possible for the normalized assessment) to determine the overall score for each student. Note that though the normalized scoring scheme was used, the unnormalized scoring scheme did not produce large differences in the statistical measures of validity of the assessment.

4.5 Criterion Validity

Another route for establishing validity beyond content and construct validty, is to perform *criterion validity* analyses, comparing assessment performance to other relevant measures [59]. These analyses must be conducted after MR piloting and scoring. For this study, overall assessment averages were compared to students' average exam scores and final course grade for a subset of respondents at a single institution (N=76).[9] We found inconclusive evidence regarding criterion validity. Correlations between students' average exam scores and achieved assessment score was 0.431, while the correlation between students' final course grade and assessment score was 0.357. Some consider the range of 0.4-0.7 to be an acceptable range for validity correlations, while others have argued coefficients within that range are too low to account for a sufficient amount of variance between scores [82].

There are several factors that could affect the data used for the above correlation coefficients. The semester these data were collected was during a remotely-taught course, in which course expectations were different from the norms of the in-person course. For example, exams in the course were take-home and open-book, while students were asked to not access any resources while taking the assessment. Additionally, for all exams with exception of the final, students were able to do test corrections to improve their exam scores. Exams were also weighted less heavily than normally done for the considered course. These course modifications may change the nature of

[9] We only had access to these measures at one institution.

what exams are testing and therefore may not represent an appropriate comparison for criterion validity analysis. Additionally, it is worth highlighting that this analysis could not be done for all respondents in the pilot because course performance data was only available from one institution.

4.6 Classical Test Theory Analysis

Classical test theory (CTT) is based on the assumption that a student's score on an assessment is composed of two scores: a true score and a score due to random error (which could be due to measurement error, testing conditions, etc.). It is assumed that the true score would be a measure of student ability and the error score accounts for fluctuations from the true score as measured by the assessment. A key assertion is that the error is random, and thus true scores of a population in aggregate can be accurately measured via averaging.

Another assumption of CTT is that the population tested is representative; this must be the case for the outcomes and measures of the assessment to hold for *any* population. However, populations are rarely fully representative of all sub-populations, especially in physics, which is predominately White and male [10]. Thus, outputs of the model, and therefore the test statistics themselves, are population dependent and are likely to change when a different populaton is tested.

To align with traditional PER assessment development, CTT analyses were done for the U-STEP. This was done, in part, to identify items that needed to be revised or removed when creating the final U-STEP. These analyses include calculations of item difficulty and discrimination, as well as overall assessment discrimination (i.e, Ferguson's delta) and a coefficient related to internal consistency (i.e., Cronbach's Alpha), as discussed in Sec. 4.6.1. Additionally, they included a differential item functioning analysis of performance differences between demographic groups in attempt to curb bias in the assessment (Sec. 4.6.2), an important aspect of assessment development.

4.6.1 CTT Validation Statistics

As mentioned above, the key CTT outputs discussed here are item difficulty, item discrimination, Ferguson's delta, and Cronbach's Alpha. Item difficulty is a measure of how difficult an item

is to answer correctly, and is reported as the average score on the item [83]. This means that higher difficulty values actually respresent easier questions, whereas low difficulty values represent more challenging questions. Discrimination refers to the extent to which an item or test can distinguish between high- and low-performing students [59, 83]. Higher discrimination values indicate better differentiation between high- and low-achievers.

Table 4.2 presents difficulty (b) and discrimination (a) measures for the Spring 2020 and

Table 4.2: CTT validation results—difficulties (b) and discriminations (a)—for Spring and Fall 2020 pilot administrations of the U-STEP items. N-values for the Spring administration change due to the different versions of the assessment piloted and varied number of institutions receiving each version. (Each version was composed of a set of 6 anchor items and 7 secondary items, which differed based on items.) N-values for the Fall pilot remain the same due to only a single version being piloted. [A]Note: The discrimination from anchor items for the Spring 2020 pilot are presented as averages across the two versions.

			Spring 2020 Pilot			Fall 2020 Pilot		
Item	Topical Area	Focal Topic	N	b	a	N	b	a
1	work	PV diagram	99	0.28	0.14	164	0.25	0.21
2	first law	PV diagram	169	0.47	0.23[A]	164	0.52	0.49
3	heat	See Fig. 3.7	70	0.51	0.51	164	0.45	0.40
4	stat. mech.	micro/macrostates, probability	169	0.54	0.35[A]	164	0.51	0.27
5	stat. mech.	multiplicity	99	0.75	0.40	164	0.76	0.24
6	equilibrium	thermal, mechanical, diffusive	99	0.72	0.41	164	0.63	0.53
7	entropy	mixing gases	70	0.58	0.45	164	0.66	0.32
8	entropy	Carnot engine & entropy	99	0.51	0.38	164	0.51	0.46
9	entropy	heat flow between solids	70	0.49	0.43	164	0.51	0.46
10	energy	degrees of freedom	70	0.60	0.48	164	0.57	0.30
11	stat. mech.	See Fig. 3.9	169	0.47	0.44[A]	164	0.42	0.56
12	stat. mech.	entropy & Z[†] for single state	99	0.61	0.42	164	0.51	0.49
13	engines	heat and work	99	0.40	0.22	164	0.46	0.40
14	engines	entropy-temperature diagram	70	0.23	0.26	164	0.31	0.36
15	temperature	isotherm of ideal gas	70	0.59	0.36	164	0.68	0.31
16	temperature	See Fig. 3.6	169	0.65	0.30[A]	N/A	N/A	N/A
17	stat. mech.	fundamental assumption [68]	169	0.52	0.13[A]	N/A	N/A	N/A
18	heat	PV diagram	169	0.40	0.30[A]	N/A	N/A	N/A
19	energy	equipartition	99	0.19	0.33	N/A	N/A	N/A
20	stat. mech.	probability	70	0.14	0.20	N/A	N/A	N/A
21	heat	See Fig. 3.6	70	0.51	0.51	N/A	N/A	N/A

[†]Partition function

Fall 2020 pilot administrations of the preliminary and final U-STEP items.[10] Difficulty is typically defined by the proportion of correct responses with respect to the total number of responses; however, this definition only makes sense in the context of dichotomous data. Here, difficulties were found by averaging all scores achieved by individuals for each item; CTT difficulties values can only range from 0 to 1. For the Spring 2020 pilot, items difficulties ranged from 0.14-0.72. For the Fall 2020 pilot (which contains the final set of items to be included in the U-STEP) item difficulties ranged from 0.25-0.76. The literature suggests ideal difficulties lie with the range of 0.30-0.90 [83], and only one item piloted in Fall 2020 fell outside this range. Overall difficulties for the two assessment versions piloted in Spring 2020 were 0.52 and 0.44. For the single version of the U-STEP piloted in Fall 2020, the overall difficulty was 0.52.

Discrimination values for items were determined using a Spearman correlation between item scores and average score on the rest of the assessment [84]. For the Spring 2020 pilot, item discrimination values ranged from 0.14-0.51. For the Fall 2020 pilot, item discriminations ranged from 0.21-0.56. The literature suggests these values should lie above 0.30 [83]. Seven items from the Spring 2020 pilot (items 1, 2, 13, 14, 17, 20, and 21), three of which were dropped for the Fall 2020 administration, fell below this threshold. Three items in the Fall 2020 pilot fell below this threshold, with a=0.21 (item 1), a=0.27 (item 4), and a=0.24 (item 5). Overall assessment discriminations, as measured by Ferguson's delta, were 0.99 and 0.98 for the two Spring 2020 versions and 0.99 for the Fall 2020 version.[11] These meet the desired requirement of $\delta \geq 0.9$ [83]. Cronbach's Alpha is a measure of internal consistency and it is recommended this coefficient has a value of $\alpha \geq 0.7$ to be considered acceptable for group comparisons [85]. The two Spring 2020 versions yeilded values of α=0.71 and α=0.69, while the value for the Fall 2020 pilot was α=0.78. This indicates the instrument is reliable at the level of group measurement as needed for this type of assessment.

[10] Note these analyses used the normalized scoring scheme discussed in Sec. 4.4.
[11] Note the unnormalized scoring scheme was used to determine Ferguson's delta.

4.6.2 Differential Item Functioning with CTT

Differential item functioning (DIF) involves statistical comparisons of subgroups within a population to investigate the extent to which items may be measuring different abilities for students with similar scores on the assessment overall. This can be one route to identify bias within items—if students in different subgroups (e.g., men and women) score similarly on the assessment, but score vastly differently on a particular item, that may indicate issues with the item that need to be addressed.

Often differences in performance based on gender and/or race are compared using averages of all students within each considered population. DIF differs from this route by separating students by ability as measured by the overall assessment and then comparing averages on individual items between subgroups in ability categories. The CTT approach to DIF used for the U-STEP involved looking at average scores on items for the top and bottom 25th percentiles based on demographic group (i.e., gender or race) and identifying items that had significant differences between subgroups.

Students were sorted into ability levels by first ranking students by overall assessment score, from highest scores to lowest. Then, the bottom and top 25th percentiles (i.e., the 25% of students with the lowest overall scores and 25% with the highest overall scores) were identified. These two groups were then split by gender or race for the analyses. Since the percentile grouping depended only on rankings of overall scores, N-values for each demographic group within these percentiles are not equal. For example, on one version of the assessment piloted in Spring 2020, the top and bottom 25th percentiles were each composed of 25 students. In the top 25th percentile, 4 were women, whereas 18 were men.[12] N-values for all DIF analyses can be found in Appendix F.

The purpose of splitting students into percentile groups was to allow for comparison of students of similar abilities. By focusing solely on averages for *all* students in a particular demographic group, possible issues of bias may be suspected when there may not actually be any. For example, if a particular group scores lower on certain items, it may just be that the students in that

[12] Others in the 25th percentile were non-binary or did not report demographic information.

group, for one reason or another, overall have lower ability.[13] Often performance differences, or "gaps," are analyzed in terms of overall group performance on items, as opposed to by *specifc ability level* as measured by the assessment. Comparing group differences in terms of *all* students within a particular group can imply bias in the assessment where it may actually be absent. This can become problematic when considering analysis and interpretations of the gaps, which often inform construction of various interventions or assessment revisions. Some DIF analyses of physics assessments, which compare students of similar abilities, have been conducted (e.g., ref. [6]), but these are typically done *after* the assessment has been formalized.

DIF analyses involving students of similar abilities are important because if gaps on average item scores appear between similar-ability students from different groups, this may indicate issues with an item. This is because one would expect students of similar abilities to achieve similar scores on each item. If this is not the case for a particular item, that indicates potential issues with the item—such as bias—that may exist, warranting a closer investigation into the structure of the item.

The analyses presented here focus on two genders—men and women—and three racial categories—Asian, White, and underrepresented minority (URM).[14] These groupings were largely informed by demographic representation in STEM [10]. Men are overrepresented compared to women in STEM. Similarly, Asian and White students are overrepresented in STEM compared to URM students. Asian and White are split into separate categories due to their distinct racial categories. This is an important distinction due to the way race influences the experiences students encounter when pursuing STEM degrees (as discussed in more detailed in Ch. 6-Ch. 9).

For the Spring 2020 pilot, only one item (item 16 in Table 4.2) resulted in statistically significant ($p<0.05$) differences of average scores between racial groups (White and Asian students, and White and URM students) of similar abilities as determined by a Mann-Whitney analysis [86];

[13] Note the term "student ability" is a measure of performance as opposed to innate ability of individuals. See Sec. 2.4.2.

[14] Students were asked to provide demographic information at the end of the assessment. Students who selected "prefer not to answer" are not included in the DIF analyses, though may have been in the upper or lower 25th percentiles. Similarly, non-binary students are not included in the gender analysis due to low N, though non-binary students may have been in the upper or lower 25th percentiles.

no statistically significant differences between men and women or Asian and URM students were detected. No changes were made to this item based on this analysis, as the item was eventually combined with item 21 to produce a single CMR item (see Fig. 3.7).

Results from the CTT DIF analyses for the Fall 2020 pilot administration are presented in Appendix F.[15] One item (item 13 in Table 4.2) was found to have differences in the mean score of men and women for both the upper and lower 25th percentiles that are statistically significant at the 0.01 level. No significant differences were found with this item for race. Item 13 relies on recall of the term *adiabatic*.[16] Analyses of response patterns to the item found similar frequencies of selection of the distractor requiring knowledge of this term (an average of 6.5% for men and 10% for women). No changes to this item were made based on this analysis. Though the definition of "adiabatic" could be provided in the prompt or response options, much like what was done with "state function" in Sec. 3.3.1.5, provision of this definition would make the problem trivial (as can be seen with considering the prompts in the item, presented in Appendix D). However, moving forward, this item will be re-evaluated for inclusion in the finalized U-STEP pending results from the final pilot administration in Spring 2021.

4.7 Item Response Theory Analysis

Item response theory (IRT) can facilitate differential item functioning, and came about to rectify some shortcomings of CTT [59]. Though CTT assumes a representative sample, this is not actually the case in most instances, leading to statistics dependent on the population with which the assessment was piloted (i.e., CTT statistics vary when the population changes). This becomes particularly problematic when it comes to fairness in testing. That is, if assessments are mainly piloted at predominantly White and male departments (which describes most large physics departments), the validation statistics are based on that population, as opposed to other populations composed of students of color and women. This can lead to the test favoring students of the sampled popu-

[15] Scoring was revised for one item during the period analysis was conducted. DIF analyses before and after this revision did not result in statistically significant gender or race performance differences.

[16] Adiabatic refers to processes in which heat neither enters nor leaves the system.

lation (because the test statistics were determined with that group), leaving students from other groups to receive different (historically lower) scores. Some scholars have discussed a hyper-focus on "achievement gaps" in assessments [87]. Utilization of only CTT validation approaches could be a factor in the appearance of these gaps due to the population-dependent nature of the approach and the fact that the populations typically used are predominantly composed of only certain demographic groups. IRT, on the other hand, allows for determination of test statistics that do not vary wildly when testing population changes, provided the initial validation pool is broad.

The two-parameter logistical (2PL) IRT model is composed of two main variables: item discrimination (a) and item difficulty (b). The three-parameter (3PL) model incorporates a third parameter accounting for guessing, but is not considered here due to the low likelihood of students guessing the correct response pattern for CMR items. According to the 2PL model, the probability of a student of ability θ scoring correctly on an item (i) is given by:

$$P(\theta_i) = \frac{1}{1 + e^{-a_i(\theta - b_i)}} \tag{4.1}$$

Unlike CTT, the item difficulty and discrimination parameters need not range from 0 and 1, and can instead take on *any* value (including negative values), though they typically take on values between -3 and 3. In general, each item has its own difficulty and discrimination values. Higher, more positive difficulty values indicate more challenging items,[17] while more negative values indicate the item is easier. Discrimination is typically positive, unless an item is not behaving as intended, and indicates how well an item distinguishes (i.e., *discriminates*) between high- and low-ability students; higher values indicate better discriminatory power. Ability[18] (θ) is defined such that higher (positive) values indicate higher ability, and more negative values indicate lower ability. Each student has a particular ability value θ.

Equation 4.1 can be used to plot Item Characteristic Curves (ICCs) for each item. ICCs allow one to interpret various properties of an item, namely difficulty and discrimination. Difficulty in IRT can also be defined as the ability level required to have a 50% chance of answering an item

[17] Note: This behavior is opposite for CTT, in which higher difficulty values indicate easier items.
[18] Note again "ability" is used here to align with the literature. See Sec. 2.4.2.

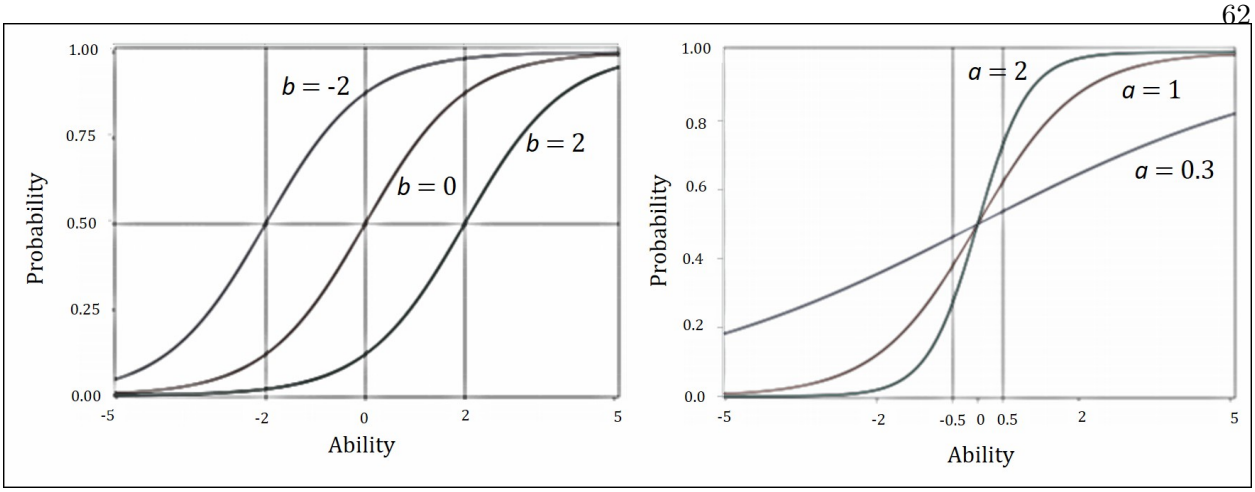


Figure 4.2: Modified from ref. [88]. Item Characteristic Curves with varying difficulties (left) and discriminations (right). The left image shows an easy (b=-2), medium (b=0), and difficult (b=2) item, all with the same discrimination. The difficulty value is the ability level at which the probability of a correct response is 50%. The right image shows ICCs with the same difficulties and varying discriminations, with a=2 having the most discriminatory power and a=0.3 having the least. Steeper slopes at the 50% probability location (inflection point) indicate higher discrimination.

correctly, located at the inflection point of an ICC.[19] Discrimination can be read from an ICC via the slope at the steepest point of the graph (i.e., the inflection point). Figure 4.2, modified from ref. [88], shows examples of ICCs with different difficulties and discriminations; each curve would represent characteristics of a different item. Test characteristic curves look similar to ICCs, but display information for the entire test as opposed to a single item.

4.7.1 The Rasch Model

The one-parameter IRT model, also known as the Rasch model, is a simplification of the 2PL model. In the Rasch model, all item discriminations are assumed to be 1 [63]. As a results of this assumption, the Rasch model reduces Equation 4.1 to:

$$P(\theta_i) = \frac{1}{1 + e^{-(\theta - b_i)}} \tag{4.2}$$

The mathematical formalism of the Rasch model assumes the probability of a student an-

[19] The definition of difficulty varies slightly for the three-parameter logistical model.

swering an item correctly is determined by a latent trait (i.e., *ability*) and item difficulty. This allows ability and item difficulty to be mapped onto the same invariant scale, leading to consistency among item difficulty measures when a new population is tested. This is beneficial for DIF analyses to identify bias. Another benefit of Rasch analyses is that they require a much smaller sample size than those required for the 3PL and 2PL models [89]. Though model fit depends on both sample size and test length, some have suggested a sample size of 200 respondents and 15 items (the same number of items on the U-STEP) provides sufficient fit statistics [90]. This is promising because after the next pilot administration of the U-STEP the sample size should meet or exceed this threshold when combined with the Fall 2020 data.

As with all IRT models, the fit of the Rasch model to the data cannot be assumed, but instead must be evaluated. Thus, before measuring outputs of the model, such as item difficulties and student abilities, it is important to check that the model applied fits the data sufficiently. This is because Rasch model outputs cannot be accurately interpreted unless the model fits. One implication of the Rasch model's assertion of a=1 is that discrimination cannot vary to better fit the data. Instead, this means in order to get a better fit for the model, one must adjust the items as opposed to the model. This has implications for designing and testing items.

4.7.2 Preliminary analysis using the Rasch Model

A restriction of the original Rasch model is that it requires dichotomous data. However, data collected from CMR items are not dichotomous, and can instead take on various values between 0 and 1. Because of this property of CMR items, and the limited number of CMR-based assessments, there are few investigations of Rasch analyses applied to CMR items. Though polytomous Rasch models exist (see Sec. 4.7.4 for an example), they require much larger sample sizes than dichotomous Rasch analyses; smaller sample size restrictions are beneficial for upper-division courses, which have a relatively small pool of students to draw respondents from.

The goal motivating the analyses presented here is to explore application of Rasch to CMR items to see if this validation method is possible or feasible for use in finalizing the U-STEP or

analyzing future CMR assessments. To explore the possibility of applying the dichotomous Rasch model to the U-STEP, we investigated dichotomizing CMR item scores. Namely, we created various score thresholds that were required to obtain a score of 1, with any score less than that threshold being converted to a score of 0. These analyses lay a foundation for novel IRT approaches for CMR assessment analysis.

The typical CMR item on the U-STEP is composed of one MC prompt and one MR prompt (or a series of MC-MR pairs). Since the scoring scheme we chose assigned 2 points for the MC portions and 3 points for the MR portions, the typical CMR item was worth 5 points (which was then normalized to 1).[20] For this exploratory analysis, we looked at a 40% threshold, 50% threshold, and 60% threshold to determine whether a score was converted to a 0 or 1. A 40% threshold would, in general, indicate the student answered at minimum the MC prompt of a CMR item correctly (i.e., a score of 2 or above on a 5-point item). The 50% threshold would indicate at least some credit for reasoning was received in addition to a correct response to the MC prompt (i.e., a score of 2.5 or above). The 60% threshold, in general, indicates students received at least one-third of the possible points for reasoning in addition to a correct response to the MC prompt (i.e., a score of 3 or above). Note these general rules vary slightly by item, as not all items were worth the same amount of raw points (which were then normalized); however, the above are the justifications used for determining the thresholds tested.

To see how well the dichotomous score transformations aligned with the polytomous scoring scheme, a Pearson correlation was conducted between total assessment scores from the traditionally scored CMR items and the dichotomously scored total scores determined using the three thresholds. Correlation coefficients were 0.960, 0.967, and 0.958 for the 40%, 50%, and 60% thresholds for the overall assessment, respectively. However, it should be noted there were several students whose scores on the traditionally-scored CMR assessment were somewhat higher or lower than their dichotomized total score, as seen in Figure 4.3.

[20] Note some items were worth more than 5 points. However, all items were normalized to 1 such that they were weighted equally.

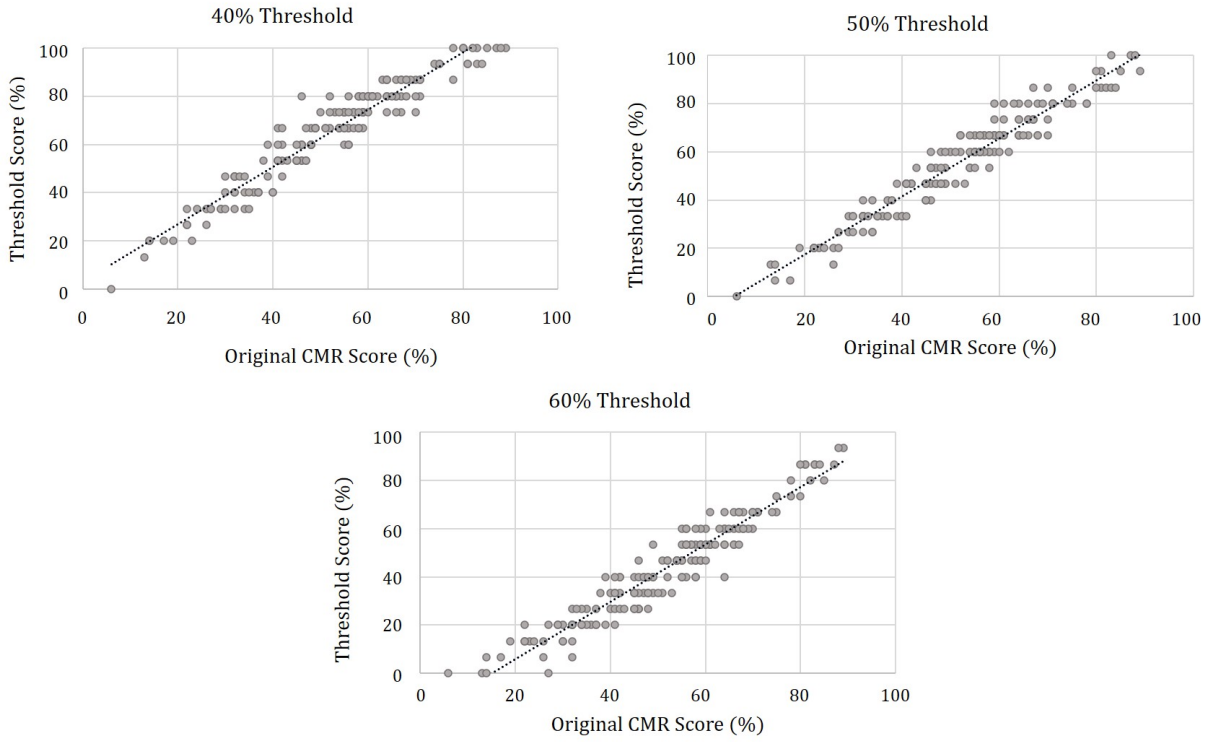


Figure 4.3: Scoring for the overall assessment for all students in the Fall 2020 pilot, comparing the original CMR scoring scheme with the dichotomously-scored schemes for the 40%, 50%, and 60% thresholds. Threshold and original CMR scoring results are presented as percentages determined by the total number of points received on the assessment divided by the total number of points possible (i.e., 15 pts.)

Exploratory Rasch analysis was done only using the Fall 2020 data, utilizing the mirt package in the computer program R [91]. We used only Fall 2020 data because the assessment piloted that semester was a single version, allowing for a larger number of responses (N=164) for analysis than could have been analyzed with the Spring 2020 pilots of two separate versions. Overall model fit was tested using an M_2 statistical test. This test was proposed by Maydeu-Olivares and Joe [92] and provides a measure of model fit for the overall assessment. This is an important step in the analysis process because it allows one to gauge reliability of model outputs, such as item difficulty and student ability. Large M_2 values correspond to lower p-values and more model misfit. Additionally, root mean square error of approximation (RMSEA) and comparative fit index (CFI) provide another measure of overall fit. Recommended cutoffs for these outputs vary across the

Table 4.3: Overall assessment fit statistics using three thresholds for dichotomizing data. Significant misfit is indicated by $p<0.05$. Root mean square error of approximation (RMSEA) less than 0.06 indicates a relatively good model fit; comparative fit index (CFI) values greater than 0.95 indicate good model fit [93]. Results combined indicate good model fit for each threshold utilized, with lower thresholds displaying better fit.

	M_2	p	RMSEA	CFI
40% Threshold	117.01	0.199	0.027	0.973
50% Threshold	118.65	0.171	0.028	0.969
60% Threshold	125.90	0.080	0.035	0.951

literature and often depend on the estimation method used [93]. However, conservative discussions suggest an RMSEA value less than 0.05 indicates a close fit [94, 95] and CIF values greater than 0.95 indicate relatively good model fit [96]. Table 4.3 presents M_2, RMSEA, and CFI values from the overall model fit analyses. Each tested threshold—40%, 50%, and 60%—yielded acceptable overall model fit according to the literature, with the 60% threshold being the closest to potential misfit. This indicates the data pulled from the Rasch analyses for these thresholds, including item difficulty and student ability measures, are reliable enough and imply these theshold methods may present promising avenues for Rasch analysis of CMR items.

The test characteristic curve for the three threshold analyses are presented in Figure 4.4. As one might expect, the difficulty of the assessment increased as the threshold values increased. This trend was also true for most items. This was not the case for all items because some scoring schemes led to identical results for different thresholds. For example, the 40% threshold and the 50% threshold difficulties and resulting ICCs were identical for items in which 5 points were possible and reasoning only allowed for integer-values of point allotment for reasoning (i.e., one could not receive a 50%, and thus the two thresholds were essentially identical). Trends in difficulty were similar for these analyses and the CTT analyses with the original CMR scoring scheme. That is, more difficult items as measured by IRT were also rated as more difficult with the CTT analysis. IRT difficulty values are presented in Appendix G.

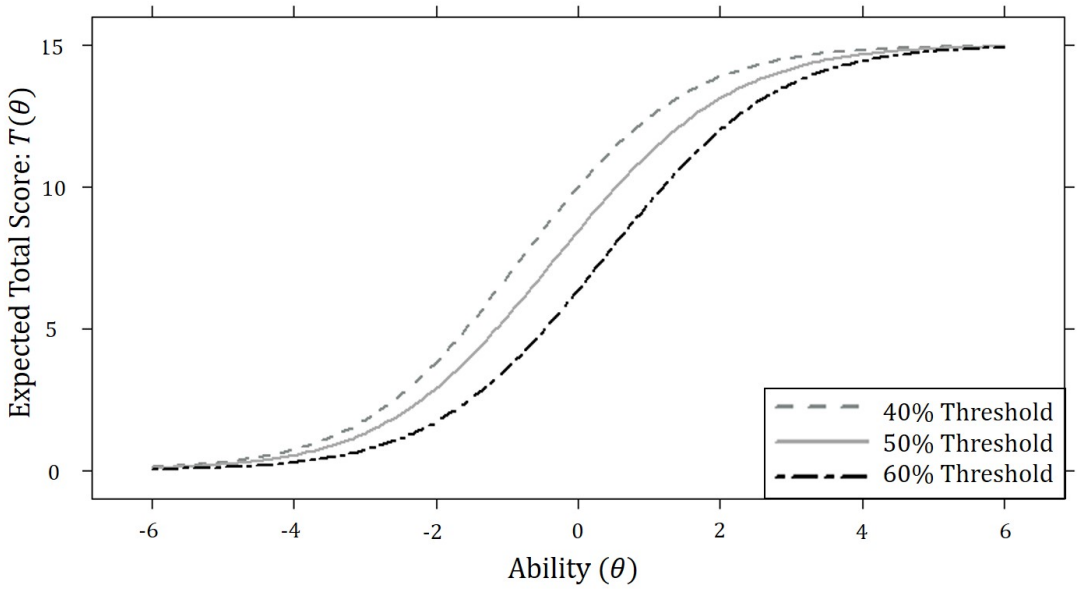


Figure 4.4: Test characteristic curve for the Rasch analysis of the Fall 2020 piloted assessment using 40%, 50%, and 60% thresholds for dichotomizing data. The expected total score ranges from 0 to 15 because there are 15 items total, each worth 1pt. The difficulty of the assessment increases as the threshold increases, as indicated by the shift to the right with increasing threshold.

Item fit statistics were determined using a Pearson's chi-squared test (S-X^2), which has been found to produce a better detection of model misfit than other measures [97]. Fit statistics, along with IRT-produced difficulty values for the 15 items included in the Fall 2020 pilot can be found in Appendix G. Higher S-X^2 values indicate a higher level of "misfit" (i.e., lower S-X^2 indicate better fit). Additionally, RMSEA values less than 0.01 indicate "excellent" fit for the S-X^2 statistic, while values greater than 0.08 indicate mediocre fit [98]. Thus, item-model misfit can be deduced from high S-X^2 values, low p-values ($p<0.05$), and RMSEA$>$0.08. Misfit items (see Table 4.2 for item descriptions) for each threshold included:

- 40% Threshold: item 14
- 50% Threshold: items 1, 7, 13, 14
- 60% Threshold: item 1, 5

Analyses for each threshold were re-ran with misfitting items removed. Appendix G presents

Table 4.4: Item fit statistics when items are removed, determined using three thresholds for dichotomizing data. Significant misfit is indicated ($p<0.05$). Items were removed if they appeared with statistically significant misfit for the entire assessment for each threshold. New/Remaining misfit items are items that appeared with misfit that did not appear in the intitial analysis or items that remained misfit before and after the associated item was removed.

40% Threshold		50% Threshold		60% Threshold	
removed item (p)	new/remaining misfit items (p)	removed item (p)	new/remaining misfit items (p)	removed item (p)	new/remaining misfit items (p)
14 (0.020)	6 (0.034)	1 (0.040)	7 (0.048)	1 (0.023)	4 (0.008)
			13 (0.007)		5 (0.003)
		7 (0.034)	13 (0.020)		13 (0.004)
			14 (0.037)	5 (0.018)	N/A
		13 (0.040)	1 (0.037)	1, 5	13 (0.013
			10 (0.032)		
		14 (0.018)	3 (0.016)		
			13 (0.016)		
		1, 7, 13, 14	5 (0.006)		

detailed reporting of the produced S-X^2 analyses, and related p-values and RMSEA values, for these analyses for all items. Table 4.4 summarizes specific removed items and the items that appeared as significantly misfit when those items were removed. When removing item 14 for the 40% threshold, a new item (item 6) appeared with signicant model mifit (p=0.034). Removal of any individual misfit item for the 50% threshold reduced the total number of other misfit items, though never eliminated all misifit. Item 13 remained misfit in all misfit-removal analyses for the 50% threshold, and even emerged as misfit in the 60% threshold analyses when item 1 was removed. This indicates this item may warrant further investigation once a larger sample size is available and more robust analyses can be completed (i.e., after the Spring 2021 pilot). This could result in revision or removal of the item. The 50% threshold initial analysis highlighted 4 misfit items, and even when all 4 of these items were removed, statistically significant model misfit still appeared (emerging from new items). This suggests that use of the 50% threshold produces the most model-data misfit and may not be ideal for future Rasch modeling of CMR items. Notably, removal of item 5 from the 60% threshold yielded no other item misfit. Removal of items for reanalysis did not significantly change difficulty values output from the model for any threshold.

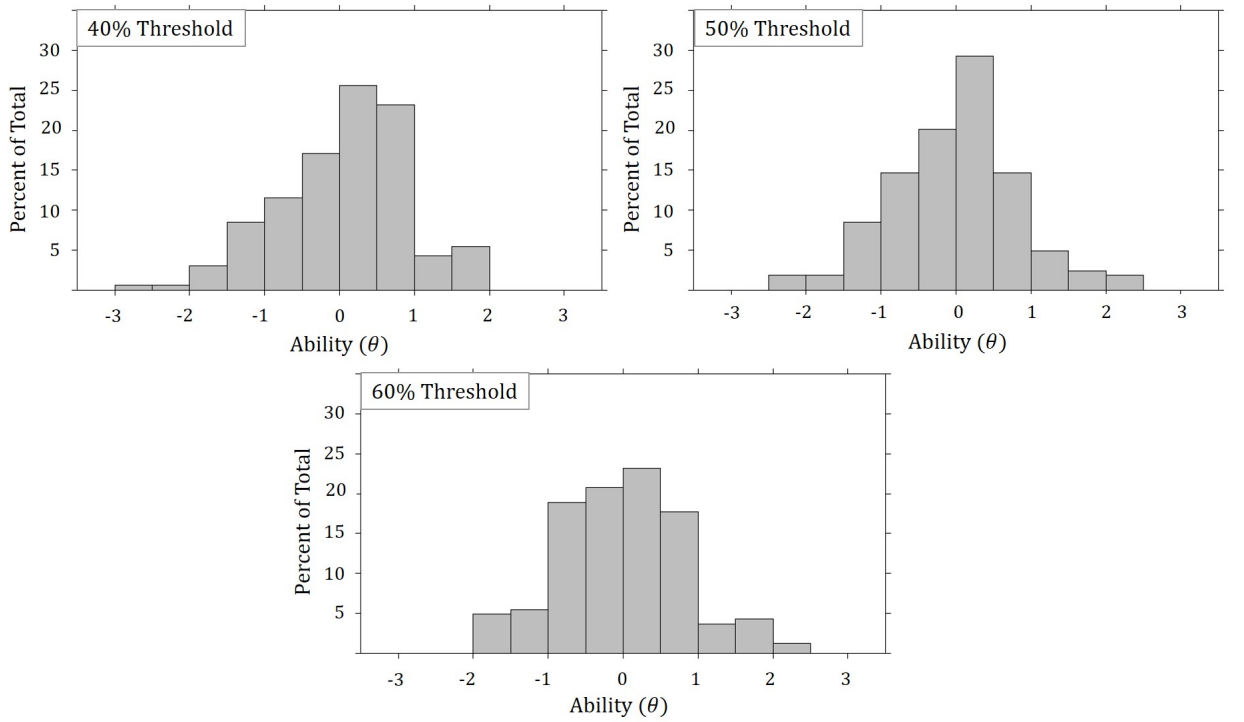


Figure 4.5: Histograms of student ability measures determined with the Rasch analysis of the Fall 2020 piloted assessment using 40%, 50%, and 60% thresholds for dichotomizing data. The average abilities across the three thresholds remains constant and centered at 0. Higher, more-positive values indicate more ability, while more-negative values indicate lower ability.

The encouraging results from these Rasch analyses led to consideration of student ability measures and how to report them to instructors. Ability estimates were determined with the three thresholds used, as presented in Figure 4.5. The lower-bound of ability measure shifts to more-positive values with increasing threshold. This is due in part to the assumptions of the model, one of which is that the mean of the abilities is 0 with a standard deviation of 1 [63]. A major goal of assessments like the U-STEP is to provide information to instructors about their students' aggregate performance on the assessment to inform their instruction. Distributions of student ability measures are one output from the analyses that could be provided to instructors as feedback regarding their class's performance. Histograms such as those presented in Figure 4.5 are one possibility of reporting these abilties to instructors while still maintaining anonymity for students. Another option is to report average performance determined by ability, though that would require

provision of detailed descriptions of how to interpret the results, which is less necessary when reporting averages determined by CTT. Reporting of student performance measures as determined by IRT while maintaining anonymity is still an open question, especially when considering smaller class sizes.

Another dichotomization method was considered in addition to the three thresholds discussed above. The method involved dichotomizing items based on whether responses were fully correct (score of 1) or not fully correct (score of 0). Rasch analysis of this scoring scheme could not be conducted for all U-STEP items, as there were two items (items 9 and 13)[21] in which no fully correct responses were achieved, leaving all students to receive a score of 0 on those items.[22] Rasch analysis without these items produced statistically significant misfit for only one item (item 7). Overall assessment fit with items 9 and 13 removed was acceptable according to M_2, RMSEA, and CFI measures. However, one must interpret these results with caution, as the assessment had to be dissected (i.e., split due to item removal) in order to be analyzed.

Despite the seemingly sufficient fit, CTT measures of difficulty using this scoring scheme were very low, with only 4 items (3 of which were pure MC) falling within the accepted difficulty range of 0.30-0.90 [83]. It is also of note that CMR items are intentionally designed to award partial credit, which can be earned through an array of reasoning patterns, and it is quite difficult to achieve a fully correct response on CMR items. Given the analysis described above, we do not recommend that this dichotomization method be used for future Rasch analyses of CMR items like those on the U-STEP because (1) the intention of the items' scoring schemes was never to have items scored as fully correct or incorrect; (2) low CTT difficulties were produced from this dochotmization method; and (3) the analysis could not be ran for the full assessment.

Some literature refers to populations of 500 respondents as "small sample sizes" for analyses of model misfit with S-X^2 [99]. Thus, the sample used in these analyses (N=164) may be too small

[21] Item 9 addresses entropy of interacting solids, composed of 3 MC-MR pairs. Item 13 addresses heat and work with an entropy-temperature diagram, composed of two MC questions and an MC-MR pair.

[22] Rasch analyses require at least *some* distribution of scores for each item. If this is not the case for all items, the analysis cannot run successfully.

to make meaningful conclusions about the model fit. However, an additional pilot of the U-STEP is planned for Spring 2021, prior to its finalization. Since only minor revisions have been made since the Fall 2020 pilot, responses to the Fall 2020 and Spring 2021 administrations will be able to be pooled, providing a larger sample to run the analysis with and putting our sample size closer to this suggested lower-bound. This preliminary work lays the foundation for future analyses of the assessment with larger sample sizes when they become available and suggests possible best-practices for dichotomizing polytomously-scored data retrieved from traditional CMR items for use of IRT for validation. The work also shows that model-fitting CMR data with a dichotomized Rasch analysis may be a promising new route in validation of MR assessments.

4.7.3 Differential Item Functioning with IRT

IRT analysis methods also allow for differential item functioning (DIF). However, it is recommended to have a minimum of 100 respondents per group for DIF analyses with IRT [100, 101]. These restrictions to this method made it not possible to run any meaningful DIF analyses with our preliminary dataset. For example, with this restriction, gender analyses would not be possible, as only 32 of the 164 Fall 2020 respondents were women. Additionally, group sizes are not near 100 for any racial category considered. For these reasons, IRT-based DIF analyses were not conducted. However, this will be a focus of future work when more responses to the U-STEP are collected.

4.7.4 The Partial Credit Model

The threshold analyses presented in the previous section were motivated by the requirement of dichotomously-scored data for the Rasch model. Though the analysis suggests promising potential analysis methods for CMR items, a benefit of using CMR items in assessment is the allowance for partial credit (i.e., items that are *not* scored dichotomously, but polytomously) to account for consistency and reasoning. Dichotomously-scored items and utilization of the Rasch model cannot fully account for these aspects of CMR items. However, there is another IRT approach that could be more applicable for these items: the partial credit model (PCM). As its name implies, the PCM

was born from traditional IRT to can account for *partial credit*. The PCM in particular (as opposed to the *generalized* partial credit model) is a variant of the Rasch model, and thus the discriminations of each item are assumed to be 1.

The PCM was developed by Masters in 1982 [102]. For the PCM, one assumes the maximum possible score can be achieved by going through steps of other score "categories" defined by the amount of partial credit. In the example provided in his 1982 paper, there is a math problem with 3 parts (see Figure 4.6). For this item, four possible scores are achievable: 0, 1, 2, and 3. These are score categories. From these score categories, there are 3 steps that can be taken to achieve the maximum score: going from score $0 \rightarrow 1$, from score $1 \rightarrow 2$, and from score $2 \rightarrow 3$. For each step in this progression, there is a difficulty associated with the step, which is not necessarily the same for each step. For example, going from $0 \rightarrow 1$ may be easier than going from $2 \rightarrow 3$, resulting in different difficulty values. Essentially, the PCM treats each "step" as a dichotomous item, which

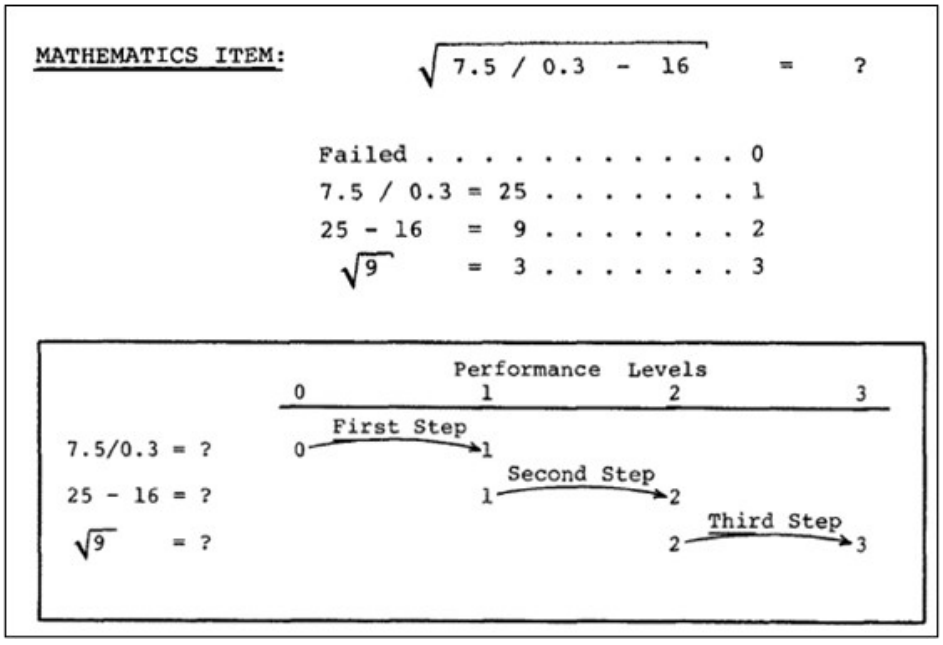


Figure 4.6: The mathematics item provided as an example from Masters' seminal 1982 paper [102] illustrating the partial credit model. As can be seen, progressing through each portion of the problem involves a steps going from one score to another, leading from lower to higher score, with a maximum possible score of 3 in this example.

can either be correct and progress a score to the next step (e.g., 1 $\rightarrow$ 2), or incorrect and leave a score at the latter step (e.g., remain at score of 1). Each of these steps is then modeled using the Rasch model (i.e., Equation 4.2), producing a difficulty value for each step. It is worth noting that the PCM may operate on the assumption that scores are ordinal, and thus build on each other. However, this is not the case for the U-STEP for all items, especially those composed of multiple MC-MR pairs (e.g., item 11) or multiple MC prompts (e.g., item 12).

Despite this possible restriction, exploratory PCM analyses were conducted for the 15 U-STEP items using the ltm package in R [103]. This analysis yeilded complicated results in many instances, due to the many number of score categories within different items. For example, item 11, which was composed of 3 MC-MR pairs and allowed for half-integer points for some reasoning selections, had 27 possible score categories. Thus, this item had 26 difficulty values, one for each score step. The lowest number of score steps was 1, which appeared for items composed of a single MC prompt.

Similar to Rasch, each *step* for PCM produces a single characteristic curve, which describes the probability of completing that step based on ability (described by Equation 4.2). An example plot of PCM item response category characteristic curves for the item presented in Figure 3.7 is shown in Figure 4.7. This particular example had 5 steps due to its 6 possible score categories: 0, 1, 2, 3, 4, and 5. Each curve respresents the probability of a student of a particular ability receiving the associated score of 1, 2, 3, 4, or 5. Notably, the curve for achieving a score of 5 closely matches the expected pattern described by Equation 4.2; higher-ability students have a higher probability of receiving the highest possible score. Similarly, the curve for a score of 1 takes a similar shape with opposite directionality; this indicates that lower-ability students have a higher probability of making the lowest-level step (0$\rightarrow$1). The curves representing scores of 2, 3, and 4 have peaks that gradually shift towards higher abilities, indicating that students need higher ability levels in order to achieve the progressively increasing scores.

The plot shown in Figure 4.7 suggests some conceptual alignment of responses to the model based on the gradual upward progression of score peaks with increasing score. However, an analysis of fit statistics indicates the model is not sufficient. The current scoring scheme leads to lower

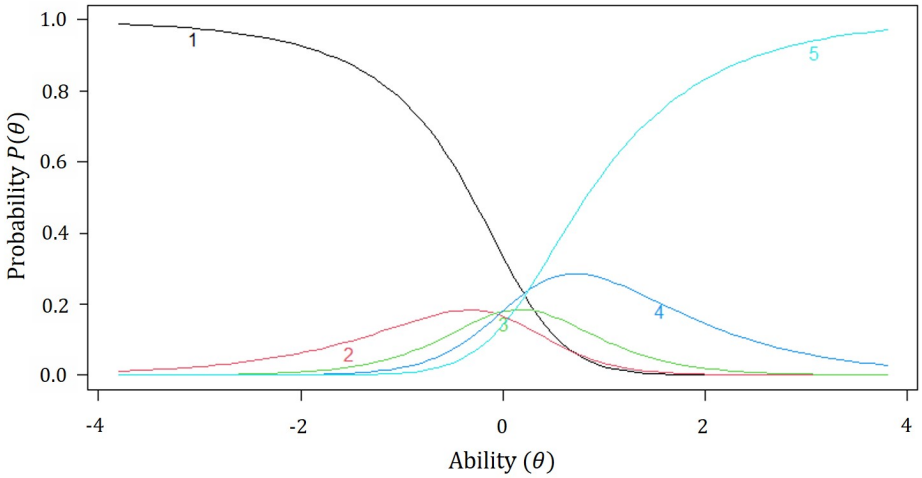


Figure 4.7: PCM item response category characteristic curves for the item presented in Figure 3.7. Each curve is labeled with its associated score, ranging from 1 to 5, and its shape is described by Equation 4.2. Higher-ability students had a higher probability of receiving a score of 5, with lower-ability students being less probable to receive that score. Similarly, lower-ability students have a higher probability of only receiving a score of 1, while higher-ability students have a lower probability of receiving that score.

N-values in each score category, largely due to complex items with multiple possible scores. Because of this, the fit statistics, which produced very large values of S-X^2 measures appearing for the overall assessment fit (i.e., S-$X^2 \gg 0$), cannot be accurately interpreted because an assumption of the model (i.e., sufficiently large-N in each bin, as is required for traditional Rasch analyses) is being violated.

To investigate the impact of reduced number of score categories for PCM analysis, item scoring was revisited such that possible scores for each item could take on only one of three possible values: 0, 1, and 2. The traditional (5pt) CMR items were scored such that 1 point was awarded for a correct MC selection and 1 point was awarded for a minimum of half-correct reasoning (i.e., if a student received at least 1.5 out of 3pts.). Pure MC items were scored as 0 or 1. Scoring for paired CMR items was more nuanced and varied by item. For example, if the MC prompts within a single item tested similar ideas (e.g., heat flow in vs. out) and associated reasoning for each prompt was similar, MC responses combined were scored as worth either 0 or 1 and combined reasoning was scored as either 0 or 1 (determined by a set threshold, depending on the item). Analysis of model

fit produced a high S-X^2 value, but was much lower than the output from the initial PCM analysis (which had more score categories). This suggests better fit, though it is likely still insufficient due to low-N in the sample. Literature suggests that use of the PCM model requires very large sample sizes (e.g., close to 1,000 respondents) [104].

Present results imply use of the PCM as explored here is not yet feasible for validation of the U-STEP, as our dataset is not large enough for a robust analysis using the methods employed thus far. PCM analysis with the original scoring scheme results in too many score categories, especially for complex items with multiple parts, making it impossible to produce valid fit statistics. Additionally, it is questionable if the reduced-category scoring approach (i.e., limiting each item to a maximum of three score categories – 0, 1, 2) is something we would want to pursue in the future. However, these analyses were helpful in narrowing down the possible IRT validation methods that could be applied to the U-STEP before it is finalized (dichotomized Rasch as determined by a threshold will likely be the final approach).

4.8 Assessment Validation: Summary

This chapter presented the overaching validation process for the Upper-level Statistical Mechanics and Thermodynamics Evaluation for Physics (U-STEP). A cornerstone of this process was pilot administrations of the CMR U-STEP in classes and interviews, which helped establish content and construct validity. Routes to establish statistical validity were discussed, including an inconclusive invetigation of criterion validity. Additionally, preliminary item response theory analyses using the Rasch model and partical credit model were presented. Though it is unclear as of now how precisely these methods will be utilized in completing the finalized U-STEP, they provide promising possibilities for validation of future CMR-based assessments.

Chapter 5

Discussion: Upper-Division Thermal Physics Assessment

Despite many available thermal physics assessments, only the assessment discussed here—the Upper-level Statistical Mechanics and Thermodynamics Evaluation for Physics (U-STEP)—addresses statistical mechanics in addition to classical thermodynamics. Though existing assessments have served as useful in identifying student alternate conceptions of thermal physics concepts, the U-STEP fills a gap that can allow instructors to gauge student understanding of statistical mechanics concepts in in additional to solely classical thermodynamics. Additionally, to the research team's current knowledge, the U-STEP is also the first *upper-division* thermal physics assessment.

The U-STEP is the first thermal physics assessment to be composed of coupled, multiple-response (CMR) items [7], though at least one two-tiered *multiple-choice* assessment exists [55]. Additionally, though the development process of the U-STEP aligned with traditional development routes, it also incorporated relatively rare techniques for PER assessment development such as utilization of differential item functioning *during* the development stage. In addition to this, the U-STEP is the first physics assessment that has explored Rasch validation for CMR items during the development phase.

5.1 Development and Validation

The process for item development and validation aligned with common techniques used for PER assessment validation [57], including classical test theory (CTT) analyses. However, the analyses also involved investigating gender- and race-based performance differences *during* the

development and validation process, which has traditionally only been conducted in PER after assessments have been finalized (ref. [6] provides an example of this). The validation process also adopted more novel approaches for analyzing physics assessments, such as Rasch analysis, *during* the development stage as opposed to post-hoc. Additionally, the U-STEP is the first CMR-based physics assessment to use Rasch approaches at *any* stage, during development or after finalization. The following sections summarize the item development and validation routes undertaken for the U-STEP and discuss implications for the presented work.

5.1.1 Item Development

The item development process for the U-STEP began by soliciting faculty input through a content survey, which was distributed to over 200 institutions. We received over 70 responses from across the United States, which were then used to inform the content included on the assessment. Utilizing results from this survey allowed for identification of key topical areas and texts that could lay the foundation for construction of the U-STEP. This was essential in the development process, as it allowed for many instructors' input into the content covered and notation used within the assessment. Additionally, this makes the assessment more likely to be useful for a broad range of instructors while remaining in a singular form. Results from the content survey can also be useful for other researchers and curriculum developers in the field of upper-division thermal physics.

After identifying key topical areas from the survey, assessment objectives that spanned the space of assessable content were developed to guide item development (see Appendix B). Items were written and iteratively revised based on results from student interviews, pilot administrations, and statistical analyses. Student interviews helped in establishing both content and construct validity. Results from the pilot administrations were used to determine item difficulty and discrimination. Additionally, we inspected student response patterns and conducted differential item functioning (DIF) analyses. Items that appeared too difficult or had poor discriminatory power were removed from the pool of possible assessment items or had multiple-response options revised. Correct multiple-response selections that were infrequently selected, or incorrect distractors that were

frequently selected, were revised or removed as well. Additionally, DIF analyses allowed for identification of potential bias, leading to more in-depth analyses of items that appeared to be problematic with regards to gender- or race-based performance differences (see Appendix F).

5.1.2 Item & Assessment Validation

A key component of the validation process was recruitment of piloting sites. We intentionally recruited piloting sites from a range of institutions, by both department size and student populations served. This means the population used for validation may be more diverse than those traditionally used for physics assessment analyses. Thus, the sample of students is closer to representative of upper-division physics students nationally than is typically used for assessment validation in PER; pilot sites are often large, predominantly-White departments due to the need for large N for validation purposes.

Validation analyses involved classical test theory (CTT), which included a DIF analysis in addition to traditional analysis that outputs item and test parameters. We argue DIF analyses are essential to the development process of assessments, as it assists in identification of differential performance on items, allowing for potential curbing of bias in the assessment. This consideration of fairness and mitigation of bias during the development process is key in reducing performance gaps caused by bias that could appear on the assessment in future administrations.

In addition to CTT analyses, preliminary item response theory (IRT) analyses, namely Rasch analyses, were conducted with the data as an additional validation method to lay the groundwork for future IRT validation once sufficient data has been aggregated. IRT analyses are beneficial over CTT analyses due to the limited population-dependence of IRT student ability measures. This means the item descriptive statistics are less fundamentally tied to the population used for validation, allowing for less ambiguous comparisons between student populations, either by demographic group or institution. This is helpful for instructors who wish to compare their students' performance to that of students at other institutions. Additionally, these comparisons become more reliable when race and gender have been considered during the assessment development and

validation process, especially for institutions with demographic compositions differ from the White male norm typically used in assessment validation.

Three thresholds for dichotomozing CMR data were considered and compared to lay the groundwork for validation of the U-STEP and future CMR assessments using IRT. Difficulty results from the Rasch analyses mapped well onto the results of CTT, such as both analyses ranking difficulty of items similarly, implying that the model outputs are behaving as expected. Additionally, model fit statistics showed results that suggest the dichotomizing methods used could be sufficient for modeling CMR data with the Rasch model. Due to CMR items' nature of allowing partial credit, an extension of the Rasch model—the partial credit model (PCM)—was explored. Though results initially seemed promising, the fit was not sufficient likely due, in part, to low-N. Additionally, the complexity of the PCM outputs make interpretation less straightforward, and may be challenging to convey to instructors. Thus, PCM may not be the ideal validation approach for the U-STEP or other CMR assessments at the upper-division level. This is especially true when considering the promising results produced from the dichotomous Rasch analyses.

These analyses have strong implications for validation of future assessments in PER, whether the format is CMR, multiple-response, or traditional multiple-choice. Rasch analyses, combined with DIF analyses, could change the way assessments in PER are developed and allow for minimization of performance differences due to item bias before they are implemented at large scales.

5.2 Future Work

The U-STEP has not yet been finalized, as it will require one additional pilot administration, which will occur in the Spring 2021 semester. This pilot administration will allow for preliminary IRT-based DIF analyses in addition to more robust CTT-based DIF methods. Approximately 18 piloting site have been recruited for this pilot administration, with a pool of over 200 students who may take the assessment. Since minimal changes have been between the Fall 2020 and Spring 2021 semesters, we will be able to pool data from both for analyses, allowing for more robust validation statistics using both CTT and Rasch analyses. Results from these analyses will be distributed

broadly. In addition to continuing the CTT and Rasch analyses for the larger data set, exploratory factor analysis will be conducted to robustly test unidimensionality of the U-STEP.

After finalization of the U-STEP, best methods for distributing the online-formatted assessment and encoded scoring scheme (to allow for streamlined scoring) will be considered. Additionally, automated methods for score reporting to instructors will be determined and implemented. After automation is finalized, the assessment will be widely publicized to instructors on PhysPort.

The U-STEP development lays the groundwork for another assessment that will address both upper-division thermal physics content *in addition to* scientific practices (the Thermal and Statistical Physics Assessment, TaSPA). The faculty content survey will inform content-coverage as well as focal scientific practices for this new assessment. The TaSPA will be constructed from complex CMR items [105] and provide descriptive feedback for instructors to inform pedaogical and curricular changes. This differs from the U-STEP in that the TaSPA will provide actionable feedback, as opposed to just items scores, and will explicitly focus on scientific practices, such as those defined by the Next Generation Science Standards [23]. The preliminary Rasch analysis presented here will be used to inform validation of the TaSPA once its CMR items are piloted. Additionally, the utilization of Rasch analysis for the TaSPA will allow avenues for creating a variable-content-coverage format, in which instructors will be able to select content and practices to include on the assessment administered in their class. This will make the TaSPA broadly applicable and useful for a variety of upper-division thermal physics courses and instructors.

5.3 Conclusions: Upper-Division Thermal Physics Assessment

The U-STEP is the first assessment of its kind, evaluating both upper-division thermodynamics and statistical mechanics content understanding with CMR items. The development and validation cycle contributes to newer approaches in PER for assessment construction, including DIF and Rasch analyses for validation. Implementing these strategies in assessment development can help curb bias and minimize performance gaps during future implementations of the assessment, facilitating fairness in testing and allowing for more reliable comparisons between institutions. The

U-STEP will set up the groundwork for construction of similar assessments in the future and could inform transformation of assessment development techniques for future work in PER.

Chapter 6

Background: Impacts of Race & Gender on Student Experiences in STEM

Part 2 of this dissertation describes a secondary project complementary to the work focused on thermal physics assessment. Results can be used to inform interpretation of performance differences on assessments despite attempts to address bias. This study focuses on various factors contributing to experiences and retention of underrepresented groups in STEM, particularly women and people of color.[1] Part 2 presents results from five core analyses: perceptions of STEM course environments (Sec. 7.1); perceptions of professor care (Sec. 7.2); student sense of belonging in STEM (Sec. 7.3); the intersections of course environments, professor care, and belonging (Sec. 7.4); and perceptions of race and gender impacts in STEM (Ch. 8). This work is meant to contribute to the large body of work that has investigated issues of underrepresentation in STEM fields, and was published in the International Journal of STEM Education [20, 21, 22].

6.1 Motivation

Women and people of color have been largely underrepresented in, and historically excluded from, most STEM fields, though recent trends are showing improvement [106, 107, 108, 10, 109]. This poses issues for meeting the growing needs in the STEM job sector, but also more general issues regarding access, inclusion, and equity. Reasons suggested for both race and gender underrepresentation span a range of factors including cultural norms, organizational structures, differential access

[1] The phrase "women and people of color" is used frequently throughout Part 2. This language has potential to exclude women of color, who belong to both groups. It is worth highlighting here that "women and people of color" is meant to encompass White women, women of color, and men of color. This phrase is used for simplicity.

to appropriate secondary school preparation, discrimination and harassment, and characteristics of individuals themselves. The goal of the study presented here is to investigate and understand some of the many factors contributing to underrepresentation of women and people of color in STEM fields. In the following section, key portions of prior research relevant to this work are presented.

6.2 Literature Review

There is a vast body of research on the differential experiences of women and students of color in STEM. Studies investigate a wide range of avenues for explaining and categorizing underrepresentation in STEM. Here, a subset of this body of literature is presented, focusing on work directly related to, and built upon, for this study.

6.2.1 Impact of Interactive Teaching on Learning and Persistence in STEM

In addition to factors mentioned above, many researchers have turned their attention to possible contributions of contemporary STEM pedagogy and curricula to the problem of underrepresentation in STEM. Poor-quality teaching has been found to be a concern of students who persist in or leave STEM majors [110], suggesting focusing on quality of teaching and instructional approaches has value in improving student outcomes. Additionally, there is robust body of research indicating that interactive teaching methods can improve the learning of all students regardless of gender or ethnicity (see, for example refs. [111, 112]), though the efficacy of these practices does depend on implementation [113, 114]. Attempts to uncover research on the differential impacts of interactive teaching approaches on women and underrepresented people of color, done as part of this study, yielded few studies.

The most extensive effort at addressing the impact of interactive vs. lecture-based teaching on underrepresented groups was undertaken by Madsen *et al.* [115]. They report on a meta-analysis of 26 published studies on the gender gap on concept inventories in physics. They found mixed results regarding impact of teaching style, with interactive methods having positive, negative, or neutral impacts on the gender gap depending on the study. Their results are consistent with the

sparse results in other fields; namely, many studies support multiple conclusions. Madsen *et al.* did not attempt to analyze the type of classroom beyond the self-reported status given by the authors of each study, as it was not possible due to the nature of the study (i.e., it was a review of literature and thus they did not have access to those data). It may be that the wide variation in type of teaching associated with the ideas about "interactive," "reformed," "research-based," or "student-centered" contributes to the mixed results.[2] In other words, there is a range of how "interactive" teaching is implemented and the specifics of the implementation impact the results. Their findings indicate that interactive teaching can improve learning outcomes for women but those results likely depend on the nature of the implementation.

One of the most robust studies of an implementation of a research-based, highly interactive pedagogy is reported by Beichner *et al.* [116]. They report results from a comparison of highly interactive studio-style classrooms to traditional lecture-based classrooms involving 15,000 students. They found all student demographic groups were more likely to pass the class in the studio style classrooms than in the lecture-based classrooms. Notably, they also found passing rates for women and underrepresented people of color were improved more than those of men and White students. While not directly considering persistence, increased passing rates imply a likelihood of increased persistence.

These results leave open the question of if and how research-based teaching styles may differentially impact students of different genders and races (including the portions of these populations that overlap; i.e., women of color). The analysis in Sec. 7.2 addresses this open question.

6.2.2 Impact of Professor Care on Students' Experiences in STEM

A relatively under-investigated area in the STEM literature is the impact of professor care on students. "Care" has a somewhat ambiguous definition across existing literature, but has ties to faculty qualities such as being approachable [117, 118], showing interest in and concern for

[2] In line with prior research in the field and this study's broad investigation of variations in teaching styles, in Sec. 7.2 the term "interactive" is used to refer to all of these pedagogical approaches.

students [118, 119], and being understanding and fair [120]. Though many university and college administrators report their faculty tend to care about students [118, 121], care is a difficult construct to measure.

Despite the difficulty in measuring and gauging professor care, general college-level studies have found that positive interpersonal relationships between faculty and students hold importance [122]; these studies have found associations between a students' sense of professor care with success measures such as grades, confidence, additional course enrollment in a discipline, and retention [117, 118, 122, 123]. The latter of these factors can be of particular interest for studies investigating representation of different groups of students, such as women and people of color, who tend to leave STEM at higher rates than men and White students [16]. However, there is a dearth of investigations into sense of professor care in STEM specifically, though some work has suggested feeling valued and cared for within a department has positive influences on STEM success [124]. The work presented in Sec. 7.2 investigates student perceptions of professor care within their department while pursuing a STEM degree.

6.2.3 Impact of Sense of Belonging in STEM

Students' sense of belonging in college, and STEM more specifically, span a vast body of literature and has been found to have significant impacts. Sense of belonging refers to "students' sense of being accepted, valued, included, and encouraged by others (teachers and peers) in the academic classroom setting and of feeling oneself to be an important part of the life and activity of the class" [125] (p. 80). Previous work indicates that sense of belonging in STEM has a significant impact on educational success and persistence, especially for women and students of color.

As Strayhorn summarized in his extensive review of research [15], sense of belonging is associated with academic achievement, retention, and persistence in college, and these impacts are frequently more pronounced for women and students of color. Recent studies have continued to confirm these findings [14, 17, 18]. Within STEM, both women and students of color, who are often stigmatized in STEM, have consistently reported less sense of belonging than men and White

students [14, 126, 127]. Socially stigmatized groups have been found to be more susceptible to belonging uncertainty [128], which arises when people feel unsure of their ability to "fit in" [14]. Such feelings may cause those students, especially women, to experience competing belonging from non-STEM fields, pulling them out of STEM [17]. Fear of confirming negative stereotypes of a group one belongs to (gender, race, etc.) can undermine performance and contribute to a lack of sense of belonging as well. This can give rise to feelings that people like them do not belong there. Additionally, external cues, such as low representation of one's group, can influence sense of belonging, particularly for women in male-dominated fields, such as most STEM disciplines [129, 130].

6.2.3.1 Factors Contributing to Sense of Belonging.

Peer interactions and interpersonal relationships have significant impacts on sense of belonging and are often seen as the most critical factor for overall sense of belonging [127]; these impacts are influenced by the quality and types of interactions students have within their STEM discipline [130]. For example, engagement in peer discussions outside of the classroom was shown to increase the likelihood of women persisting in STEM [131]. Women of color in particular are highly influenced by the presence of peer relationships and peer support overall [131, 132]. Even for high-achievers, minority students without such connections to faculty and peers tend to leave STEM [16].

Participating in social interactions has been linked to students' identity development as well, specifically in STEM [133]. Part of this science identity development for students involves building relationships not only with their peers, but also with their specific STEM discipline [134, 135]. Science identity is rooted in both private and public identification as someone being a "science person," e.g., whether students see themselves, and others see them, as this "kind of person" [136]. Science identity can be a predictor for higher grades and persistence in STEM fields, and this identity can be mediated by belonging in the field [137, 138]. Some argue that science identity is particularly important for minority students who contend with uncertainty in whether or not they belong in STEM, and that it can bolster this sense of belonging and persistence [16, 137, 139].

Students' perception of and confidence in their own understanding of STEM content and ability to succeed in the field has been associated with belonging as well. One could consider this in terms of self-efficacy, which refers to beliefs about one's ability to perform in their field and the extent to which they have control over what happens in their future and success [140, 141]. Studies show women tend to have lower sense of self-efficacy beliefs about themselves in STEM [142, 143], aligning with women's commonly lower senses of belonging in STEM fields [141, 144].

6.2.4 Impact of Race & Gender on Students' Experiences in STEM

As part of the study presented here, we asked students if the experience of being a STEM major varies for different races and genders. A vast body of research suggests the answer to this question is "yes." Students are significantly impacted by both their race and gender as they pursue STEM. For example, in addition to lower sense of belonging and performance "gaps" compared to men and White students, research indicates that both women and people of color are more likely to report discrimination, microaggressions, and harassment in STEM [145, 146]. These experiences have been found to affect the physical and mental health of women and people of color, their ability to thrive in STEM fields, and their willingness and ability to continue in STEM [145, 147, 148].

It is important to note that the question asked to students in the interviews (presented in Ch. 8) is not about their opinion on these matters but their *perceived experiences* with racialized or gendered issues in STEM. By the time students are seniors in college, they are highly likely to have had the opportunity to experience and or witness impacts of both race and gender is STEM. Thus, the presented analysis seeks to answer the question: to what extent are students aware of these impacts and how does awareness differ by demographic characteristics of the interviewee?

6.2.4.1 Differential Race- & Gender-Based Experiences

Though much is known about race- and gender-based issues in STEM experiences, less is known about *perceptions* of both dominant and underrepresented groups regarding race and gender in STEM. This section provides information supplemental to the literature presented thus far to

describe what is known about general trends regarding perceptions of race and gender in the broader context of the United States. Little literature exists regarding perceptions of the race- and gender-based differences in experiences in STEM, especially in college settings. The presented findings are used as a grounding for the investigation of perceptions of race and gender in STEM-specific domains presented in Ch. 8.

Generally, studies of perceptions of the impacts of race and gender in STEM report on the perceptions of a marginalized group only (i.e., women's perceptions of gender discrimination in STEM or students of color's experiences in STEM that are impacted by their race). Studies comparing the perceptions of the same environment across *multiple* genders and races are relatively rare, and it is difficult to find studies regardings students' perceptions in the academic literature, especially regarding the perceptions of students from dominant groups in STEM. Studies regarding students' *own* experiences in STEM are more prevalent, but there are limited studies regarding students' awareness of the experiences of others that are not in their same demographic group.

The Pew Research Center has conducted polls regarding perceptions of race- and gender-based differences in the United States [149]. These polls consistently show White people perceiving fewer race impacts than people of color and men perceiving fewer gender impacts than women. For example, a recent poll of STEM professionals about their workplace environment [149] reports consistent gaps between the perceptions of discrimination between women and men. The researchers found that 38% of women in male-dominated STEM fields felt women are usually treated fairly in their workplace in opportunities for promotion and advancement compared to 78% of men. Likewise, they report a gap among races. In response to a question about fairness in promotion and advancement for Black people, 37% of Black people reported fairness compared to 75% of White people. The same poll also found that reports of discrimination in STEM are higher than in non-STEM fields. Women in male-dominated STEM fields reported experiencing discrimination at work more than women in non-STEM jobs (50% vs. 41%) and Black people in STEM fields reported experiences of workplace discrimination due to race more than Black people in non-STEM jobs (62% vs. 50%).

As described here, there is a pattern of evidence that more privileged groups fail to *recognize* their own privilege. There are also indications that, despite evidence to the contrary, privileged groups may believe they *themselves* are the disadvantaged group. This is presumably in reaction to policies and practices like affirmative action that acknowledge and attempt to explicitly address inequity. For example, a recent study demonstrated that on average White people in the US believe anti-White bias is more prevalent today than anti-Black bias [150]. That same study offered evidence that White people believe reverse-racism has increased over time. In general, these perceptions may be the result of a lack of contact with members of other groups or a result of White people feeling threatened by diverse contexts or wishing to protect their position of privilege [151]. Perceived threats may also be psychological, including the suggestion that their position has not been the result of their own merit or that they are part of a group that has unfair advantages.

While research consistently finds that members of marginalized groups are more likely to recognize the impacts of their demographic status on life chances than those of the privileged group, studies such as the Pew Research Center poll described above [149] also find a significant proportion of the marginalized group that does not report an experience with discrimination. This discrepancy has been noted and theorized by researchers across varying fields.

One explanation for the disconnect between perception and experiences is the denial of personal disadvantage. In Crosby's foundational paper on the denial of personal disadvantage [152], she identifies the discrepancy between women's acknowledgment of discrimination in general but denial that they are personally discriminated against, despite evidence to the contrary. Crosby proposes two barriers preventing women from acknowledging their own disadvantage. First, she suggests the cognitive bias associated with analysis of large numbers vs. individual cases obscures discrimination of an individual. Secondly, she suggests the emotional discomfort people experience when confronting their own victimization as a barrier for acknowledging disadvantage.

Another notable explanation for marginalized groups not recognizing their disadvantage is system justification theory. Jost *et. al.* review system justification theory [153], which offers insights into why people justify the status quo even when it is against their own self-interest. They

suggest people have a tendency to support and defend existing social structures due to a need to view themselves and their group positively. They argue members of marginalized groups engage in system-justifying behavior because such engagement serves a palliative function to minimize the harm they encounter in circumstances they cannot change.

6.2.5 Missing from Current Literature

The study presented in Part 2 aims to fill gaps in the current literature regarding race- and gender-based experiences in STEM. Specifically, it aims to contribute more to the literature by analyzing different aspects of STEM experiences, including an intersectional approach that includes women of color, men of color, White women, and White men as separate groups. The study includes investigating student perceptions of instructional methods; perceptions of professor care in STEM contexts; sense of belonging in STEM; and perceptions of race- and gender-based differences in STEM experiences. Though literature exists in some of these domains, few studies address these areas in the same way this study does, as the presented study focuses on student experiences while explicitly analyzing by intersectional race and gender cohorts.

As an example of this phenomenon, many studies regarding gender differences in belonging do not investigate racial identity in conjunction, as many studies are done at predominately-White institutions and people of color are notoriously underrepresented across STEM disciplines. Thus, often studies produce data about "women" that is only true for White women, leaving the experiences of women of color in STEM classrooms largely unexamined. There are a few studies that look at the experiences of women of color regarding belonging, i.e., [127, 154], but this body of literature remains largely underdeveloped. Other studies about belonging in STEM fields that discuss racial identity often omit a gender analysis, additionally excluding women of color from the literature. There exist other intersectional works that investigate women of color's experiences, e.g. [155], but it is difficult to find literature that focuses on women of color's sense of belonging in STEM. A look into current literature resulted in little-to-no studies taking on an intersectional analysis approach when investigating instruction style and professor care.

In addition to the lack of intersectional analyses along the lines of race and gender, studies often investigate instruction style and race/gender differences through the lens of researchers and practitioners, as opposed to soliciting student views and perspectives of their experiences. Further, when looking into the current literature, no studies investigating profesor care in STEM were found. The work presented in Part 2 of this dissertation fills a space in the literature that is currently lacking.

6.3 Key Terms

Throughout Part 2, several terms not commonly invoked in traditional STEM literature and/or that have differing meanings in colloquial language are used. To avoid confusion about these terms and their uses, specific definitions for them are presented below.

Students of Color and Underrepresented Minority Students. Throughout Part 2, two similar, but different, terms used enough that they deserve explicit discussion here: students of color and underrepresented minority (URM) students. *Students of color* refers to students of any race that is not White, including Asian, Black, Hispanic, Native American, and multiracial. On the other hand, the National Science Foundation defines *URM* races as minority groups underrepresented in STEM, including Black, Hispanic, and Native American (Asian is excluded from this categorization) [10]. In the presented study, multiracial is also included in the URM category. It is worthy of note that the term "URM" has been critiqued in recent years, largely due to the implications of combining multiple races into a single category, such as erasing the unique experiences of certain races [156]. This can be viewed as a racist tool of oppression, as it places Whiteness as a norm to be compared to (e.g., use of "minority" in comparison to numbers of White people, as opposed to referring to individual racial categories on their own) [156].

Due to low-N in the study sample across different racial groups, and emergent commonalities within multple racial categories, URM and students of color are included as singular, distinct groups throughout Part 2. The terms "students of color" and "URM" are not perfect terms, and critiques of their use are well-founded. Unfortunately, there are limited alternatives to these terms' use in

the literature that can encompass students of many different races who are observed to encounter similar experiences. The use of the term URM does lead to limitations of the presented work, and conclusions should be interpreted with caution—the intent is not to say that all students within the URM and students of color categories have the same experiences or are a homogeneous group; readers should be cognizant of this when reading and cautious in their takeaways.

Intersectionality. A major aspect of this study is an intersectional approach to data analysis. Intersectionality refers to the idea that aspects of one's identity (e.g. race, gender, class, sexual orientation) are not unitary and mutually exclusive, but instead interact to construct one's identity and the way one experiences oppression [157, 158]. Here, "intersectionality" is used to refer to the intersections of identities and investigate how those complex identities influence students' experiences in STEM. Considering race and gender as single axes of identity without analysis of the intersections of those groups can lead to the erasure of some identities, such as for women of color [157, 159]. In the study presented in Part 2, two axes of identity—race and gender—and the intersections of these identities are considered: women of color, men of color, White women, and White men. Though there are other axes that could be considered, these are the only axes of identity recorded for the study, which is a limitation. However, the intersectional approach taken here will avoid the erasure of some students' complex identities (e.g., those of women of color).

Privilege, Power, Prejudice, Sexism, and Racism. *Privilege* refers to a set of unearned advantages, such as access to resources or social power, available only to certain people because of their membership in a social group [160]. Privilege can exist among many dimensions (race, gender, religion, age, sexual orientation, etc.) and varies between societies and cultures. Having privilege does not mean someone did not have to work for their success or never experienced adversity, nor does it mean someone is intentionally oppressing people belonging to other groups. Instead, it refers to unintended systemic advantages certain groups experience that simultaneously disadvantage others without these ascribed characteristics. People with privilege do not necessarily have things easier because of their race or gender, but instead do not experience additional hurdles or oppression due to their race or gender.

Those who have privilege are considered to hold more *power* in society. Power typically manifests from social hierachies, and the amount of power people hold can vary depending on context. For example, a female faculty member in physics may be in a position of less power within her department compared to male faculty, but holds more power compared to students in the context of a classroom in which she is the instructor. However, it should be acknowledged that complex power dynamics can come into play when a female faculty member (who holds more institutional power) interacts with a male student (who may be considered to hold more power in society outside of academia).

Prejudice is the conscious or unconscious thoughts, beliefs, assumptions, and cultural stereotypes of individuals about the superiority or inferiority of certain groups that underlie discrimination for or against people in those groups. Under this definition, prejudice is an individual's set of beliefs about people who share characteristics of the "other" group (or their own favored group), and may be the basis of discriminatory behaviors. *Sexism and racism*, on the other hand, refer to *systems of disadvantage* based on gender and race. In Part 2 of this dissertation, racism and sexism are based on a *prejudice plus power* definition. References to sexism and racism are different from prejudice in that they go beyond simple bias by individuals and instead are embedded in the power structures of society. It is worth emphasizing that one can uphold sexist and racist norms and values unintentionally; it is possible to perpetuate racist and sexist ideologies without holding any personal beliefs about the inferiority of other races or genders, i.e., without being prejudiced. For an in-depth discussion about this phenomenon with regards to race, see ref. [161].

Consider the following examples that illustrate these terms. An individual man holding the belief that women are less mathematically able than men is *prejudice*. Utilizing standardized tests constructed with items that overestimate the mathematical ability of men while underestimating the mathematical ability of women to determine admission to college or graduate school is an example systemic or institutional *sexism*. This thereby confers *privilege* onto men as they are more likely to gain admission with less merit. For an example of this phenomenon, see ref. [162].

Marginalization. Marginalization is tied to structures of privilege and power. The concept refers to the relegation of people from certain groups to the *margins* of a society's organizations, institutions, and cultural system by denying them an active voice, identity, or place within the given social context [163]. People from these excluded groups can be described as marginalized in these contexts. For example, women and people of color can be considered marginalized in STEM, due to their underrepresentation and relative lack of power in the fields compared to White men.

6.4 The Roots of STEM Success Project

Part 2 of this dissertation reports findings from interviews collected as part of a larger mixed method study, the Roots of STEM Success Project (https://clas-pages.uncc.edu/rootsofstem/). Other analyses using the same dataset may be found in refs. [154, 164, 165].

The Roots of STEM Success Project was designed to study experiences in STEM of students who are traditionally not represented in STEM fields, namely women and students of color. The Roots of STEM Success Project includes a large quantitative dataset with administrative data from middle school through college graduation of students who graduated from North Carolina high schools in 2004 and matriculated to one of the 16 campuses of the University of North Carolina (UNC) system. Many of these data are related to STEM success. In addition to the quantitative data, in 2013 the Roots of STEM Success Project conducted more than 300 interviews with college seniors who were asked to reflect upon their family, community, secondary school, and college experiences related to their decisions in choosing their college majors.

Prospective interviewees were identified by distributing an email recruitment survey to seniors at all 16 UNC campuses in January 2013. On these surveys, students could designate up to three major fields of study. Based on student responses, 201 respondents were catergorized as STEM majors (those currently majoring in STEM) or STEM leavers (those who began college in a STEM major but later elected to switch to a non-STEM major). The remainder of interviewed students were considered STEM avoiders, who expressed interest in STEM in college but never pursued it. STEM major are defined using the National Science Foundation Advance Program cat-

egorization (http://www.nsf.gov/crssprgm/advance/index.jsp) where majors such as engineering, physical sciences, earth, atmospheric or ocean sciences, mathematical and computer sciences, and biological and agricultural sciences are considered to be within the STEM category. A student was considered a "major" if their current major, at the time of the interview, fell within these STEM fields. Social sciences were excluded from the definition of STEM fields because of the study's focus on underrepresentation of women.

The sample was restricted to students who attended public school (K-12) in North Carolina and who were younger than 30 years of age. These selection criteria were designed to align the interview sample with the quantitative data in the larger project. Students were asked to identify their racial/ethnic group, as well as their gender, on the survey. Once the potential interviewees were identified, researchers on the project reached out to them via email to set up an interview (either by Skype, phone, or in person). They attempted to match interviewers with interviewees according to gender and race, although that was not always possible. The interviews lasted between 30 minutes and 1 hour and were recorded for transcription purposes. Students were paid US$25 for participation in the interview.

6.4.1 Interview Sample

In this study, students underrepresented in STEM were oversampled, and thus the sample is not representative of STEM students in either race or gender. However, racial demographics of the interviewed sample closely mirror the demographics of the geographic region in which the interviews took place. (Note the UNC system is composed of approximately 40% students of color.) Of all students interviewed, 66% identified as female and 34% identified as male.[3] Additionally, 48% of the students identified as White and 52% identified as students of color: 31% Black, 8% Asian, 7% Hispanic, 2% multiracial, and 1% Native American. Self-identified race is used when presenting findings of this study.[4] All participants were seniors about to graduate with a degree.

[3] Gender is referred to here as male and female. Gender identity of students was indicated as male or female on the survey; "male" and "female" are used as genders in this study to respect the identities that students self-reported.

[4] Demographic information was collected for the Roots of STEM Success Project in 2013.

6.4.1.1 A Note on N-Values in Various Analyses

Given the interest in intersectionality and how it helps to better understand STEM outcomes, samples are disaggregated by race, gender, and status (major vs. leavers) in some analyses. This resulted in some categories with very small numbers (e.g., the sample includes only four Hispanic women who were majors). Therefore, in many instances categories were collapsed in order to report findings in meaningful ways. When looking at responses across different racial and ethnic groups, similar patterns emerged across different groups. For instance, in Sec. 7.2 the responses of Hispanic students and Black students looked more similar compared to the responses of White students, and thus Hispanic and Black students were included in the same group (i.e., "students of color"). Additionally, the unique position of Asian students, who are both students of color and well-represented in STEM was carefully considered. Asian students' placement in analysis groups was done such that they were placed in the group their responses most closely matched.

Due to low numbers when parsing out by certain identities, not all analyses presented in the following chapters are intersectional in the same manner. These analyses are included because the partial-intersectional approach helps avoid the erasure of some students' complex identities, such as those of women of color, by allowing the reader to see both gender and race analyses. Though there are other axes beyond race and gender that could be considered, these are the only axes of identity recorded for this study. Additionally, due to variations in individual interviews, some questions were not asked in all interviews or received uncodable responses. Thus, the N-values presented are not the same for each analysis.

6.4.2 Interview Protocol

The interview protocol was developed specifically for the Roots of STEM Success Project. Questions were designed to elicit a recounting of the participants' history with STEM and factors that influenced their decision to major in STEM. Interviewees were asked about family and peer influences, childhood informal educational experiences, secondary school and college experiences in

and out of the classroom, beliefs about the self, attitudes toward STEM, and the students' reasons for pursuing or not pursuing a STEM major. The analyses presented in the following chapters are based on a subset of questions asked in the interviews. The presentation of those questions is reserved for their corresponding sections.

6.4.3 Interview Analysis

Conventional procedures were used for analyzing the qualitative data captured in the interviews. Using a partial-grounded theory approach to qualitative data analysis [166], two researchers independently read the responses and coded under broad categories, some of which were determined a priori and others that emerged from the data. The researchers then compared their independently created codes. Through discussion, the codes were reorganized, collapsed, and expanded in an iterative process until a coding scheme for each set of interview questions was developed. Using the final coding schemes all interviews were coded again by two researchers and compared; discrepancies were aligned through discussion. Analysis and interpretation was done primarily by White women, with social and physical science backgrounds.

Due to the qualitative nature of the data, as well as the number of comparison groups (often with low-N), statistical analyses for significance are not presented for these studies. Statistical analyses were only performed for some of the presented studies (i.e., data from Ch. 7), and these analyses are not presented for the purpose of consistency of presented results. The interested reader can find statistical analyses for a subset of the presented work in ref. [21]. Due to the lack of statistical analyses, results should be interpreted with caution, as they may be true of the sampled population but not necessarily generalizable to all genders or races.

Chapter 7

Instruction Style, Professor Care, and Sense of Belonging in STEM

This chapter presents analyses of three sets of questions in the Roots of STEM Success Project interviews. These analyses focus on: (1) perceived and preferred instruction style (Sec. 7.1); (2) perceptions of professor care (Sec. 7.2); and (3) student sense of belonging in STEM (Sec. 7.3). An analysis of the intersections of these factors is also presented (Sec. 7.4). This work was published in first-author manuscripts in the International Journal of STEM Education [20, 21].

7.1 STEM Course Environments

Interview analysis investigated students' perceptions and preferences of instructional styles in their STEM courses. Findings regarding instructional style of STEM courses are based on responses to two questions from the interview protocol:

(1) *To what extent did your math and science teachers lecture vs. use more active approaches such as, encouraging student discussion, cooperative learning, and hands on activities?*

(2) *Would you have preferred a different emphasis?*

Due to low-N for intersectional race-gender cohorts, results are presented for underrepresented minority (URM)[1] leavers and majors; White leavers and majors; female leavers; and male leavers. Due to their complex identities as students of color who are well-represented in STEM (e.g., neither

[1] See Sec. 6.3 for discussion on use of this term.

White nor URM), and relatively low numbers in our sample, Asian students are not included in the analysis presented in this section.

Responses to these questions were coded based on the interviewees' reported instruction style of their STEM professors and students' stated preference for instruction style in their STEM courses. Student responses were coded in three ways: (1) interactive instructional style, (2) lecture-based instructional style, and (3) mixed instructional style. Responses indicating a mixed instructional style meant that some interactive methods were reported, but lecture was also highly present in the course. For the most part, when a student was coded for having mixed experiences they reported that the main class was lecture-based but there was a lab component that was interactive. For example, one student stated,

> *"You have the class and then you have the lab. The lab is the place for you to work hands-on with your classmates or whatever and then class is just lecture."*
> – Black female leaver

Sometimes a student reported different experiences in different courses. For example, one student described her chemistry class:

> *"In the entry level chemistry courses where you have a larger class, most of the time the teacher ends up lecturing. But, once you get into the junior and senior level courses it's no longer so much concentrated on lecturing as it is on, 'here's some problems, this is the basic outline and how you should be thinking about these problems, now work together or work alone and help each other figure out these problems.'"* – White female major

In this situation her response was coded as both interactive and lecture-based, due to her describing two separate courses with different instruction styles.

Students' preferences of instruction styles were coded the same as above, noting whether students desired a different emphasis or were comfortable with the reported approach to instruction. It is worth noting that all responses are based on student perceptions of interactivity and may not align with the instruction style professors thought they employed or with the perceptions of other students in the same course.

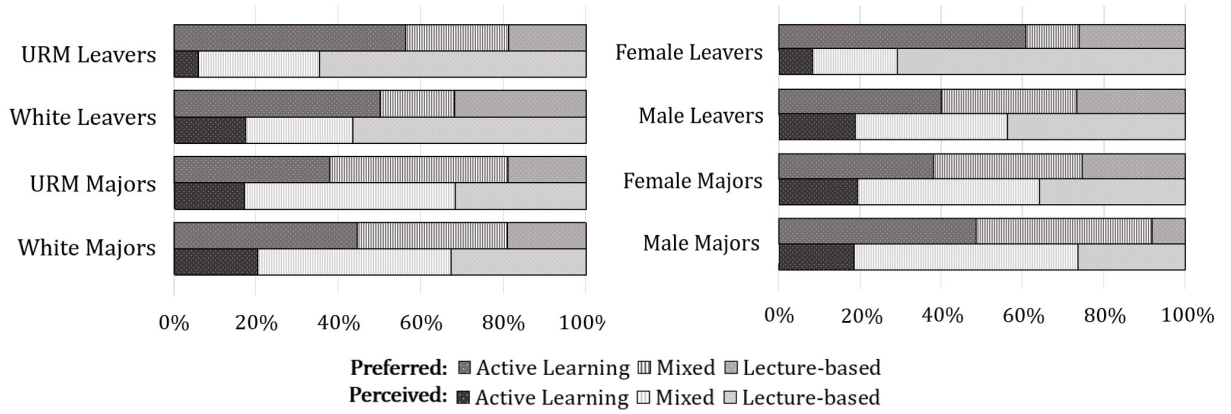


Figure 7.1: Preferred and perceived instruction styles by race, gender, and major status. Mixed instruction styles refer to instruction styles implementing combinations of both active learning and lecture-based approaches. N-values for each groups are as follows for preference and perception, respectively: URM Leavers (16, 17), White Leavers (22, 23), URM Majors (37, 41), White Majors (63, 64), Female Leavers (23, 24), Male Leavers (15, 16), Female Majors (63, 67), and Male Majors (37, 38).

7.1.1 Perceived vs. Preferred Instruction Style

Results for perceptions and preferences of instruction style for both gender and race are shown in Figure 7.1. Interactive instruction was the least frequently reported instruction style for all groups. Female leavers reported the lowest levels of interactive instruction. In contrast, female leavers reported lecture at higher rates than other groups. Male leavers and female majors reported similar instruction styles. Notably, male majors reported encountering mixed instruction styles more frequently than lecture-based instruction, whereas female majors did not have a significant difference in frequency of mixed or lecture-based instruction.

As can be seen in Figure 7.1, interactive instruction was the least frequently reported instruction style encountered for all racial groups. URM leavers reported encountering interactive instruction at lower rates than other groups. In contrast, the number of students reporting lecture-based instruction was much larger than that of interactive teaching. The responses of leavers diverge substantially from those of majors for both mixed and lecture-based instruction. URM leavers were more likely than other groups to encounter lecture-based instruction, though White

and URM leavers reported lecture at comparable rates.

Of note, students generally report preferring more interactive-based teaching than they report encountering. Women leavers stand out among those with the greatest preference for interactive methods compared to other groups, despite reporting encountering that method in their STEM courses the least. As seen in Figure 7.1, White and URM majors had similar preferences for instruction style. Leavers, however, did not. Notably, URM leavers report the greatest preference for interactive teaching compared to other groups, whereas White leavers had the greatest preference for lecture-based instruction. Interactive approaches to instruction are the most popular while lecture-based instruction is the least favored approach among all students.

7.1.2 Summary of Instruction Style

The presented data shows a large discrepancy between preferred instruction style and reported instruction style. In particular, preferences for instruction style lean towards more interactive approaches. It is worth noting this preference for more interactive teaching is present across all demographic groups. Female leavers and URM leavers were the most likely to prefer interactive instruction styles but were the least likely to report experiencing it. This finding suggests that students from underrepresented groups who are leaving may have been affected by a mismatch in their preferred instruction style and the instruction styles they perceived in STEM courses.

Because these data are qualitative, direct correlation between instruction style and persistence cannot be concluded. However, another study in STEM relating perceived instruction style to persistence, which looked at calculus students, concluded that those who do not persist are more likely to perceive their classrooms to be less interactive than students in the same classroom who persisted [167]. An additional study investigating students persistence in STEM found pedagogical approaches of STEM instructors was a factor in students' decisions to leave STEM majors [110]. These findings are consistent with those presented here. Leavers who report lecture-based instruction may not have experienced more lectures than their counterparts who report more interactive classrooms, but leavers perceive their classrooms to be less interactive. The data presented here

indicate this is especially the case for underrepresented groups. Leavers from underrepresented groups report a greater preference for interactivity than majors of all demographics, but it cannot be concluded that, compared to other leavers or majors, they experienced different classroom pedagogy. These findings suggest potential positive impacts on the retention of underrepresented groups by reforming teaching so it is more interactive. This is an area of potential importance for which more research is needed.

7.2 Perceptions of Professor Care

In the interviews, participants were asked:

(1) *Do you think your {major or dropped major} instructors cared about you and your learning?*

Answers were coded broadly as either "instructor cared" or "instructor did not care." This section simply reports on students' perceptions of professor care and does not present sub-coding based on reasons given. Students who reported both having professors who cared and professors who did not care were counted twice, once for each response. For example, a math major may report that they had a calculus instructor who cared about their learning and a linear algebra instructor who did not care about their learning. Their responses would be counted twice, once for each code. Findings are reported based on race and gender, as well as by the race and type of STEM field (i.e., biological sciences versus physical sciences).[2]

The analysis presented in this section includes responses from 135 majors and 38 leavers. The data show majors were more likely than leavers to report a professor who cared about their learning, as shown in Figure 7.2.

7.2.1 Gender, Race, Representation Status, & Professor Care

Analyses based on respondents' gender and race are presented in Figure 7.3, for both majors and leavers. White women majoring in STEM were the most likely to report feeling cared about

[2] These fields are disaggregated due to the differential representation of women in the field. In biological sciences, women are well-represented, whereas this is not the case in most physical science fields [10]

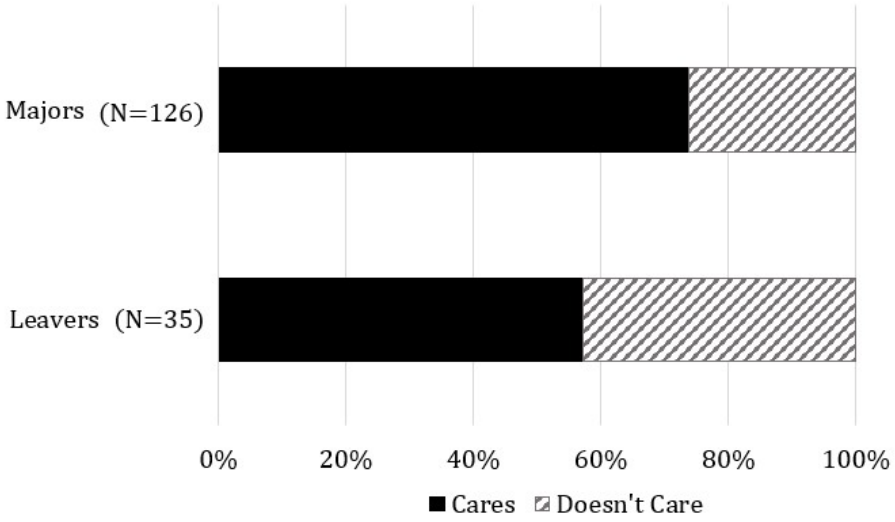


Figure 7.2: Perceived professor care among majors and leavers.

by their instructors while women of color (both majors and leavers) were the least likely. Notably, women of color majoring in STEM report less care than the White women who dropped their STEM major. Men's responses, in contrast, did not vary significantly based on race. Nearly identical results are seen for both White men and men of color majoring in STEM, with only a small racial difference among male leavers. The findings indicate small variations in perceived professor care across the gender by race cohorts. All leavers tend to perceive less care than majors

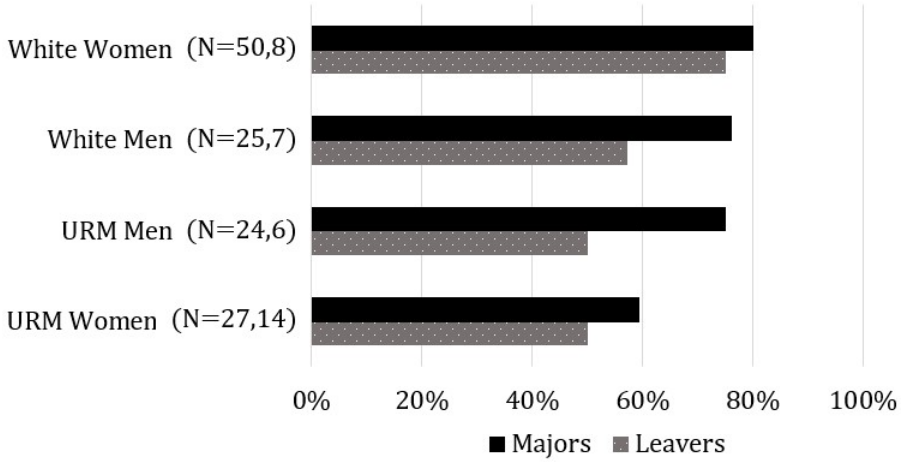


Figure 7.3: Perceived professor care by major status, race, and gender. N-values are report for majors and leavers, respectively.

of the same race and gender, and White female leavers reported as much care as male majors of any race. Women of color, whether majors or leavers, report less professor care than any other cohort.

Perceptions of professor care were also analyzed by discipline. As can be seen in Figure 7.4, students in physical science (pSTEM) fields were less likely than those in biological science fields to feel their professors cared about their learning. It should be noted that these students' responses were frequently about STEM instructors in general, and does not necessarily refer to perceptions of care solely within the major itself.

7.2.2 Summary of Professor Care

Most students perceive that their professors care about their learning. However, across all demographic groups, majors are more likely than leavers to report feeling their professors cared about their learning or them personally. Among all majors, 75% reported feeling their STEM professors cared about them and their learning. While it is not surprising that majors reported higher levels of care than leavers, note that all of these students were seniors nearing graduation

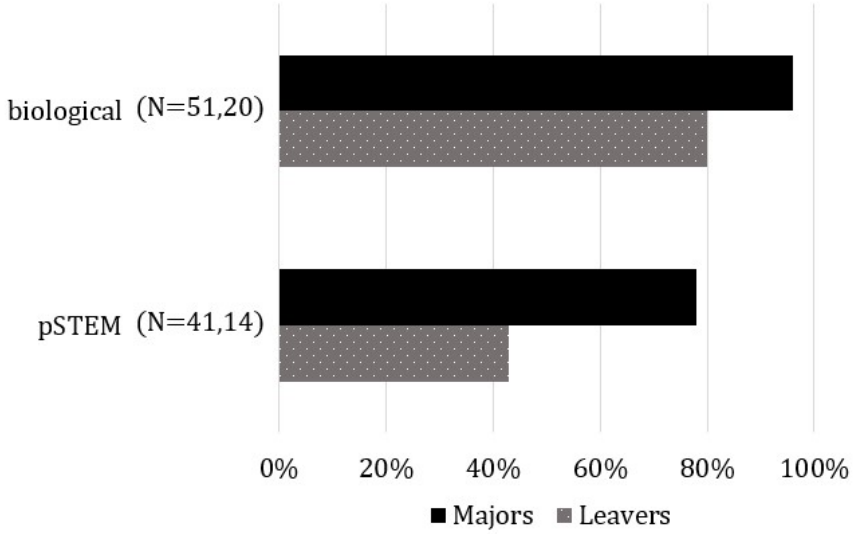


Figure 7.4: Perceived professor care by field: biological sciences and physical sciences (pSTEM). N-values are report for majors and leavers, respectively.

in their field. However, nearly 25% of students about to graduate with a STEM degree reported feeling their STEM professors did not care about their learning, which raises an issue of concern to be investigated further. While the results for White men and men of color were nearly identical, White women reported the most care while women of color reported the least. Only 60% of the women of color in this sample reported feeling their professors cared about them, despite being seniors about to receive their degrees in a STEM field.

Reasons given by women of color for not feeling their professor cared varied. A common reason was poor or distant teaching. For example, one leaver stated:

> *"I felt like they were there just to teach us what we were supposed to know and if we didn't grasp it then it wasn't their problem. I would get outside tutoring help and they would try to help me out as best as I could but because my teacher wasn't teaching the basics it was just hard for me to grasp onto the concepts."*
> – Black female leaver

Large class sizes were also a common reason students gave for feeling lack of care. One major noted:

> *"I didn't feel they [cared] because there are so many kids in the class. I think it's hard to really connect with your teachers."* – Multiracial female major

Unfriendly interactions with individual teachers were also frequently mentioned. For example, one women told a long and involved story of feeling unsupported by her professor after returning to school after a family death.

The extent to which students, particularly women of color, feel their professors care about them and their learning may influence their persistence. This analysis is not meant to question whether the faculty care about their students and their learning, and analyzing the literature provided no research-based evidence that faculty do not care. However, these results suggest that many students do not feel this care. It is notable that it is not unusual for introductory science courses to have large enrollments where there is little opportunity for personal contact between the professor and students, and it is likely many leavers only encountered these courses in their STEM trajectories.

7.3 STEM Students' Sense of Belonging

This section presents findings focusing on responses that were coded broadly as a student feels they "belong in STEM" or "does not belong in STEM." These codes were applied to any part of an interview in which a student made reference to belonging. Typically, sections coded as one of these two codes appeared as an answer to one or more of the three interview questions asked consecutively in the interview protocol:

(1) *Do you feel like you belong/belonged in {your STEM major}?*

(2) *Did you ever feel out of place?*

(3) *Has this feeling changed over time, and if so, what led to these changes?*

Note that here "{your STEM major}" is used as a placeholder—in the interviews this was replaced with students' current STEM major (for majors) or previous STEM major (for leavers). Students reported either feeling they belonged or did not belong; others reported mixed feelings of belonging.

7.3.1 Race & Gender Impacts on Students' Sense of Belonging

This section reports student responses for sense of belonging in their STEM major by gender, race, and representation of demographic group in their STEM major (i.e., biological sciences vs. pSTEM). Students frequently gave multiple reasons for belonging or not belonging. Therefore students' overall belonging status was coded as: (1) belongs in STEM (positive belonging status, only reported feelings of belonging); (2) does not belong in STEM (negative belonging status, only reported feelings of not belonging); or (3) mixed (reported both feelings of belonging and not belonging). This scheme permitted coding all students uniquely into one of the three categories.

Not all students who reported their belonging status gave explanations for their sense of belonging while some gave multiple reasons. Because of this pattern, there is a discrepancy between the total number reported for belonging statuses and number of belonging explanations.

7.3.1.1 Effects of Race on Belonging

Results of the race analysis for both majors and leavers are shown in Figure 7.5. As mentioned in Sec. 6.4.3, there was a need to consolidate racial categories due to small numbers when dividing among many different groups. Here, three racial groups are considered: White students, Asian students, and URM students. We report Asian students separately because while they are an ethnic minority, like URM students, they are generally overrepresented in STEM in proportion to their share of the population, like White students [10].

Among majors, there were responses from 70 White students, 11 Asian students, and 50 URM students. Figure 7.5 shows that students of color (i.e., both Asian and URM students) who major in STEM are less likely to report a sense of belonging than White majors. As can be seen from responses of the 24 White students, 5 Asian students, and 33 URM students who were leavers, a large fraction of leavers, regardless of demographic group, reported they did not have a sense of belonging in their STEM major.

7.3.1.2 Effects of Gender on Belonging

Results of the gender analysis for both majors and leavers are shown in Figure 7.6. A total of 131 analyzable responses from majors were coded (51 men and 80 women). It can be seen in

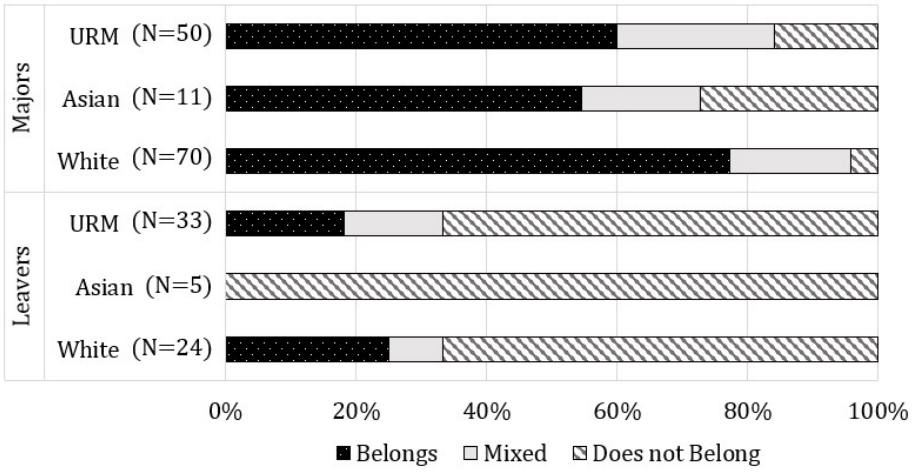


Figure 7.5: Student sense of belonging in STEM by race and major status.

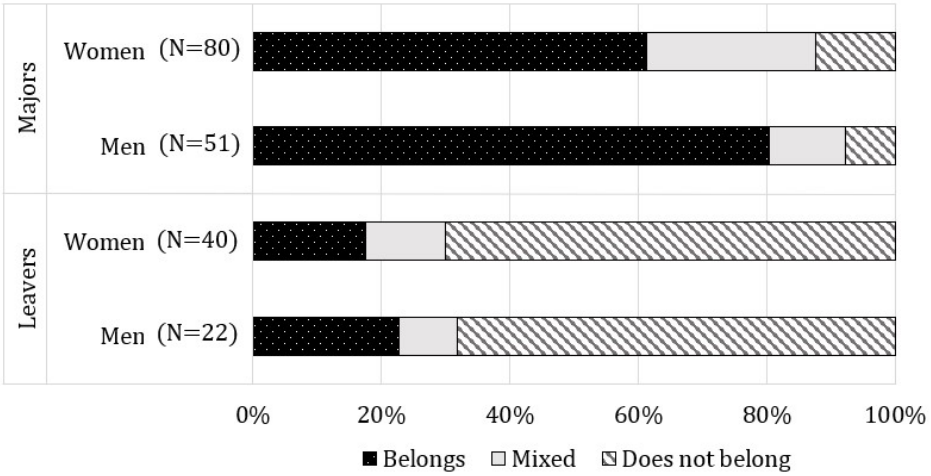


Figure 7.6: Student sense of belonging in STEM by gender and major status.

Figure 7.6 that female majors were less likely than male majors to report they felt they belonged in STEM. A total of 62 analyzable responses from leavers were coded (22 men and 40 women). Similar results are seen between male and female leavers. Perhaps unsurprisingly, leavers were more likely to report they did not feel they belonged in the STEM field they left.

7.3.1.3 Effects of Race *and* Gender on Belonging.

Figure 7.7 presents the relationships between belonging, race, and gender among majors. Due to the smaller number of leavers and large frequency of a negative sense of belonging among those respondents, only majors are focused on in this section. Not surprisingly, with women reporting a lower sense of belonging than men, and students of color reporting a lower sense of belonging than White students, it can be seen that women of color were the least likely to report a sense of belonging when compared to all other students. White men were the most likely to report a sense of belonging when compared to all other students.

These results reflect the importance of simultaneously considering the intersections of gender and race in any analysis such as this study. While most majors feel they belong, when we consider both gender and race together, we see that most of the STEM majors who report perceptions of not belonging are women of color and men of color. Notably, these students persisted in their STEM

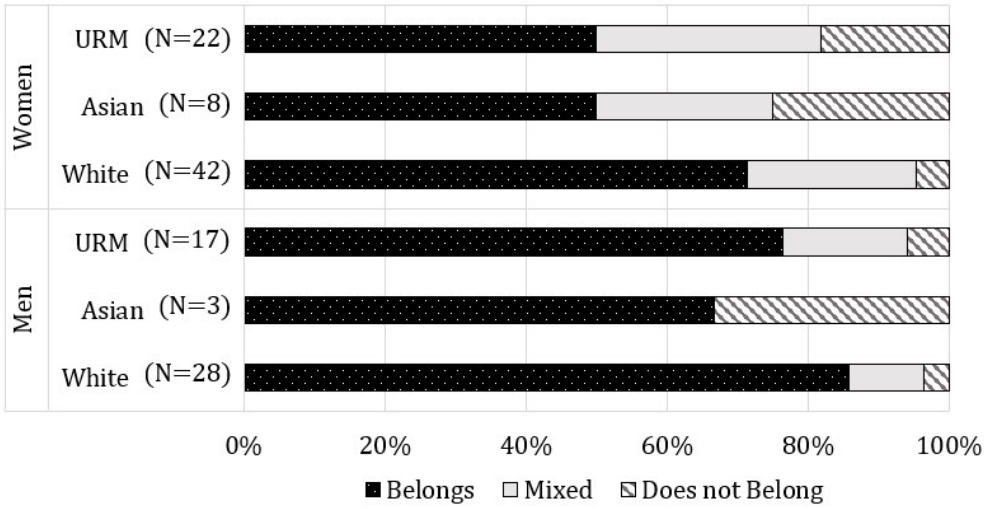


Figure 7.7: Majors' sense of belonging in STEM by gender and race.

majors despite not feeling as if they belonged there.

Asian Students and Belonging. While the size of the presented racial subsamples are small, the general patterns for Asian students match those of URM students more closely than those of White students for most analyses. This result suggests that although Asian students may be represented in STEM, their experiences may not align with those of the other racial group that is well-represented (i.e., White students). For the remainder of Sec. 7.3, racial groups are collapsed into only two categories: White students and students of color. We combine Asian students with other students of color because their responses (shown in Figure 7.5 and Figure 7.7) match closer to those of URM students than White students and the number of Asian students in this self-selected sample is quite low.

7.3.1.4 Sense of Belonging & Representation Status in the Major.

Sense of belonging within STEM fields was analyzed by gender and race. Biological science (biological) and the physical sciences (pSTEM) were disaggregated because the representation of women varies between fields of study in STEM. The data for men is not disaggregated by STEM majors, as men are generally represented, or highly represented, in all STEM fields. Figure 7.8

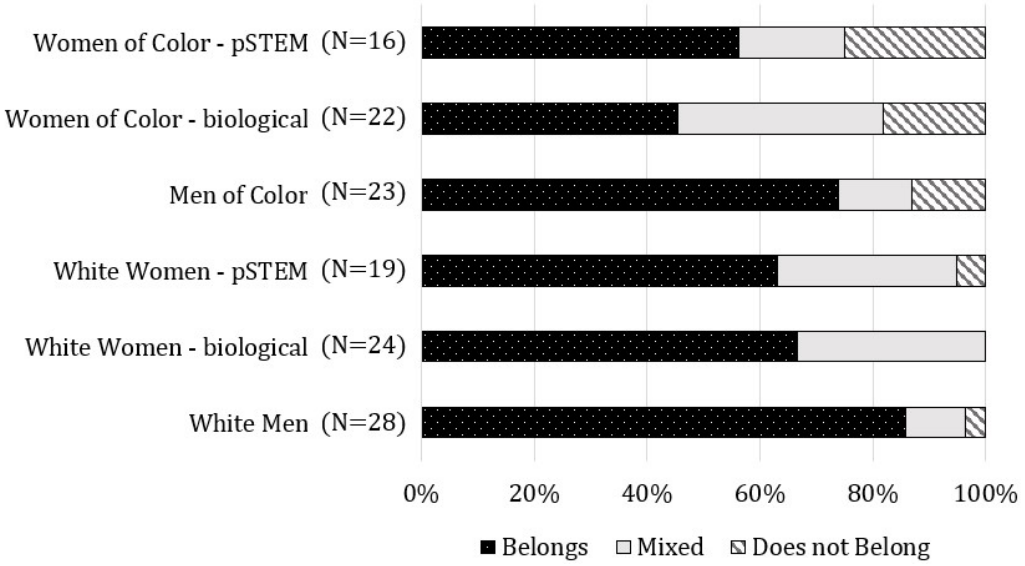


Figure 7.8: Sense of belonging in STEM field by race and gender. Women are disaggregated by major, biological sciences ("biological") or physical sciences ("pSTEM"), due to the differential representation of women between those fields.

presents levels of belonging by gender, race, and major discipline (for women only). It shows that women of color in pSTEM, where they are highly underrepresented, report belonging much less frequently than White men.

Due to small N, comparisons between groups should be made with caution. However, note that there is a general tendency for feelings of belonging to follow patterns of representation. As a student's demographic group becomes less represented, the less likely a person appears to report a sense of belonging. It is worth noting that lack of sense of belonging was most commonly reported by people of color, suggesting that race significantly impacts belonging, perhaps even more than gender.

7.3.2 Explanations for Sense of Belonging

This section presents analysis of common factors cited for belonging and lack of belonging. Factors cited by majors for their reported sense of belonging (N=70) and for a lack of a sense of belonging (N=55) are discussed, as well as factors cited by leavers for their low sense of belonging (N=37). Explanations for leavers who felt they belonged in STEM are excluded because very few

leavers reported a positive sense of belonging in STEM majors. Due to low numbers for majors and leavers, results based on race and gender are presented, but the intersections of race and gender are not. It is noteworthy that not all interviewed students gave explanations for their belonging status while some gave multiple reasons.

Four broad themes emerged during the coding of answers to questions regarding belonging, which were labeled as: interpersonal relationships, science identity, personal interest, and competence. Students either had or lacked the aspects encompassed by the codes. For example, students can attribute their positive sense of belonging to having interpersonal relationships or attribute their negative sense of belonging to a lack of interpersonal relationships. It is worth noting that some students' responses were labeled with more than one theme. Table 7.1 presents all belonging explanation codes.

Table 7.1: Summary of coding scheme used for explanations for sense of belonging in STEM. Each code is defined for students who had or lacked the reason described.

Code	Reason for belonging	Reason for not belonging
Interpersonal relationships	Feels socially connected with peers and/or faculty members. May share common interests with peers.	Lacks a social connection with peers. Feels socially different, does not fit in.
Science Identity	Science is a part of their identity as a person.	Lacks a personal connection to the major or material.
Personal interest	Expresses personal interest in course subject or major.	Explicit lack of interest. May find the material boring or unrelated to their reason for choosing the major.
Competence	Feels like they understand major-related material or receives good grades in major-related courses.	Feels like they do not understand major-related material well or receives poor grades in major-related courses.

7.3.2.1 Interpersonal Relationships & Belonging

In his theory of undergraduate socialization, Weidman defines interpersonal interaction as one of three processes of socialization. This includes relationships with peers or faculty as well as the frequency of interactions and intensity of those relationships [132]. In this study, the code interpersonal relationships (IRs) encompasses any personal relationships that students have with other members of their associated STEM department, such as faculty or fellow students. Having IRs means that a participant feels socially connected or similar to those around them in their STEM major. For example, a major coded as having IRs explained,

> *"I can really relate to the other biology majors. Most of my friends are biology majors. I feel like it's where I belong."* – White female biology major

Another interviewee related IRs and belonging to knowing people in the major, as can be seen in his response to the question about feeling out of place:

> *"At first I was just starting to get used to everything because I didn't know everyone but now I just fall right in."* – White male information technology major

These students were both coded as having IRs in STEM that contributed to their positive sense of belonging in their major.

On the other hand, a lack of IRs indicates a lack of social connection or similarity to those around them. Responses for a lack of IRs included differing hobbies from peers and social isolation, among other things. For example, one leaver coded as not belonging due to "lacks interpersonal relationships" said that, though he enjoyed the course work associated with his STEM degree, the social environment made him uncomfortable:

> *"I enjoyed the classes. I just did not enjoy the atmosphere. When I looked around, I saw all these people, all these people that I didn't fit in with, and I didn't feel comfortable there."* – White male exercise science leaver

A female leaver described how sometimes the lack of belonging was related to her demographic status:

> *"[I felt out of place] especially because I was like 1 of 2 girls at the time that was a physics major. Even that other girl that was a physics major with me, I think she changed to a math major."* – Asian female physics leaver

7.3.2.2 Science Identity & Belonging

In contrast to connections developed through interpersonal relationships, science identity is more focused on the individual student. Science identity as defined in this study is related to one's personal connection to their field, meaning science is closely connected to their sense of self. This definition overlaps in part with Carlone and Johnson's definition of research scientist identity, which relates to excitement for uncovering the natural world and scientific knowledge [155]. In short, science identity in this study encompasses one's feeling of being a "science person." Participants who expressed belonging based on having a positive science identity describe their major as an integral part of their life and who they are. When asked about whether he felt he belonged in his field, one major responded:

> *"Absolutely. I feel like this is exactly where I belong, and this is the type of work that I want to do."* – White male engineering technology major

This conveys that engineering technology is an important part of his life, as he says that is "exactly" where he belongs and what he wants to do. Students with a science identity, like the major quoted below, often expressed feelings of passion for the major as well:

> *"[I fit into my major] because I am passionate and I love what I'm doing. I would never change my major even for a slight second."* – White female engineering major

Stating that she would "never change [her] major even for a slight second" indicates that her major is an important aspect of her life and thus important to who she is.

Many who lacked a science identity expressed feeling like there was no connection between the major and who they are as a person. When asked if he felt he belonged in his previous STEM major, one leaver answered "no" and explained:

> *"I didn't feel like I was the type of person. Again, I'm not a nerdy guy. Not all scientific people are nerds, obviously, but I'm just a person who questions things just to understand, and that's why I think I'm a lot better as a journalist because [what] you need to [do is] ask questions, and they weren't people who answered questions well."* – White male physics leaver

Another leaver, in direct response to a question about belonging in her previous major (biology), commented that she felt out of place due to her lack of passion:

> *"I definitely kind of felt a little weird because everyone that was around me was so much more excited about what we were doing than I was. And I kind of felt like that was a problem because if it was something that I really loved then I should be just as excited."* – White female biology leaver

7.3.2.3 Personal Interest & Belonging

While science identity is closely tied to students' sense of self, personal interest relates to one's interest in major-related material or the major in general. Typical responses coded in this category were "I enjoy it" and "the major fits my interest." This differs from science identity because it lacks a connection to passion and sense of self. Responses categorized as personal interest focus on interest in the field, independent of how they view themselves as a person. Someone who was coded as having a science identity made a personal connection between themselves and the field, whereas someone coded as having personal interest expressed interest in a way that was not connected to who they are as a person. For example, as a direct response to the question about belonging, one participant said she belongs when she is interested in the material and does not feel she belongs when she is not interested:

> *"When I was in the classes I care about I feel like I belonged but when I'm classes that's like, we're talking about plants, vertebrate zoology, and stuff I feel like I don't belong there."* – Black female biology major

Here, she doesn't express having or lacking passion, but instead a lack of interest in a particular subfield that diminishes her sense of belonging. Majors and leavers both expressed a lack of personal interest as a factor contributing to their lack of belonging. A lack of personal interest

expressed by majors was often attributed to an emphasis in the degree program that differed from what they were interested in. For example, a biology major cited a lack of personal interest because the degree program focuses on anatomy and genetics while her personal biology interests lean more towards marine biology. One student expressed a similar perspective when asked if she felt she belonged:

> *"To be quite honest, not really... I have learned a lot of math but realizing that I am more applied math, the math department at [my school] on the whole is not an applied math department. It is much more theory-based."*
> – White female math major

Here, this major encountered a discrepancy between the content she is interested in and the content she encountered; she lacked personal interest in the content emphasized in her department, which contributed to her perception of not belonging in the major.

7.3.2.4 Competence & Belonging

In addition to interest in the subject matter, feelings of belonging or lack of belonging were also influenced by students' perceived competence in the subject matter. Competence in this study refers to people's *perception* of their own performance and understanding. This definition aligns with Carlone and Johnson, who defined competence as one's perceived grasp of scientific concepts and material [155]. Competence is captured by grades, conceptual understanding, and ability to communicate understanding to others. Participants frequently cited competence as a reason why they belonged. For example, when asked if she felt she belonged in biology one major replied,

> *"I do now. (laughs) Because I know that I can have a good understanding of everything I've been learning and it's like I know that because I can teach others. I can help others understand."* – Black female biology major

Numerous students cited their struggle to understand concepts as a reason why they did not feel they belonged. When asked if she ever felt out of place, one woman said she sometimes felt out of place or uncomfortable and gave the explanation:

> *"[My classmates had] a lot of practical knowledge and ... I didn't have all that knowledge and I was trying to learn it... I think that was a struggle for me."*
> – Hispanic female computer engineering leaver

This leaver felt that her understanding of the material didn't measure up to that of her peers; her feeling out of place was due to a lack of competence. Another student's explanation for his lack of sense of belonging echoes the previous student's explanation:

> *"I feel out of place because I think some of [the other majors] know more than I do and I wonder how because we have taken the same classes."*
> – Black male information technology major

Another student described mixed feelings of belonging that varied depending upon his self-perceived competence at the moment:

> *"Sometimes I do feel out of place, for example, with that group project... I didn't really know that much, but with another group project... I felt like I belonged because I had good ideas and contributed to the group and people listened to me... It kind of varies."* – Black male information technology major

It should be noted that students' responses about poor grades may not align with common expectations of what a poor grade is. For example, several female STEM leavers interviewed considered receiving any grade less than an A in a class as a bad grade. Thus, a high-achieving student may report low competence despite getting high grades in their STEM courses, and would be coded as lacking competence. Our coding of competence is based on students' *perception* of their own understanding and performance; we did not have access to information regarding students' actual grades or course performance.

Frequency of themes. In order to gauge the frequency with which these themes appeared in the interviews, appearance of each code was counted for majors and leavers. Figure 7.9 shows common responses for positive sense of belonging among STEM majors. Students could offer several reasons for belonging. Interpersonal relationships were the largest factor cited for each demographic group, aligning with the literature [15]. Competence was the second most commonly cited factor for majors' positive sense of belonging, and was cited at similar frequencies for all demographic

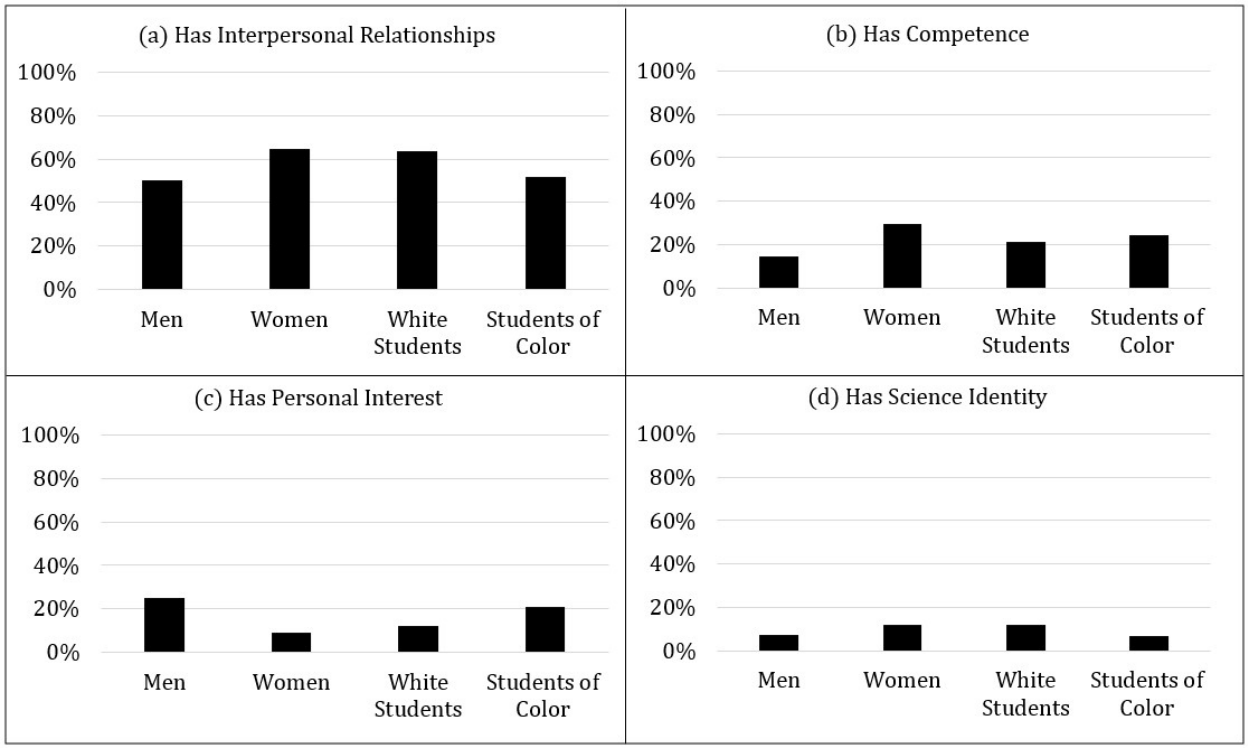


Figure 7.9: Reasons majors cite for belonging in STEM by race and gender. N-values for respondents are: Men (28), Women (34), White Students (33), Students of Color (29).

groups. Personal interest was the next most frequently cited by majors. Science identity was cited by only a small percent of interviewees. There were no significant differences between responses of men and women or White students and students of color.

Figure 7.10 shows common responses among majors and leavers who reported a lack of a sense of belonging in STEM. The most frequently cited factor contributing to a lack of belonging among majors was the absence of interpersonal relationships. This was true for all demographic groups. Lack of competence was the second most frequently cited explanation for all majors. No majors cited a lack of science identity as an explanation for their negative belonging status. Absence of interpersonal relationships was the most commonly reported factor contributing to leavers' negative belonging status. Leavers rarely cited lack of personal interest for their lack of belonging in their STEM major.

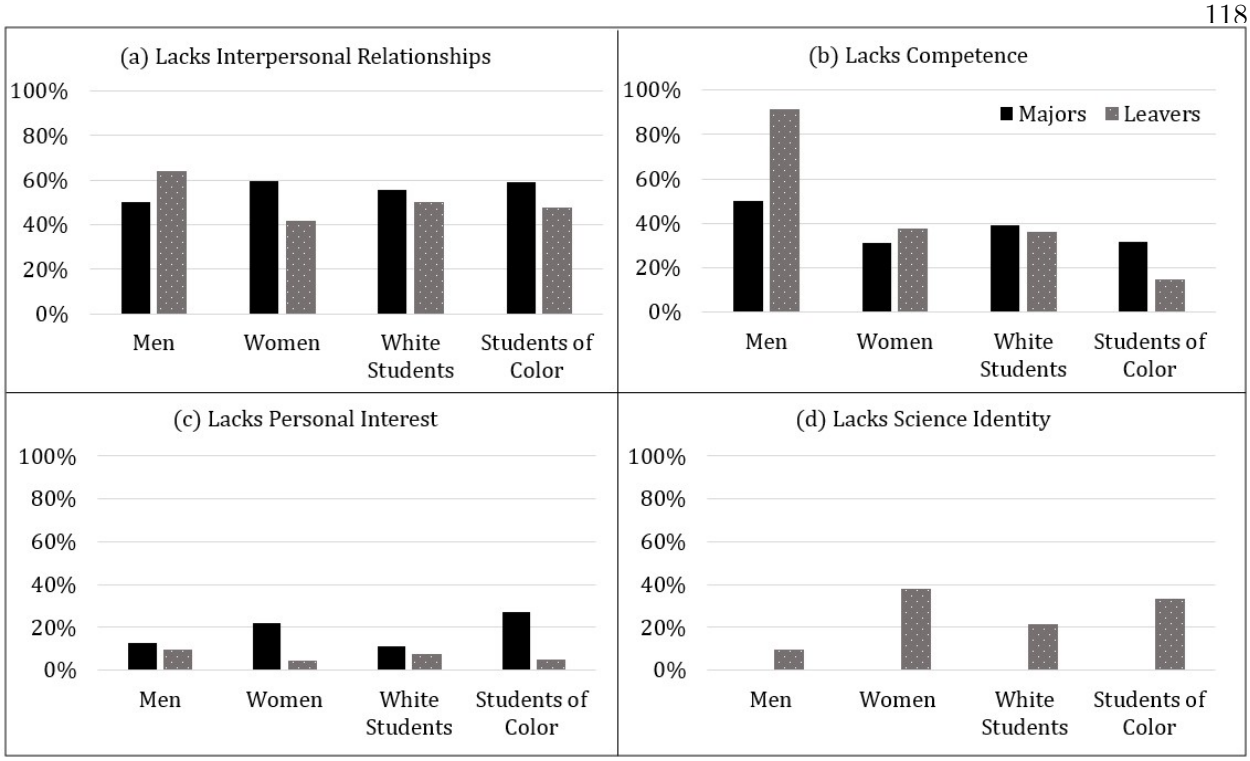


Figure 7.10: Reasons majors and leavers cite for not belonging in STEM by race and gender. N-values for respondents for majors and leavers, respectively, are: Men (8, 11), Women (32, 24), White Students (18, 14), Students of Color (22, 21).

7.4 Intersections of Professor Care, Instruction Style, and Sense of Belonging

In this section, the ways that instructional style, perceived professor care, and sense of belonging intersect for majors and leavers is considered. Because not all interviews could be coded for all questions, the number of responses for the intersections analyses are often lower than others presented in this chapter. Figure 7.11 presents results for all analyses including intersections of (a) instruction styles and professor care; (b) instruction style and belonging; and (c) professor care and belonging.

Interactive Teaching May be Associated with Greater Feelings of Professor Care. The connection between encountered instruction style and perceived professor care was investigated. There appears to be a relationship between reported instruction style and perceived professor care: as the level of course interactivity increases students in our sample are more likely to report their

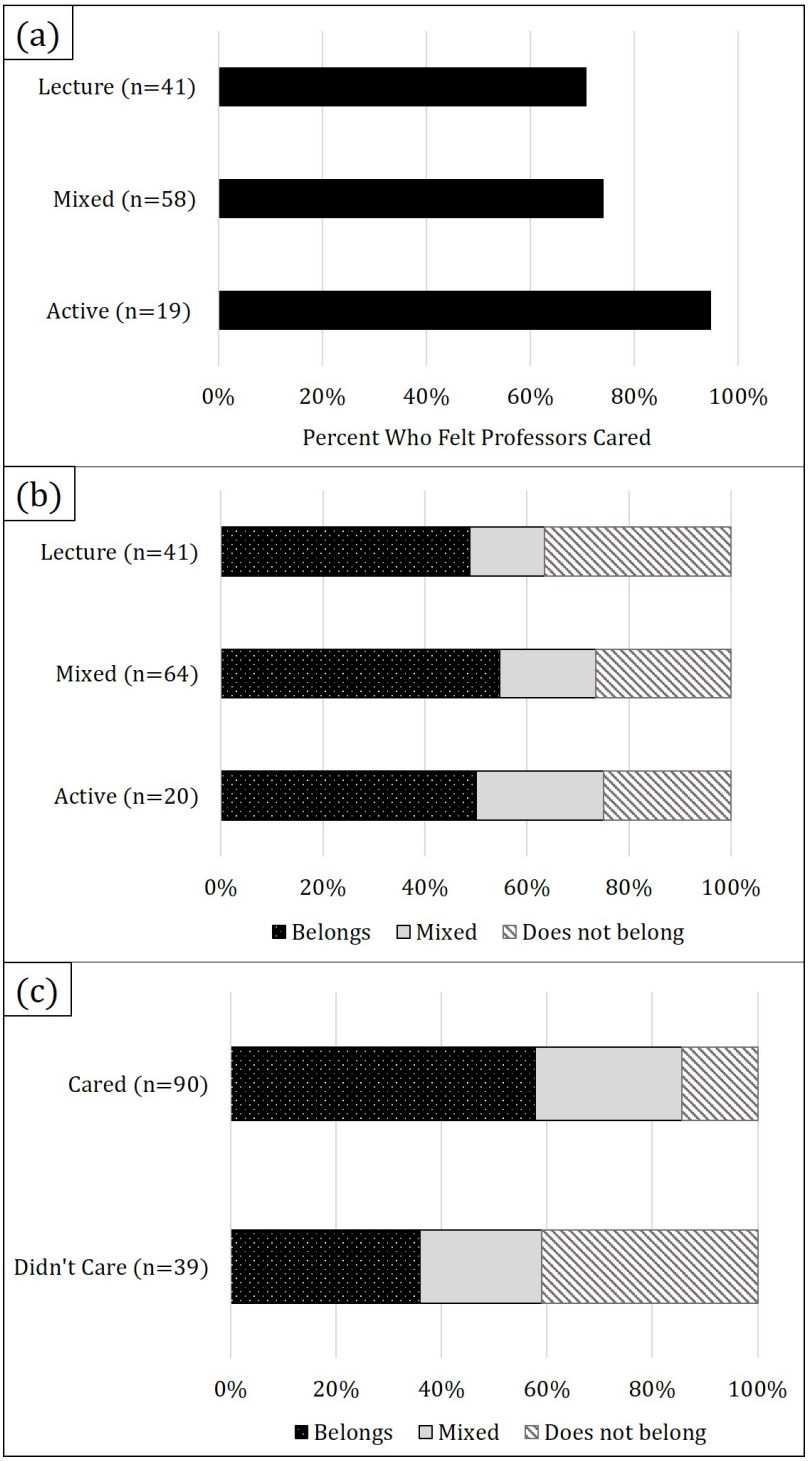


Figure 7.11: Results from analyses of intersections across different factors investigated: (a) Perceived instruction style and student perceptions of professor care in STEM; (b) sense of belonging in STEM and perceived instruction style; and (c) perceived professor care and student sense of belonging in STEM.

professors care. Students who perceived more interactive approaches also perceived more caring from their STEM instructors than those who reported lecture-based instruction.

Lecture-Based Instruction May be Associated with Decreased Belonging. Analysis was done to see how belonging and instruction style are connected. Our data shows a slight increase in lack of belonging for those in lecture-based courses, though this trend is not statistically significant.

Belonging and Feeling Professor Care Appear to be Associated. The relation between belonging responses and the care students perceived from their professors was also analyzed. Similar trends were seen for both majors and leavers, so the aggregated data are presented. The data show that those who reported professor care were more likely to report feeling they belong in STEM than those who did not feel cared about.

Sense of Belonging, Caring Professors, and Interactive Classrooms. These data suggest a connection between students' perceptions of their professors' instruction styles, feelings of belonging in their major, and whether their professors care about their learning. There are distinct differences in sense of belonging among STEM majors who felt professors cared and those who didn't. Those who felt professors didn't care were far more likely to report feelings that they did not belong in their major. The data also show a relationship between perceived instruction style and perceived professor care. Students who experienced interactive teaching were more likely to report feeling cared for by their professors. Moreover, gender and racial differences in these perceptions are consistent with demographic patterns of underrepresentation in STEM.

Chapter 8

STEM Students' Perceptions of Race & Gender Impacts

This chapter comes from second-author work published in the International Journal of STEM Education [22] and presents an analysis of responses students gave to two specific questions asked near the end of the interview:

(1) *Is the experience of being a {STEM major} major different for people of different genders?*

(2) *Is the experience of being a {STEM major} major different for people of different races?*

Here, {STEM major} refers to the STEM major chosen or dropped by interviewees such as biology or chemistry. Students' responses were analyzed using an iterative coding process. Two researchers independently coded a subset of responses identifying themes and then compared. In the first pass, ideas students suggested as impacts of gender and race were collected under common labels. They were then collapsed, expanded, and refined through a collaborative process. Then another subset of interviews was independently coded and again compared, further collapsing, expanding and defining the codes. Disagreements in coding were discussed until full agreement was reached. Through this repeating process, a coding scheme in which the majority of responses could be coded was finalized. Codes were collapsed under three broad categories descriptive of the source of the impact. These codes and categories are summarized in Table 8.1 and elaborated on in the presentation of results.

At the broadest level, responses can be categorized as either, "notices gender/race differences" or "doesn't notice gender/race differences." If a respondent indicated any difference in impact, that

Table 8.1: Summary of coding scheme used for students' perceptions of gender and race differences in the experiences of STEM students.

Doesn't Notice Differences	Not aware of race/gender impact		
Notices Differences	Not sexism/racism	Impacts due to differences in individuals not attributed to systemic factors	• *"Men and some races are more interested in or value science more."* • *"Women naturally work harder than men."*
	Sexism/Racism	Impacts due to underrepresentation	• *Intimidation* • *Pressure to work harder* • *Feeling out of place*
		Impacts due to discrimination	• *Negative impacts for employment* • *Bias against women and/or people of color*
		Students of color lack social or cultural capital	• *High school preparation* • *Socioeconomic status* • *Informal networks*
	Differential benefits	Impacts due to advantages women and students of color receive because of their gender and/or race	• *Scholarships* • *Job opportunities*

response was coded as noticing a difference. A small fraction of the "notices a difference" responses indicated a belief that women or students of color benefited based on gender/race. Such responses are reported separately. For those responses coded as "notices differences," we then divided them

into those which allude to structural/systemic factors and those which did not (difference was attributable to differences in genders/races and not systemic effects). For those coded as "notices differences," it was possible for a response to be coded in multiple categories if respondents expressed multiple beliefs. Responses coded as identifying sexism/racism impacts fall into three main categories: (1) impacts due to being a minority; (2) impacts due to discrimination; and (3) impacts of what we refer to as social capital (which appear only with respect to race). Table 8.1 summaries these categories.

8.1 Analysis

Analyses were conducted separately for student responses about gender and race differences. The total number (N) of respondents for both the gender and race analysis are shown in Table 8.2. For reporting, race is divided into students of color and White students. This is because reporting results by specific race/ethnicity (i.e., Black, Asian, Hispanic, mixed-race, etc.) results in sample sizes too small for useful comparisons. Additionally, when looking at data from students of color, patterns across racial groups emerged that are similar and distinctly different from White students. Finally, separating out by White students and students of color essentially divides the groups into those of racial privilege (White students) and those of non-privilege (students of color), which are essential identities relative to questions about racial impacts. It is of note that while Asian people are generally not underrepresented in STEM, which may call into question their inclusion in the non-privileged group, there is ample data suggesting their experiences in STEM are more aligned

Table 8.2: Gender and race demographics of respondents coded for gender and race analysis.

		Students of Color	White Students
Gender analysis respondents	Women	53	53
	Men	31	32
Race analysis respondents	Women	56	54
	Men	32	30

with people of color than White people (i.e. they do not hold the same privileged status as White people). This is especially true for Asian women [168, 169].

Analysis for perceptions of race and gender impacts includes responses from 183 total interviews. Due to variations in comprehensiveness of individual interviews (a question may have been deemed uncodable for various reasons), the sample size is different for each analysis: N=169 for gender and N=172 for race.

In the following sections, respondents' responses to the two interview questions regarding race and gender impacts are described, as well as overall frequencies for the broad categories found in student responses.

8.2 Lack of Perception of Race & Gender Impacts

If a response was coded as "doesn't notice gender differences" or "doesn't notice race differences" the respondent explicitly said there was no difference in the STEM experience for students of different genders or races or they stated they were unsure if there were differences. Many of these responses were of the form "I don't think so" or "not really."

Several respondents cited equal academic responsibilities for all students as an explanation for the lack of gender and race differences. For example, one Black female major said

> *"No I don't think it is. I think they're the same. Same course load, I don't think they get it any easier or any harder."*
> – Black female electrical and computer engineering major

Also common was an appeal to the objective nature of science to justify a lack of impact relative to race or gender. For example, a White male major said

> *"I feel like especially in the world of science we are being kind of objective about who we are and what we are doing, I think we are kind of on the same path."*
> – White male environmental science major

While most respondents in this category gave explicit responses about a lack of gender and race differences, some simply said they were unsure and did not commit to an answer. For example,

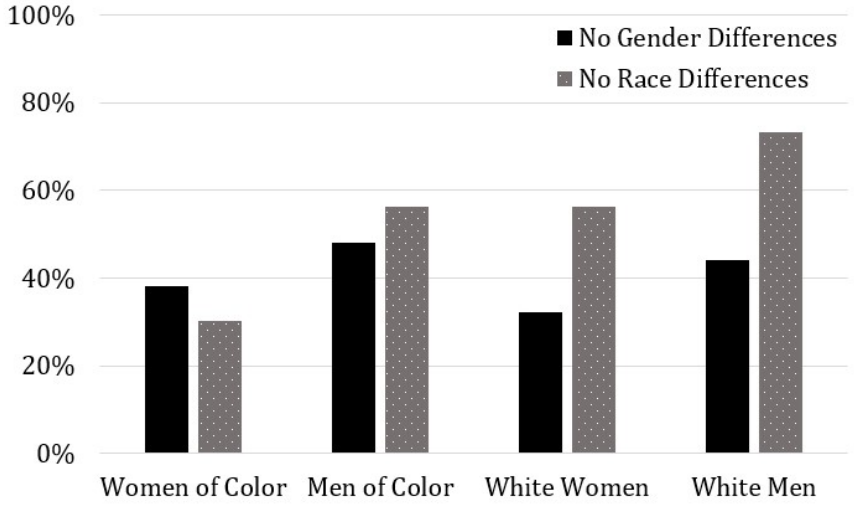


Figure 8.1: Percent of respondents by demographic group who did not notice gender and race differences. N-values for each group for race and gender analyses, respectively, are: Women of Color (53, 56), Men of Color (31, 32), White Women (53, 54), and White Men (54, 30).

a White woman majoring in biology stated *"No I don't think so. I mean I couldn't imagine how it would be"* and a White man majoring in aerospace engineering said *"I don't know. It's kind of hard to say for me because I don't have any experience"*

Distributions of respondents from each demographic group are shown in Figure 8.1. White men were much more likely than other groups to not recognize gender and race differences for students in STEM and women of color were the least likely. Except for women of color, students were less likely to report perceiving an impact due to race than gender. Respondents coded in this broad category of "doesn't notice differences" were not coded for any of the subsequent broad categories which required them to identify a gender or race impact.

8.3 Race & Gender Impacts from Individual Differences

Some students indicated that they perceived differences in STEM students' experiences related to gender or race, yet they did not attribute these differences to any cultural or structural systems in which STEM education takes place. Statements in which differences were noted based on individual characteristics that were not directly or reasonably inferred to be attributed to systemic

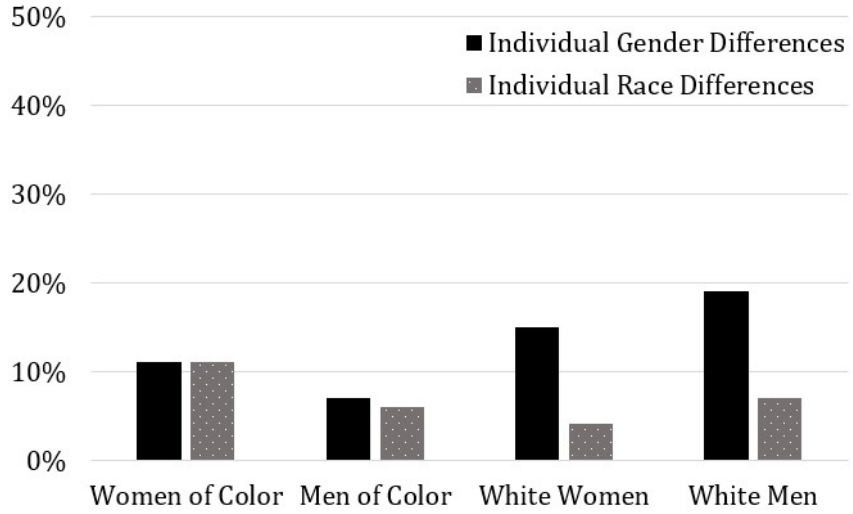


Figure 8.2: Percent of respondents by demographic group who cited gender and race differences due to individual differences. N-values for each group for race and gender analyses, respectively, are: Women of Color (53, 56), Men of Color (31, 32), White Women (53, 54), and White Men (54, 30).

causes fell under the code "Individual Differences." Figure 8.2 shows the distribution of responses in this category. Under this broad category student responses were categorized into various theme groups, each of which are described below for both gender and race.

8.3.1 Differences in Individual Characteristics of Men & Women

Some respondents described individual characteristics of men and women as the source of women's different experiences in STEM, without inference to these characteristics being driven by sexism. Below the two most common individual characteristics cited are described. White men were the most likely to see individual differences as the key factor in differing experiences of different genders.

8.3.1.1 Belief that Women are Not as Interested in Science as Men

After being asked if there were gender differences in their major, some students (nearly all men) noted the gender imbalance in STEM and offered women's ostensible lack of interest in science

as an explanation. For example, one student explained

> *"We do an open house and every time I talk to girls, with the exception of one out of the hundreds of families I have seen walk by me, only one seemed interested in electrical engineering and she knew that is exactly what she wanted to do. Others seemed to be afraid of the subject, they are probably afraid it would be too challenging or hard, but I suppose they probably don't know much about the major."*
> – White male electrical engineering major

8.3.1.2 Belief that Women Willingly Work Harder than Men

Some students (nearly all women) proposed that gender differences were due to effort expenditures. That is, the idea that women work harder than men. In order to be included in this category the response appeared to be a quality of the gender and not attributed to a systemic cause (i.e., women work harder because of discrimination). Those citing that women work harder due to a systemic cause are discussed later in another category.

For example, a White woman majoring in biotechnology said *"What I've seen is most of the girls at our school are a lot more willing to put in the extra work to study versus some of the guys."* In many cases, this involved the idea that women are more academically motivated than their peers who are men:

> *"I feel like guys like to take the easier route and it's not a lot of guys that like to be challenged, and for a female it's kind of like they will work harder instead of taking the easier route"* – Black female biology major

8.3.2 Differences in Individual Characteristics of Different Races

Some respondents described individual characteristics of different races as the contributor to racial differences in experiences in STEM without inference to these characteristics being driven by racism. Statements were coded in this category when the interviewee gave no indication of larger systemic causes to the differences between races (e.g., stating something is just a characteristic of Asian culture). The most common of these was the belief that some races value science more. In particular, many students mention people from Asian descent come from cultures that encourage

the pursuit of science. As one Black woman stated

> *"Not every African American would want to be a doctor, some people want to pursue other careers and same thing with Caucasians and Asians. I think there's a larger number of Asians who go into the math and science and technology kinda stuff than other races."* – Black female biology major

Additionally, a White male majoring in computer science expressed *"I can see it being encouraged a lot in the Asian communities because they put such a focus on both successful careers and like math-oriented careers."*

8.4 Differences in Experiences due to Sexism & Racism

This section focuses on those who attributed impacts to structural or cultural systems (i.e., sexism and racism), either explicitly or implicitly. As mentioned above, if a student listed any impact that could be categorized this way they were counted in this category. Therefore, this category represents the students who noted any sexism or racism impacts even if they also noted impacts not attributable to sexism and racism.[1] Under this broad category, student responses were categorized into three main categories: (1) impacts due to minority status; (2) impacts due to discrimination; and (3) impacts due to social and cultural capital (for race only). Each of these are described in detail below for both gender and race.

8.4.1 Impacts of Being One of the Few

A number of responses were coded as describing impacts related to the underrepresentation of some groups. For both gender and race, respondents talked about women and students of color feeling intimidated, feeling pressure to work harder, and feeling out of place due to their underrepresentation. Responses related to gender and race tended to be similar, therefore the discussions of gender and race are combined below. Figure 8.3 presents the number of students who fell into this broad category for gender and race impacts. A description of the most common

[1] The terms "sexism" and "racism" were rarely used explicitly by students.

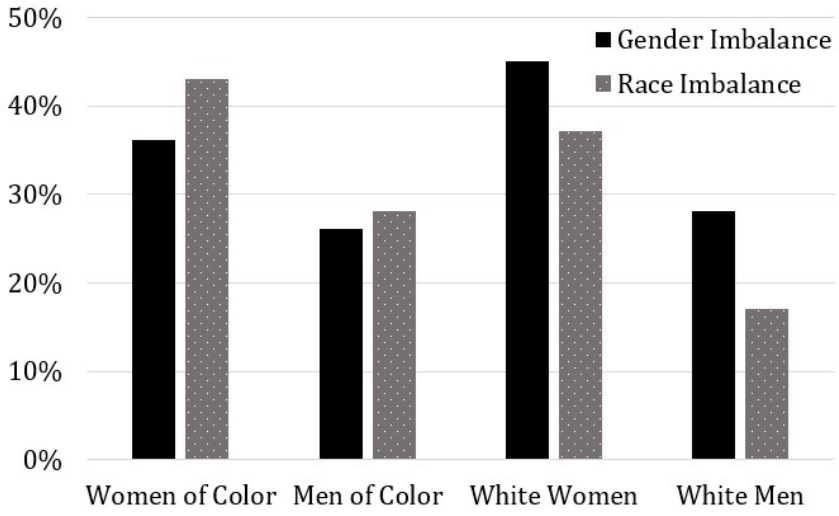


Figure 8.3: Percent of respondents by demographic group who cited imbalance in representation for gender and race differences in experiences in STEM. N-values for each group for race and gender analyses, respectively, are: Women of Color (53, 56), Men of Color (31, 32), White Women (53, 54), and White Men (54, 30).

impacts attributed to the minority status of women and people of color in STEM are described in detail below.

8.4.1.1 Feelings of Intimidation

A common theme under the impacts due to being a member of an underrepresented group was that those in the minority feel intimidated due to their underrepresented status, either directly when pursuing their degree or in job markets or internships. Often this came up as an expression of lacking the confidence to pursue STEM or of feeling discouraged. For example, when discussing her math department, one female major said:

> *"I have been surrounded by several males and sometimes I have gotten intimidated by it just because there are only six [female students] all together being outnumbered by these males... Sometimes I do get intimidated myself."*
> – White female math major

Nearly all responses in this category for gender were given by women, with White women making up the majority.

8.4.1.2 Pressure to Work Harder

Sec. 8.3 presented a theme revolving around the belief that women naturally work harder than men as a reason given for gender differences in STEM. Relatedly, some students mentioned that due to low representation of women and people of color in STEM, women and students of color would feel *pressure* to work harder. Although superficially similar ("women work harder"), these two categories are distinctly different due to the mechanism proposed (individual vs. structural) and are therefore reported separately. In a later section this same theme of working harder surfaces again in relation to bias.

Frequently, responses about women and people of color working harder were related to disproving stereotypes about women and people of color in science and feeling a need to prove themselves in the field. For example, one White woman double majoring in mathematics and computer science said *"It feels strange if I'm the only female in a class... it feels like I have to keep up... like, if I'm not as good, then it says something bad."* Another White woman spoke of pressure due to stereotypes as well, stating

> *"It makes me feel like there is more pressure because I feel like if I don't do as well or better that it will be like, turned into a gender thing... I don't want to be treated differently... I want to kind of prove myself like I'm not only as good, I'm better than you guys."* – White female biology major

After discussing the low numbers of women in engineering, one White man majoring in aerospace engineering expressed feeling sorry for women in engineering because they would be *"out of their comfort zone"* to be in a class of all men. He went on to say *"I guess that [women] would have to work a little bit harder, I would think, to prove themselves."*

Many people said that students of color likely feel pressure to work harder due to their low representation in STEM as well. Many of these comments had to do with combating or disproving stereotypes about certain races, such as Black and Hispanic, in STEM. For example, one Hispanic man majoring in engineering technology said *"I kind of push myself a little bit more just because ... the majority in STEM are White, White males, so being the only Hispanic in my class kinda makes*

me wanna try a little bit harder... I wanna be seen as equally competent as a White male." One Black man majoring in mathematics expressed feeling there was "equal opportunity" in STEM, but also remarked: *"I have the drive to try to prove others wrong and I'm here in this field and I'm successful in this field so far and I'm just going against the stereotype and against the statistics."*

What these students are describing is an experience of stereotype threat [170]. Stereotype threat research demonstrates that a person whose group is numerically marginalized in a classroom may perceive a spotlight on their "performance." If there is also a negative cultural stereotype associated with their group, the anxiety that spotlight triggers, combined with the negative cultural stereotype, can lead to diminished effort and/or lower performance. Students from underrepresented groups may decline to try hard; this way, any "failures" in the class are due to their diminished effort rather than confirmation of the stereotype. The widespread stereotype threat associated with certain racial/ethnic and gender groups has been demonstrated as a barrier to success for some underrepresented groups in STEM. Importantly, a number of respondents indicate that they fight the stereotype and try harder, rather than diminish their efforts.

8.4.1.3 Feeling Out of Place

Many students speculated that because women are outnumbered in STEM, they may feel out of place. For example, one White women majoring in computer science said about her major, *"I know it's easy for women to feel out of place or alienated just because it's not a very popular profession for women."* Many women voiced this feeling after being directly asked if they'd ever felt out of place in their degree program. For example, one Hispanic woman majoring in computer science said *"Sometimes [I feel out of place] because I am pretty much the only girl, I think in all of my classes this semester."* When asked if the feeling had changed over time, she said "no" and continued *"I think from the very beginning everyone is always like, 'oh a girl in computer science—this never happens' so I am probably used to it at this point."*

Although no White men expressed such views, some participants either experienced or observed that low representation of certain races and disproportionate representation of others can

make students of color feel out of place. One Black woman spoke from her direct experience as a Black woman in biology, stating *"Sometimes I feel out of place because I'm Black and that's a minority in biology."* Another White woman spoke from experience as well, but spoke from a place of knowing an African-American man in her biology program: *"I remember there was like one African-American in our class and I always felt like, bad for him because I was like 'oh, he's like the only one...' He kind of felt uncomfortable and he made a couple comments about [how] he felt alone."* Her acquaintance felt uncomfortable and alone because he was the sole Black person in the class. Similarly, a Hispanic woman expressed discomfort because she was the only Hispanic person in her statistics class, saying

> *"I was very uncomfortable... It was just mostly because I think I was maybe the only Hispanic in that class... I would always sit like in the back of the classroom so that like nobody would notice me... And in my other classes, I would always just stay toward the front of it."* – Hispanic female computer science major

8.4.2 Notices Impacts of Sexism & Racism as Discrimination

Students mentioned impacts that can be characterized as forms of discrimination against women or people of color. Most responses in this category either related to employment discrimination in STEM (finding employment and/or being paid fairly) or in cultural bias in favor of men and White people based on the stereotype that men and White people have more ability in STEM or are expected and encouraged to pursue STEM. Figure 8.4 indicates the number of students coded as expressing belief in discrimination as likely impacts. The following sections discuss the various subcategories of discrimination.

8.4.2.1 Employment Discrimination Against Women and People of Color

Some students reported the belief that the employment experience likely is different for women and people of color. Responses acknowledged women and people of color can experience discrimination when seeking a job in STEM or in terms of being paid fairly. For example, an Asian major said

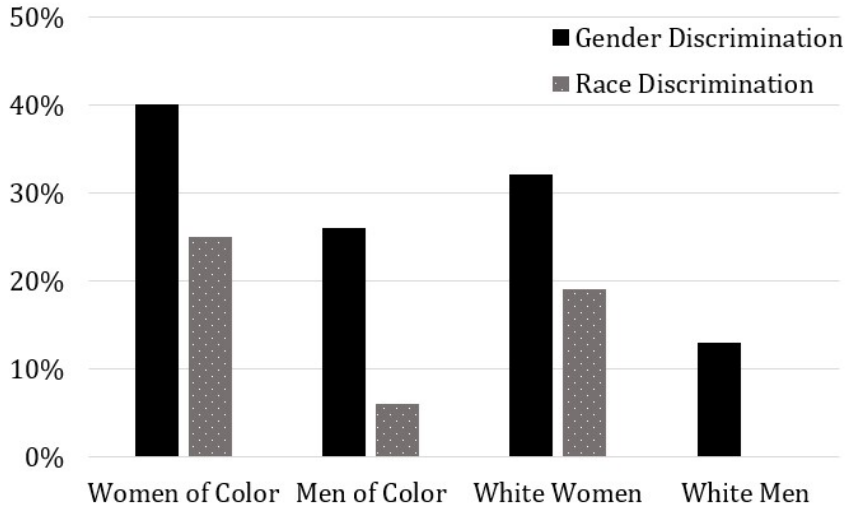


Figure 8.4: Percent of respondents by demographic group who cited discrimination impacts as explanations for gender and race differences in experiences in STEM. N-values for each group for race and gender analyses, respectively, are: Women of Color (53, 56), Men of Color (31, 32), White Women (53, 54), and White Men (54, 30).

> *"I think that out in the professional world there is still a bias towards men and against women when it comes to the science type field. It doesn't matter the degree you got, where you got it from, or the grades. I think that when it comes to science they are still going to prefer the male over the female. Why, I don't know. I feel like the women have to work harder and get more experience in order to get the job."* – Asian female chemistry major

A female information technology major echoed the difficulties of getting a job and also mentioned unfair pay:

> *"I have heard different stories that men get paid more than women... I have heard that men actually get paid more and women have a harder time trying to put themselves out there and trying to prove themselves when you are competing against men in the IT field."* – Black female information technology major

The idea that there is discrimination in employment was most commonly associated with gender-based discrimination with only a few students mentioning it regarding race.

8.4.2.2 Bias in Favor of Men and White people

Many students identified aspects of bias in STEM either in favor of men and White people or against women and people of color. Typically this was in the form of assumptions made in society at large about the competencies of women and people of color in STEM or of the expectations of careers women and men pursue. Students very commonly added that when faced with bias, women and/or people of color felt a need to work harder to prove themselves against these societal expectations.

8.4.2.3 Assumptions of Ability

Numerous students cited the stereotype that men and White people are more competent in STEM fields. For example, a female biology major stated

> *"I feel like our male teachers don't really expect females to do well with the subject; they kind of think that males are just all knowing and all seeing everything. I feel like sometimes they think we are not capable of doing certain things."*
> – Black female biology major

Or, as another Black woman majoring in chemistry stated: *"I've had some professors who make it plain and clear that, 'no, because you're Black you're not focused and I'm not really trying to, you know, give you the time of day.'"* In every case where a specific race was mentioned as being perceived to have less ability, Black people were the target.

Interviewees sometimes attributed the perception about differing abilities to stereotypes about who does science. When speaking about race students frequently referred to stereotypes about Asian and Indian people being more science inclined, for example a White man who left engineering stated, *"I know there are probably some cultural stereotypes that favor the Asian being known for math and engineering and especially now in today's world and also the traditional you know Caucasian engineers as your traditional stereotype."*

8.4.2.4 Assumptions About Who Pursues Science

There were a number of comments indicating a bias in the expectation that STEM is for men and White people to pursue. As a Hispanic woman majoring in geology expressed, *"I don't think that there is intentional racism. I think that it is probably just seen as more of a White field."* Many students indicated that the bias toward men and White people resulted in women and students of color being explicitly discouraged from pursuing STEM. As a White male majoring in computer science expressed:

> *"Women just aren't encouraged to make such a career choice growing up. I mean it's totally fine if dudes sit on the computer all day and play video games and stuff like that, which for our generation is easily one of the bigger influences on computer scientists but it's not okay for girls to do it... they are losers if they sit around and play video games."* – White male computer science major

8.4.3 Social & Cultural Capital

One emergent theme unique to discussions of race had no parallel for gender: social and cultural capital. Social and cultural capital refer to resources available to students, often linked to their higher socioeconomic status. Social and cultural capital can include superior high school preparation (higher quality teachers, AP courses), parental or extended family experiences with STEM subjects and occupations (role models), informal family and community STEM learning (science museums, summer camps, parental assistance with STEM homework), prior knowledge about applying to and succeeding in college (parents' education), informal networks as resources, and encouragement from peers or family [171, 172]. For example, when asked if there are race differences in pursuing a STEM degree, one Black female leaver said

> *"Yes, and mainly because [my school] is in a poverty stricken area. I think people from like places where it's more financially secure and stuff they have a better advantage over students with like poverty backgrounds, which would be like the Native American students or some of the Black students here, but with that being said... I think if you put in the same amount of work I don't think you would have a problem with that disadvantage."* – Black female chemistry leaver

Another Black woman majoring in biology discussed socioeconomic status and how the occupation of your parents can influence your pursuit of a STEM degree. She spoke of her friends' "richer families" with parents who were lawyers and doctors, whereas her parents were a truck driver and social worker. She then said *"I didn't really know anything about science also maybe because of that, because they {my parents} didn't have those types of jobs."*

Students of color and White women were the only students who mentioned something related to social or cultural capital as an explanation for disparate lived experiences as a STEM student. None of the White men interviewed made connections between race and social or cultural capital in their reflections on the STEM experience.

8.4.4 Bias Resulting in Women & Students of Color Having to Work Harder

Commonly, when bias was identified, students said that a consequence of this was that women and students of color had to work harder to prove themselves. For example, a White female biostatistics major stated *"I think sometimes you have to work a little harder to be taken seriously I think as a blonde female."* Similarly, a Black female biology major expressed, *"I think that is mostly a male dominated field so I feel that women that do choose to go into this career have to work harder to prove that they are as qualified."*

The notion that women and students of color work harder was common among all codes within the "notices differences" broad category. This finding is elaborated on in Sec. 8.6.

8.5 Perception that Underrepresented Students Benefit

Approximately 3% of students (most of whom were White women) felt that women and students of color had advantages over men and White students because of the current cultural and political emphasis on expanding STEM undergraduate populations to groups underrepresented in STEM. Comments about women and students of color benefiting from their demographic profiles were often related to job prospects, scholarship opportunities, and general encouragement. For example, a Black woman majoring in electrical and computer engineering commented on the expe-

rience of being both a woman and a racial minority and the benefits that would come along with that identity in the job market. She spoke about companies needing to meet a "minority quota" and how she was both female and African American. She went on to say *"I've gone to career fairs and walked by a booth and a guy is going 'no, no,... come here, come here, we're looking for African American females' and I don't know anything about their company and I might get an interview."*

Another common theme focused on underrepresented groups' receiving extra encouragement. As articulated by a White female major,

> *"I feel like as a woman at [my university] I have been especially encouraged and I don't know if that's just my experience but I feel like teachers are so afraid to be sexist that they sort of over put the emphasis on helping the women in their class."*
> – White female mathematics major

White women were the most likely to comment on how people of color, a group they are not a part of, benefit from their race. However, they were also the most likely to comment on themselves being able to benefit due to their gender.

While there is much discourse about programs to encourage underrepresented groups to pursue STEM, including scholarship opportunities specific to these groups, the available data (see Ch. 6) suggest large and negative impacts on these groups that limit access and ability to succeed. Even when ignoring factors that contribute to a general chilly climate for underrepresented groups and focusing only on the hiring process there still is no evidence that women or people of color are advantaged [173].

8.6 Perception that Women & Students of Color Work Harder

As discussed above, a common theme throughout the interviews was the belief that women and/or students of color work harder as a result of their demographic group's marginalized status. Figure 8.5 shows the distribution of this perception by demographic group. Sometimes this was attributed to a characteristic of a group without an attribution to any cultural or systemic factors that marginalized the group. Other times, the theme of extra effort by women and students of color

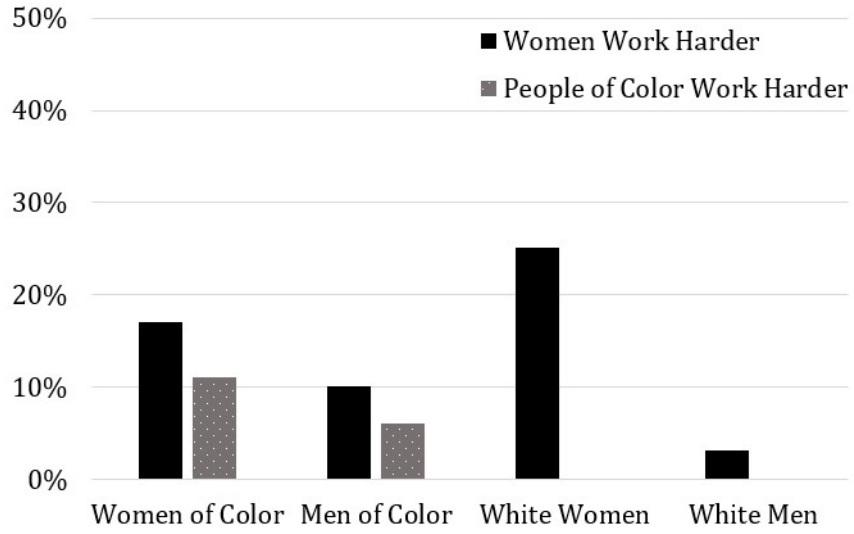


Figure 8.5: Percent of respondents by demographic group who expressed the theme that women or people of color work harder. N-values for each group for race and gender analyses, respectively, are: Women of Color (53, 56), Men of Color (31, 32), White Women (53, 54), and White Men (54, 30).

appeared as an expression of feeling pressure to do well as a result of being unusual or standing out. The third category in which we saw this theme arising was in students' expressed belief that women and students of color felt a need to prove themselves or work extra hard to do better than their White/male peers due to a perceived bias against them. White men were the least likely to report this impact and respondents were more likely to report the impact for women than for students of color.

A number of students observed the impact of needing to exert more effort had on the mental state of a student, either allowing men to be more relaxed, or resulting in women and students of color being more stressed. For example, a woman majoring in civil engineering noted her perception that men were more relaxed:

> *"I think it's more challenging for women... just looking at the guys they seem pretty set and you know they're going to pass and they're going to get a job, it seems pretty laid back. Whereas you see the women in civil engineering and they are really working hard and striving to outdo the guys because there's just such a stigma."* – White female civil engineering major

Another White female major noted the extra stress this places on women:

> *"I feel like with women . . . in chemistry are trying to prove themselves and push to show that they are able and capable of handling the workload and doing the work. . . I would say just in looking at the men and women in my major, the women are more ...I don't want to say stressed out but like they are more ...meticulous and anxious and organized about things whereas the men are more relaxed."*
> – White female chemistry major

This finding is highlighted here because it has important implications. At no time were students ever asked a direct question that would elicit the expression of women or students of color feeling a need to exert more effort. Yet, a significant number of students brought this up on their own. Most commonly they attributed extra effort from women and students of color to a desire to prove themselves because of bias against them and noted it leads to significant stress.

Chronic stress is associated with poor mental health (including anxiety, depression, and mood disorders) as well as lower learning outcomes [174]. The literature on mental health of STEM students relative to their demographic is sparse but consistent with the findings that women and students of color report more behavior associated with stress. Studies indicate women in STEM have lower overall mental health [175], more anxiety [176], and more depression than men [177]. In addition to the impacts of chronic stress on women and students of color in STEM, there is also evidence [14] that the perception of having to work harder than men in STEM leads to women to be less likely to feel they belong and to be less motivated to pursue STEM.

Yet, many of the participants identified their extra effort as a badge of honor. For example, a Black male mathematics major explained *"I have the drive to try to prove others wrong and I'm here in this field and I'm successful in this field so far and I'm just going against the stereotype and against the statistics."*

8.7 Summary of Gender & Race Impacts

Unlike prior research about each of the demographic groups in this study, these findings permit the comparison of perceptions of racism and sexism among members of several race-by-

gender cohorts of college seniors. In summary:

- Undergraduate students in this sample were often unaware that gender or race impacts the experiences of students in STEM despite extensive research indicating they have all experienced such impacts.
- Students who were aware of impacts, frequently attributed the effects to differences between genders and people of different races and not to any systemic or cultural causes.
- Among students who were aware of differences, the most common examples given were that men and some races are more interested in STEM (and therefore more likely to pursue and persist) and have different work ethics.
- Some students who were aware of impacts attribute them to sexism and/or racism, though none used these terms explicitly. Most commonly respondents stated that being a minority can lead to students feeling intimidated, feeling pressure to work harder, or feeling out of place, and these feelings have a negative impact on their educational experience. Students also identified ways women and students of color are discriminated against, especially in relation to experiences in employment and in bias against them.
- Differing levels of social and/or cultural capital were cited as an explanation for why some races are more represented than others. In doing so, students often conflated race with social class differences in opportunities to prepare for STEM.
- A small number of students, mostly White women, believed women and students of color benefit from their underrepresented status in STEM due primarily to societal efforts encouraging more women and students of color to pursue STEM.
- The belief that women and students of color work harder in STEM appeared across categories including: working harder is a characteristic of the group, it is due to being a minority, and it is due to discrimination.

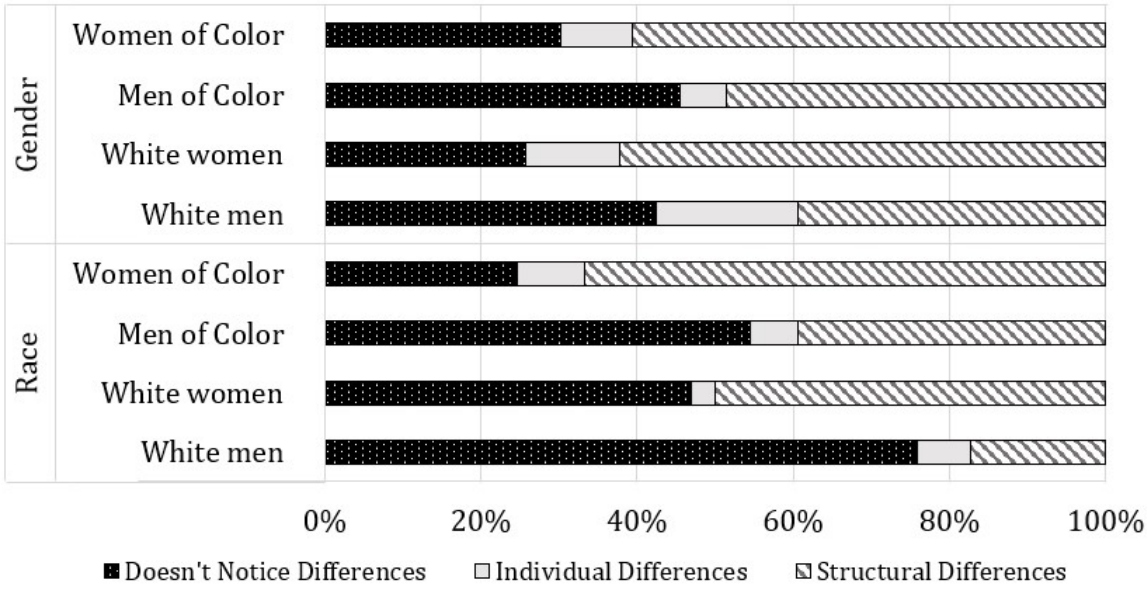


Figure 8.6: Overall distribution of responses to the question of whether there are differences in experiences due to gender and race. N-values for each group for race and gender analyses, respectively, are: Women of Color (53, 56), Men of Color (31, 32), White Women (53, 54), and White Men (54, 30).

Overall, these findings can be broken down into three broad categories: (1) respondent doesn't notice differences; (2) respondent notices internal differences between groups (not sexism/racism); and (3) respondent notices differential experiences and identifies them as grounded in structural factors (sexism/racism). These categories are summarized in Figure 8.6. Below two important trends in these findings are discussed.

Except for women of color, students were less likely to perceive race has an impact. Most students in our sample reported perceiving fewer impacts due to race than to gender. The research literature indicates that both race and gender significantly impact students' experiences in STEM, and the presented data do not allow for explanations of this gap in perceived impacts. It is worth noting that the students interviewed were attending schools with a racially diverse population (the overall system is approximately 40% students of color). This may affect their awareness of race impacts compared to students in other areas of the country where students of color are less represented at their universities.

The enhanced awareness of gender impacts may be due the greater attention given to issues of women in STEM in the media. It may be that students have simply had their attention directed more to issues of gender in STEM than race. However, it is beyond the scope of this study to investigate this possible reason. The finding of lower awareness of race impacts than gender impacts suggests there is a need to raise awareness of the influence of race, racism, and overall racial climate for students pursuing a STEM degree. It is more difficult to address inequity if there is not widespread recognition of its existence.

White males were the least likely to perceive impacts of race or gender on the experience of majoring in STEM. White men in our sample were the least likely to recognize race or gender impacts on students' experiences in STEM. This finding is illuminating, as White men are the majority population in many subfields of STEM and are overrepresented in positions of power and influence in STEM (i.e., they are more likely to achieve tenure and fill leadership positions). Only 40% of White men in the sample acknowledged any impact of sexism in STEM and only 20% acknowledged any impact of racism on the experience of pursuing a STEM major. The majority of White men in our sample appear to be unaware of the ways in which gender and race shape both their own and others' experiences. This has implications for STEM equity work because if the majority of individuals in the dominant demographic group do not recognize the disparate experiences of other demographic groups, it will be difficult to create the disciplinary cultural and structural conditions necessary to foster greater equity.

Research around equity typically focuses on the experiences of those who are in the minority rather than those in the majority. This implicitly views inequity in the context of deficits in individuals in minority populations rather than in the context of the structures and culture created by the majority population. Efforts that focus on people from underrepresented groups are likely limited in their impact on equity because the ideas and practices of the majority population also need to be understood, challenged, and changed. Thus, there is a need for shifting the focus of equity-based research in STEM away from deficit perspectives of students of color and focusing more on the ways in which dominant STEM culture can be changed to improve equity.

Chapter 9

Discussion: Impacts of Race & Gender on Student Experiences in STEM

Despite improving trends, women and many people of color remain underrepresented in STEM fields. The growing body of research on the problem suggests many factors contribute to a complex dynamic that results in the underrepresentation of women and students of color in STEM. The Roots of STEM Success Project advances understanding of the persistence of underrepresentation among women and students of color by analyzing interviews conducted with over 300 North Carolina college students who discussed if they perceived their STEM professors cared about their learning; the pedagogy they experienced and preferred in their STEM courses; whether they felt they belonged in these courses; and whether they were aware of differential experiences of students in STEM due to gender and race. This chapter presents a discussion of key findings and takeaways from these analyses.

9.1 Professor Care, Instruction Style, & Belonging

Professor care has been shown to have several positive impacts for students [117, 118, 122, 123], as has sense of belonging [15, 14, 17, 18]. Additionally, work has shown possible positive impacts of interactive course instruction for all students, regardless of gender or racial background [111, 112], though some work has suggested the implementation of this instruction impacts its efficacy [113, 114]. The study presented here adds to the literature by considering students' perceptions of care from their professors and sense of belonging in STEM, as well as comparing students' perceptions of STEM instruction styles to their preferences for instructional styles. Additionally,

the data expand previous work by comparing these factors for those who major in STEM and those who left STEM fields. Key findings from our study related to these factors include:

(1) STEM majors are more likely than leavers to report feeling their STEM professors cared about their learning.

(2) Women of color, whether STEM majors or leavers, perceive less professor care in STEM than students from any other gender/race cohort.

(3) Students in a physical science field are less likely to report feeling their STEM professors cared about them and their learning than those in biological sciences.

(4) Women who left STEM show the strongest preference for active instruction approaches but reported encountering it least frequently in their STEM courses.

(5) Students who feel they belong in STEM are more likely than students who did not feel they belonged to also report their professors cared about them.

(6) Students who recount encountering more active approaches also feel more care from their professors.

Findings presented suggest a discrepancy in the preferred and encountered instruction style for underrepresented students in our sample, in particular women and underrepresented minorities who left their STEM majors, who preferred more active approaches. Additionally, women of color were the least likely to feel their STEM professors cared about their learning. A sense of professor care appeared to be associated with active learning courses and a higher sense of belonging in STEM. Though these data are correlational, they point to the notion that increasing use of active-teaching approaches is a possible way to support students underrepresented in STEM and increase retention of those students. However, this is not a magic bullet, and there have been mixed results for underrepresented students when active-engagement methods are incorporated in STEM classrooms [115]. It is important to recognize that the implementation of these instructional styles can have a large effect of the efficacy and success of these approaches [113, 114].

The data also show feelings of belonging are associated with having personal relationships with college peers; students' confidence in their abilities to complete the requirements of the major (i.e., a sense of competence); their interest in their major field; and whether or not the student has a science identity. Sense of belonging was reported more among men, White students, and majors. STEM students' sense of belonging also appeared to be correlated with the number of members of the students' gender who also are in their major. The visible presence of learners "like me" renders the student's presence in this STEM environment as normative rather than an aberration—essentially conveying to the learner that "I belong" here.

Interviewees who reported personal relationships with others in the major were less likely to leave it. The influence of a personal relationship on staying or leaving the major holds for women of all races, White men, and men of color. Across race and gender categories, perceived competence in the field was the second most commonly cited reason given for feelings of belonging and/or lack thereof. STEM majors are more likely to attribute their science identity as a reason for their sense of belonging (as opposed to a lack of it being a reason for not belonging) and remaining in their major. Levels of personal interest in the STEM discipline contributed a small degree to major's sense of belonging, but its magnitude is not markedly different than levels among STEM leavers, with the exception of women and students of color, whose personal interests contribute more to their reasons for belonging than interest contributes for men and White students.

The main reasons leavers gave for their low levels of belonging were a lack interpersonal relationships and weak sense of one's competence. Student interviewees cited lack of confidence in their capacities to complete the major at a certain level as a motivation for leaving. However, unlike majors, leavers also report a lack of science identity as a significant reason for not belonging, and this lack of science identity is more prevalent for leavers who are women and students of color than for other leavers.

The finding that STEM leavers across all demographic groups in our sample report a lower sense of belonging indicates an association between feelings of belonging and persistence in STEM. Additionally, students from underrepresented groups in STEM overall were less likely to feel they

belong. Considering the intersections of race and gender illuminates this particularly striking association among women of color, who are more likely than any other demographic group to feel they do not belong in STEM, a sentiment found even among women of color who persist in the field.

Key implications of results regarding sense of belonging in STEM are discussed in more depth in the following sections.

9.1.1 Demographic Isolation & Sense of Belonging

Analyses based on levels of gender or racial representativeness in a STEM field in relationship to students' sense of belonging indicate that being demographically similar to others in the field may positively impact belonging. In fields where gender parity has been reached (i.e., women in undergraduate biology programs), students from demographic groups still underrepresented in other STEM fields report high levels of belonging. This raises the possibility that the problem isn't that women or people of color inherently feel they don't belong in STEM but, rather, they are responding to the unbalanced representation of certain demographic groups in some fields of STEM. Such a response to demographic imbalance implies that even if all other individual, familial, or academic preparatory factors are aligned for a student's STEM success, until STEM fields become more demographically diverse, those in the majority group (generally White men) will remain privileged by the culture and organization of the discipline in ways that sustain their sense of belonging while undermining the sense of belonging of students from underrepresented groups. This interpretation suggests the underlying problem is systemic (i.e., based in the cultures and organization of STEM rather than individuals' characteristics or preferences) and will require a systemic solution crafted to change institutional cultures and organizational structures.

9.1.2 Science Identity & Sense of Belonging

Based on our analyses, the presence of a science identity differentiates those who stay and those who leave a STEM major. A key finding in this work is that the absence of a science identity

is greater among underrepresented groups. No STEM major in our sample indicated lack of science identity as a reason for not belonging. In contrast, approximately one-quarter of leavers did so. These findings align with work by others who demonstrate the connection between science identity, race/gender, and persistence [135, 155].

The data are correlational and sample sizes are relatively small, so the extent to which a lack of science identity motivates dropping out of STEM and the extent to which those who drop out then feel a lack of identity cannot be directly determined. However, among leavers in our sample, women were more likely to report a lack of science identity than men, and students of color were more likely to report a lack of science identity than White students. No White male (major or leaver) in our sample reported a lack of science identity; all reports describing a lack of science identity were from students in marginalized groups. This relationship suggests that science identity is strongly connected to both race and gender. Given that students of color and women leave STEM at higher rates, our work suggests that science identity may play a role in persistence.

9.1.3 STEM Interest & Sense of Belonging

Leavers in our sample rarely cited their lack of personal interest in STEM to explain why they left their major. However, majors were more likely to mention it, especially students of color and women—even though interest in science contributed only a small part to interviewee's sense of belonging. This is an important finding to note because it challenges the idea that women and people of color are inherently less interested in science and instead suggests that structural and disciplinary cultural factors, rather than individual preferences, should be the focus of future examinations of representation in STEM.

9.1.4 Interpersonal Relationships & Sense of Belonging

The relationships students form with faculty as well as with peers in STEM are important in relation to persistence for students in our sample. The presence or lack of interpersonal relationships was the most common reason cited for belonging or lack of belonging. This implies that

for departments interested in increasing persistence rates among students from all demographic backgrounds, taking inventory of the extent to which members of the department—faculty and students—form positive connections and providing structures that encourage strong interpersonal relationships are likely to be productive strategies.

9.2 Student Perceptions of Gender and Race Impacts

The findings presented in the previous section confirm and add to the large body of literature that shows students have differential experiences in STEM along lines of race and gender. These studies provide numerous examples of ways in which women and people of color have experiences as they pursue STEM that limit their access and engagement in the field.

There are few studies, however, that have investigated students' awareness of these differences, especially differences for those outside of their gender and/or race. This study suggests that women of color may be the most aware of gender and race differences in experiences, while White men may be the least aware. This has important implications in STEM, in which White men are overrepresented and thus hold much of the power and privilege within these fields. With power comes more ability to implement large-scale systemic changes. If White men are largely unaware, or in denial of, negative impacts of STEM culture on women and students of color (as our data suggest), they are less likely to implement changes that can have strong positive impacts on the culture of STEM and underrepresented populations.

White male undergraduates' lack of awareness of what their peers experience can become part of the "chilly climate" with which women and students of color must contend en route to their STEM degrees and communicates that they *perceive* White male ownership of the field. Addressing the deficiency in awareness, and the underlying culture that supports the lack of awareness, should be part of any effort to address equity in STEM. Equity in STEM extends beyond individuals striving for their degrees. Greater equity in STEM fields can begin to address the loss of talent among historically excluded groups and the issues of distributive (in)justice related to STEM careers and social mobility.

9.3 Study Limitations

In the work presented here, the sample is not representative of all STEM students along several dimensions. It is not geographically diverse (all students were from North Carolina); is limited in socioeconomic diversity (all students were attending a public institution); deliberately overrepresents students of color and women compared to their representation in STEM; and was subject to self-selection bias.

Due to the problematic nature of dividing the sample into overly descriptive categories, analyses categorize students of color or underrepresented minority students into one racial category. Comparisons between different racial groups within the categorization of "students of color," or between students pursuing different majors within STEM, was not possible due to low N and the need to maintain anonymity of participants. As a result of this, findings must be interpreted and applied only at a general level. One cannot assume that the experiences of subgroups within certain racial categories (e.g., Black and Hispanic students) all encounter the same experiences, nor can they assume that the presented findings are true of all students of color, all White students, all women, or all men. One must recognize the unique experiences each student encounters, but also consider the ways in which their racial and gender identities may influence their experiences in STEM environments.

It is worth acknowledging the limited gender options provided on the initial survey when recruiting students; they were prompted to select either *male* or *female*. These data were collected in 2013. In years since this study began, more inclusive language has been continually adopted throughout academia and society to expand definitions of *gender* (e.g., man, woman, non-binary; cisgender or transgender), and distinguish this from *sex* (e.g., male and female).

The study did *not* collect demographic information beyond race, gender,[1] and major. Though the analyses presented are intersectional across race and gender, it is worth highlighting that

[1] Gender here refers to male and female. In hindsight, our study should have provided *man* or *woman* as gender identity options. However, gender identity of students was indicated as male or female on the survey; "male" and "female" were used as genders in this study to respect the identities that students self-reported.

these are not the only aspects of one's identity that may influence one's experience in STEM. For example, LGBTQ+ identities were not collected, but recent studies have shown that belonging to the LGBTQ+ community can have significant impacts on experiences in STEM [178].

Identities lie along many dimensions, and these identities interact in complex ways that cannot and should not be reduced. The aim of the presented study is to provide information that can illuminate and better inform discussions revolving around equity, diversity, and inclusion in STEM. One should not interpret this work as law and understand the variable nature of students' unique experiences.

9.4 Implications

The ultimate question for researchers in this area is "What can be done to increase the representation of women and people of color in STEM?" While causal claims cannot be made with the presented data, together these findings shed needed light on the persistent demographic underrepresentation of women and students of color in most STEM fields. Results can be used to inform retention efforts in STEM fields; approaches taken, along with lenses used to interpret those results, can be used to inform future studies regarding representation of women and students of color in STEM.

9.4.1 The Deficit Model of Women & People of Color

While rarely explicitly stated, many change efforts around equity operate under a deficit model of change. The deficit model views the problem of underrepresentation through the lens of individuals instead of larger cultural systems [179, 180]. Under the deficit model, it is assumed that women and people of color lack something (academic preparation, social capital, role models, assertiveness, confidence, experiences, encouragement, mathematical skills, etc.) that hinders their participation in STEM. This model supposes that the problem of underrepresentation can be addressed by providing supplements to those who are underrepresented, such as extra training and opportunities for experience or mentoring. Under this model, the assumption is made that gender

and ethnic/racial equity in STEM will be reached when those who are marginalized (women and people of color) change to fit the system that is in place. This model does little to question the culture or structures that contribute to the marginalization of the groups who are being asked to adapt to the current system of STEM education.

This model is flawed because it fails to recognize larger cultural systems and places responsibility for change on those who are marginalized, while it leaves intact the cultural climate, organizational structures, and classroom practices that contributed to creation of the problem in the first place. Though efforts to support the academic experiences of women and people of color can be valuable, they may not be sufficient for significant and lasting change to occur. In contrast to the deficit model of change is a *systems* model of change. Under the systems model, it is the environmental and cultural structures that are the target units for change as opposed to individuals; structures within the system can act to privilege some and thereby disadvantaging others. This model shifts away from a "fix the underrepresented people" model to a "fix the culture and systems" model.

The presented findings support a need to view equity efforts through a systems model. Specifically, the interviews in the presented study suggest that the classroom environment likely is experienced differently by those students from underrepresented groups than students from more privileged gender and racial backgrounds. While all students found an association between perceptions of a classroom as active and perceptions that their professors cared about their learning, women and students of color who leave STEM are the most likely to report a disconnection between perceived classroom environment and what they desire. Further, women of color in the sample were significantly less likely to feel their professors cared about them or feel they belonged in their field.

These results have the following implications for equity work:

(1) Efforts to address inequality in STEM should not disregard White men. As discussed, the analysis indicates the majority population in STEM may be unaware of the experiences of marginalized groups. Lasting systemic change will necessitate research to document,

challenge, and leverage the perspectives of White men. As others are increasingly noting [181, 182], reform efforts often operate from a deficit model, assuming the marginalized population needs to change in order to achieve success (e.g., mentoring programs, enrichment activities, scholarships, etc.). This perspective disregards the impacts of the majority population by instead attributing the problem with, and assuming the solution should come from, marginalized groups. Additionally, this perspective devalues the perspectives of the marginalized group in favor of the norms of the privileged group, without questioning the value of those norms. The focus on White men in STEM as setting the norm to be achieved, while disregarding them as a point of change, is a shortcoming that needs to be addressed. If the group that most influences the dominate culture is unaware of that culture or its impacts on others, and if little effort is put on identifying and changing the dominant group's understanding, one must wonder by what mechanism change would be expected.

(2) Greater attention is needed to issues of race in STEM. There was less awareness of race impacts among students in the sample. This may be due to a lack of awareness rather than actual experience, as much of the discourse among STEM research has focused on the experiences of women in STEM. There is ample evidence that race has a significant impact on one's experience in STEM [10, 183] and no indications that it is less than the impact due to gender. For inequity to be better addressed, a greater acknowledgement of the inequities experienced based on race is needed.

(3) More work is needed on the relationships among racism, sexism, and the stress they trigger among underrepresented groups. As discussed, a large number of students commented that those who are marginalized work harder, which was often connected to feeling more stress and anxiety in response to marginalization. In a search for literature regarding this topic, little literature appeared that addressed what appears to be a common and impactful issue.

9.4.2 The Importance of Intersectional Analyses

The work presented here utilizes an intersectional approach with regards to gender and race. These types of analyses are important in the field of equity research and highlight key phenomena for certain groups who could otherwise be overlooked. Unintentional exclusion of experiences can occur when cross sections of identities are not considered.

Research that examines only race or only gender differences in outcomes may mask the ways intersections of race and gender shape experiences of students. For example, the presented interviews indicate that both race and gender moderate the experiences that impact sense of belonging for STEM students. Importantly, we see large differences between women of color and both White women and men of color in our sample. Women of color reported the feeling a sense of belonging less frequently than any demographic group. Only a bit more than half of the women of color majors reported consistent feelings of belonging. These were women who were nearing graduation with a STEM major and yet frequently did not feel they belonged in the field in which they were about to receive a degree. The extent to which this group struggles with belonging can be overlooked when race and gender are not considered together. Additionally, the lack of belonging reported by men in the sample was primarily experienced by men of color. Combining the experiences of all men obscures this finding.

Intersectional analyses are absent from much of the STEM education literature, which suggests that findings reported in the literature about women's experiences are most likely reporting the experiences of White women and reports on people of color are most likely reporting the experiences of men of color. Not disaggregating students by both gender and racial identity masks the unique experiences of women and men of different races, which has consequential implications for effective strategies in improving experiences for underrepresented students.

9.4.3 Implications for Teaching

Taken together, findings from this study suggest that a possible effective way to address representation issues is to address classroom environments. Recent surveys indicate that while there is a growing trend toward more active teaching approaches, the majority of faculty report extensive lecturing. For example, the 2010-2011 HERI survey of faculty [184] found that, in STEM, the majority of faculty (70% of male faculty and 50% of female faculty) report using extensive lecture most or all the time. Surveys in individual disciplines have found similar results. A 2010 survey of calculus faculty found 80% of instructors lecturing "very often" or "often" [185]. Literature suggest the majority of STEM faculty use extensive lecture the majority of the time. Results of this study indicate that shifting toward more active approaches may increase the persistence of women and students of color.

It is of note that as teaching becomes more active and less lecture-based, the level of interpersonal interactions both between students and the student and their professor increases. As meaningful interpersonal interactions increase it is likely the perceptions of care from faculty will increase as well, which can support a greater sense of belonging, especially among marginalized students. However, it is important to recognize the influences of effective implementation in reaping the benefits of active instructional approaches [114].

9.5 A Note on the Persistence of Women of Color

Prior studies have found associations between a students' sense of professor care with several success measures [118, 122, 123], including rentention [117]. The same can be said for student sense of belonging and retention [16, 15, 131]. This work supports the idea that those who lack care or belonging may leave the field.

The data presented here shows women of color reported the lowest ratings of professor care and belonging in STEM. However, all students in this sample were college seniors nearing completion of their degree. This presents a narrative counter to those frequently found in the literature; it

indicates that, despite encountered barriers in STEM, the women of color in this sample found ways to successfully complete their degrees and overcome challenges that often push women of color out of the field. Though it is beyond the scope of this work, there exists literature, e.g. ref. [183], that investigates women of color's persistence and resilience in STEM. This is an area that could benefit from further investigation.

9.6 Roots of STEM Success Project: Summary

The presented study is unique in several ways that go beyond many studies in the literature. Race and gender, as well as their intersections, are included in analyses in contrast to many studies that consider only one dimension or the other. Additionally, the data come from a self-selected sample of college students with a wide range of backgrounds who attended one of the 16 public universities in North Carolina, while many studies report data from only one institution or from more selective institutions. Differences across different STEM fields are also included in some analyses, instead of looking at only a single field or collapsing all majors into a single STEM category. Impacts of *perceptions* of instructional style on persisting in the STEM major are also addressed, a topic that is rarely studied. Sense of belonging and whether the students perceive their professors as caring (an understudied area in STEM education on its own) is connected to the issues of instructional approaches in STEM as well.

The findings of this portion of the Roots of STEM Success Project focus on how certain features of the organization and culture of STEM disciplines contribute to the problem of underrepresentation. The presented findings underscore the necessity of taking a systems approach to change, rather than attempting to equip individual students from marginalized groups better to withstand aspects of the culture of the STEM disciplines that they—as well as some students from more privileged groups—find problematic. With a systems approach to improving equity in STEM, issues of underrepresentation can more effectively be addressed. As part of this systemic approach, this work suggests a need to identify and better understand the ways in which the organization, norms, and cultural climates of STEM disciplines work to support those in the privileged

group while often discouraging those from marginalized groups. In the biggest of pictures, taken together, these findings point toward cultural and structural factors that impact experiences of STEM students, which can impact persistence.

It is important to note that White men who pursue STEM likely benefit from the privileges of being in a field in which there are many others similar to them and where they frequently fit favorable stereotypes about who pursues STEM. Even when students stated they belonged in STEM, they sometimes expressed feeling others did not feel they belonged due to not fitting those stereotypes. Additionally, having commonalities with others privileges White men in forming interpersonal relationships, which data show are important for persistence.

White men were the least likely to recognize the well-documented differences in experiences for women and students of color in STEM. As they hold much of the power and privilege in STEM disciplines, educators and researchers must work to better inform White men about issues regarding equity, such that effective structural and systemic changes can be made. Research regarding equity in STEM should turn to examining dominant STEM culture to investigate ways in which the cultural norms and systemic barriers in STEM can be dismantled. This, along with more intersectional analyses of beliefs and experiences, can lead to more effective changes in STEM that support marginalized students.

Chapter 10

Conclusion

Understanding of the ways in which race and gender influence experiences while pursuing STEM degrees can inform development of research-based assessments and interpretation of their results. The work presented in the second half of this dissertation (Ch. 6-Ch. 9) showed several ways in which identity mediates experiences in STEM, including whether one feels they belong, whether they feel their professors care, and if they encounter the instruction styles they prefer. The intersectional analysis approach in Part 2 provides insight into the experiences of a women of color, who are often overlooked in studies that look at only race or only gender. Some findings would be obscured if gender-by-race cohorts were not investigated, such as the finding that White women feel the most care from their professors while women of color feel the least. Additionally, we found most men who lacked belonging were men of color, and that women of color reported lower sense of belonging than White women.

Despite the multiple barriers facing women of color in our sample, from less care from professors to lower sense of belonging, they remained persistent in pursuing their degrees, completing their STEM programs and nearing graduation. This is something that should not be overlooked, and more studies looking at the resilience of women of color could be beneficial in supporting women of color and other people from underrepresented groups in pursuing and persisting in STEM.

Many students who left STEM reported feeling their professors did not care about them or their learning. Even many senior STEM majors had the same seniment regarding professors' lack of care. Combined with the differential perceptions of care by women of different races, this

has implications for how STEM professors engage with their students to foster belonging and retention of underrepresented populations. This could be addressed by faculty through developing interpersonal relationships with students as well as modifying their instruction. Students who left STEM had high mismatch between their encountered and preferred instructional strategies, namely prefering more active instructional approaches and encountering more lecture. This was especially true for women and URM students who left STEM. These results indicate that incorporating more active instructional strategies could be one helpful avenue for improving retention of students from groups underrepresented in STEM.

Effectiveness of instruction is often gauged using research-based assessments. When these types of assessments produce performance gaps, a myriad of conclusions could be drawn, some of which could be problematic. The most damaging takeaways appear in the form of assumptions about who can and cannot do physics or STEM. This could lead people to justify underrepresentation or exclusion of certain groups in STEM, diminishing efforts to increase and address equity in the field. Alternatively, this could lead to creation of interventions or new instructional strategies that may be unproductive or misaligned with reality. Appearance of performance gaps could be due to bias in assessment items, as has been found in the *Force Concept Inventory* [5, 6]. Thus, addressing bias in assessment items presents potential for diminishing performance gaps and possible resulting problematic conclusions that could be drawn from them. This calls for reevaluating how assessments are designed and validated.

The assessment developed as part of this dissertation—the Upper-level Statistical Mechanics and Thermodynamics Evaluation for Physics (U-STEP)—discussed in Ch. 2-Ch. 5 was directly influenced by knowledge of the differential impacts that race and gender have on STEM students. This informed the ways in which bias was intentionally addressed during the item development phase, and the ways in which participants were solicited at every step of the process. In the U-STEP, a relatively novel approach was taken for design and validation of items. The goal to address potential bias in the assessment was at the forefront throughout the development cycle. First, thoughtful and deliberate recruitment at every phase of the development process ensured that our

data came from a diverse pool of participants, based on student populations served and department size. We solicited perspectives from many instructors across the country and incorporated their input into the writing of assessment objectives, use of notation in the assessment, and development of items. This allowed us to narrow the scope of the assessment such that it could be useful for a broad range of instructors and institutions.

We also incorporated differential item functioning (DIF), which fostered identification of bias in items during the development process through analyzing differential performance between students of similar abilities but different races or genders. This approach helps in curbing appearance of gaps due to bias, as is often seen with the FCI. In addition to DIF rooted in classical test theory (CTT), we also layed the groundwork for Rasch item response theory analyses to allow for less population-dependent validation techniques and future, more robust DIF analyses. This will aid in the endeavor to make the assessment more broadly usable, more reliable for comparisons across groups and institutions, and less susceptible to bias once it is finalized after the Spring 2021 pilot administration.

The U-STEP will be the first thermal physics assessment that addresses both classical thermodynamics and statistical mechanics. It also adds to the limited pool of upper-division physics assessments composed of coupled, multiple-response items, which allow for automated, streamlined scoring and partial credit. Though traditional routes for validation such as CTT were utilized, the U-STEP is one of the first upper-division physics assessments to investigate DIF and Rasch analyses during the development phase.

The studies presented here propose new best practices in assessment development and equity work in PER. Development of validated, research-based assessments, and interpretations of their results, should be informed by qualitative work regarding differential impacts of identity on STEM students. Despite evidence of these differential experiences based on race and gender in STEM, we found that many students are unaware of these differences. This is especially true for White men, who we found to be the least likely to recognize race- and gender-based differences in STEM experiences. Since White men hold the majority of power in STEM, due in part to their overrep-

resentation [10], they have the most power in influencing changes in STEM culture and practices. Thus, White men (along with others in STEM) should be informed about power, privilege, and differential impacts on students pursuing STEM degrees based on race and gender. This will foster more effective change in the realms of both equity and assessment design. The work presented here can be used to inform changes in dominant STEM culture, resulting in evolved norms with regards to addressing underrepresentation and equity, in addition to assessment development practices.

Bibliography

[1] D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," The Physics Teacher, vol. 30, no. 3, pp. 141–158, 1992.

[2] R. K. Thornton and D. R. Sokoloff, "Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula," American Journal of Physics, vol. 66, no. 4, pp. 338–352, 1998.

[3] Multiple-choice assessment for upper-division electricity and magnetism, Physics Education Research Conference, (Portland, Oregon), 2013.

[4] A. Madsen, S. B. McKagan, and E. C. Sayre, "Resource letter RBAI-1: research-based assessment instruments in physics and astronomy," American Journal of Physics, vol. 85, no. 4, pp. 245–264, 2017.

[5] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis, "Gender bias in the force concept inventory?," in AIP Conference Proceedings, vol. 1413, pp. 171–174, American Institute of Physics, 2012.

[6] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, "Gender fairness within the force concept inventory," Physical Review Physics Education Research, vol. 14, no. 1, p. 010103, 2018.

[7] B. R. Wilcox and S. J. Pollock, "Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics," Physical Review Special Topics-Physics Education Research, vol. 10, no. 2, p. 020124, 2014.

[8] C. Walsh, K. N. Quinn, C. Wieman, and N. G. Holmes, "Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking," Physical Review Physics Education Research, vol. 15, no. 1, p. 010135, 2019.

[9] B. Pollard, M. F. J. Fox, L. Ríos, and H. J. Lewandowski, "Creating a coupled multiple response assessment for modeling in lab courses," in Physics Education Research Conference Proceedings, Physics Education Research Conference, (Virtual Conference), July 22-23 2020.

[10] National Science Foundation, "Women, minorities, and persons with disabilities in science and engineering," 2017.

[11] L. McCullough, "Gender, context, and physics assessment," Journal of International Women's Studies, vol. 5, no. 4, pp. 20–30, 2004.

[12] R. R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," American Journal of Physics, vol. 66, no. 1, pp. 64–74, 1998.

[13] R. R. Hake, "Relationship of individual student normalized learning gains in mechanics with gender, high-school physics, and pretest scores on mathematics and spatial visualization," in Physics Education Research Conference, vol. 8, pp. 1–14, 2002.

[14] J. L. Smith, K. L. Lewis, L. Hawthorne, and S. D. Hodges, "When trying hard isn't natural: Women's belonging with and motivation for male-dominated stem fields as a function of effort expenditure concerns," Personality and Social Psychology Bulletin, vol. 39, no. 2, pp. 131–143, 2013.

[15] T. L. Strayhorn, College students' sense of belonging: A key to educational success for all students. Routledge, 2012.

[16] E. Seymour and N. M. Hewitt, Talking about leaving: Factors contributing to high attrition rates among science, mathematics & engineering undergraduate majors: final report to the Alfred P. Sloan Foundation on an ethnographic inquiry at seven institutions. Ethnography and Assessment Research, Bureau of Sociological Research, 1994.

[17] D. B. Thoman, J. A. Arizaga, J. L. Smith, T. S. Story, and G. Soncuya, "The grass is greener in non-science, technology, engineering, and math classes: Examining the role of competing belonging to undergraduate women's vulnerability to being pulled away from science," Psychology of Women Quarterly, vol. 38, no. 2, pp. 246–258, 2014.

[18] G. M. Walton and G. L. Cohen, "A brief social-belonging intervention improves academic and health outcomes of minority students," Science, vol. 331, no. 6023, pp. 1447–1451, 2011.

[19] K. D. Rainey, M. Vignal, and B. R. Wilcox, "Designing upper-division thermal physics assessment items informed by faculty perspectives of key content coverage," Physical Review Physics Education Research, vol. 16, p. 020113, Aug 2020.

[20] K. Rainey, M. Dancy, R. Mickelson, E. Stearns, and S. Moller, "Race and gender differences in how sense of belonging influences decisions to major in STEM," International Journal of STEM Education, vol. 5, no. 1, p. 10, 2018.

[21] K. Rainey, M. Dancy, R. Mickelson, E. Stearns, and S. Moller, "A descriptive study of race and gender differences in how instructional style and perceived professor care influence decisions to major in STEM," International Journal of STEM Education, vol. 6, no. 1, pp. 1–13, 2019.

[22] M. Dancy, K. Rainey, E. Stearns, R. Mickelson, and S. Moller, "Undergraduates' awareness of White and male privilege in STEM," International Journal of STEM Education, vol. 7, no. 1, pp. 1–17, 2020.

[23] N. R. Council, Next Generation Science Standards: For States, By States. The National Academies Press, Washington DC, 2013.

[24] A. Savinainen and P. Scott, "The Force Concept Inventory: A tool for monitoring student learning," Physics Education, vol. 37, no. 1, p. 45, 2002.

[25] B. W. Dreyfus, B. D. Geller, D. E. Meltzer, and V. Sawtelle, "Resource letter TTSM-1: Teaching thermodynamics and statistical mechanics in introductory physics, chemistry, and biology," American Journal of Physics, vol. 83, no. 1, pp. 5–21, 2015.

[26] J. W. Clark, J. R. Thompson, and D. B. Mountcastle, "Comparing student conceptual understanding of thermodynamics in physics and engineering," in AIP Conference Proceedings, vol. 1513, pp. 102–105, American Institute of Physics, 2013.

[27] S. Yeo and M. Zadnik, "Introductory thermal concept evaluation: Assessing students' understanding," The Physics Teacher, vol. 39, no. 8, pp. 496–504, 2001.

[28] H.-E. Chu, D. F. Treagust, S. Yeo, and M. Zadnik, "Evaluation of students' understanding of thermal concepts in everyday contexts," International Journal of Science Education, vol. 34, no. 10, pp. 1509–1534, 2012.

[29] C. Tanahoung, R. Chitaree, C. Soankwan, M. Sharma, and I. Johnston, "Surveying Thai and Sydney introductory physics students' understandings of heat and temperature," (Sydney, Australia), pp. 29–53, Proceedings of The Australian Conference on Science and Mathematics Education (formerly UniServe Science Conference), 2012.

[30] P. Wattanakasiwich, P. Taleab, M. D. Sharma, and I. D. Johnston, "Construction and implementation of a conceptual survey in thermodynamics," vol. 21, no. 1, 2013.

[31] B. Brown, Developing and assessing research-based tools for teaching quantum mechanics and thermodynamics. PhD thesis, University of Pittsburgh, 2015.

[32] C. H. Kautz, P. R. Heron, M. E. Loverude, and L. C. McDermott, "Student understanding of the ideal gas law, Part I: A macroscopic perspective," American Journal of Physics, vol. 73, no. 11, pp. 1055–1063, 2005.

[33] C. H. Kautz, P. R. Heron, P. S. Shaffer, and L. C. McDermott, "Student understanding of the ideal gas law, Part II: A microscopic perspective," American Journal of Physics, vol. 73, no. 11, pp. 1064–1071, 2005.

[34] K. Bain, A. Moon, M. R. Mack, and M. H. Towns, "A review of research on the teaching and learning of thermodynamics at the university level," Chemistry Education Research and Practice, vol. 15, no. 3, pp. 320–335, 2014.

[35] P. G. Jasien and G. E. Oberem, "Understanding of elementary concepts in heat and temperature among college students and K-12 teachers," Journal of Chemical Education, vol. 79, no. 7, p. 889, 2002.

[36] M. Sözbilir, "A review of selected literature on students' misconceptions of heat and temperature," Boğaziçi University Journal of Education, vol. 20, no. 1, pp. 25–41, 2003.

[37] A. A. Alwan, "Misconception of heat and temperature among physics students," Procedia-Social and Behavioral Sciences, vol. 12, pp. 600–614, 2011.

[38] R. Leinonen, M. A. Asikainen, and P. E. Hirvonen, "Overcoming students' misconceptions concerning thermal physics with the aid of hints and peer interaction during a lecture course," Physical Review Special Topics-Physics Education Research, vol. 9, no. 2, p. 020112, 2013.

[39] D. E. Meltzer, "Investigation of students' reasoning regarding heat, work, and the first law of thermodynamics in an introductory calculus-based general physics course," American Journal of Physics, vol. 72, no. 11, pp. 1432–1446, 2004.

[40] D. E. Meltzer, "Investigation of student reasoning regarding concepts in thermal physics," APS Forum on Education, pp. 4–5, 2005.

[41] M. E. Loverude, C. H. Kautz, and P. R. Heron, "Student understanding of the first law of thermodynamics: Relating work to the adiabatic compression of an ideal gas," American Journal of Physics, vol. 70, no. 2, pp. 137–148, 2002.

[42] T. I. Smith, W. M. Christensen, D. B. Mountcastle, and J. R. Thompson, "Identifying student difficulties with entropy, heat engines, and the Carnot cycle," Physical Review Special Topics-Physics Education Research, vol. 11, no. 2, p. 020116, 2015.

[43] B. R. Bucy, J. R. Thompson, and D. B. Mountcastle, "What is entropy? advanced undergraduate performance comparing ideal gas processes," in AIP Conference Proceedings, vol. 818, pp. 81–84, American Institute of Physics, 2006.

[44] W. M. Christensen, D. E. Meltzer, and C. Ogilvie, "Student ideas regarding entropy and the second law of thermodynamics in an introductory physics course," American Journal of Physics, vol. 77, no. 10, pp. 907–917, 2009.

[45] R. Leinonen, M. A. Asikainen, and P. E. Hirvonen, "Grasping the second law of thermodynamics at university: The consistency of macroscopic and microscopic explanations," Physical Review Special Topics-Physics Education Research, vol. 11, no. 2, p. 020122, 2015.

[46] M. Loverude, "Identifying student resources in reasoning about entropy and the approach to thermal equilibrium," Physical Review Special Topics-Physics Education Research, vol. 11, no. 2, p. 020118, 2015.

[47] M. Sözbilir, "What students' understand from entropy?: A review of selected literature," Journal of Baltic Science Education, vol. 2, no. 1, pp. 21–27, 2003.

[48] E. M. Carson and J. R. Watson, "Undergraduate students' understandings of entropy and gibbs free energy," University Chemistry Education, vol. 6, no. 1, pp. 4–12, 2002.

[49] F. L. Lambert, "Disorder-a cracked crutch for supporting entropy discussions," Journal of Chemical Education, vol. 79, no. 2, p. 187, 2002.

[50] R. Wei, W. Reed, J. Hu, and C. Xu, "Energy spreading or disorder? understanding entropy from the perspective of energy," in Teaching and learning of energy in K–12 education, pp. 317–335, Springer, 2014.

[51] D. E. Meltzer, "Observations of general learning patterns in an upper-level thermal physics course," in AIP Conference Proceedings, vol. 1179, pp. 31–34, American Institute of Physics, 2009.

[52] T. I. Smith, D. B. Mountcastle, and J. R. Thompson, "Student understanding of the Boltzmann factor," Physical Review Special Topics-Physics Education Research, vol. 11, no. 2, p. 020123, 2015.

[53] T. I. Smith, J. R. Thompson, and D. B. Mountcastle, "Student understanding of Taylor series expansions in statistical mechanics," Physical Review Special Topics-Physics Education Research, vol. 9, no. 2, p. 020110, 2013.

[54] T. A. Moore and D. V. Schroeder, "A different approach to introducing statistical mechanics," American Journal of Physics, vol. 65, no. 1, pp. 26–36, 1997.

[55] C. Kamcharean and P. Wattanakasiwich, "Development and application of thermodynamics diagnostic test to survey students' understanding in thermal physics," International Journal of Innovation in Science and Mathematics Education, vol. 24, no. 2, 2016.

[56] B. Brown and C. Singh, "Development and validation of a conceptual survey instrument to evaluate students' understanding of thermodynamics," Physical Review Physics Education Research, vol. 17, no. 1, p. 010104, 2021.

[57] B. R. Wilcox, M. D. Caballero, C. Baily, H. Sadaghiani, S. V. Chasteen, Q. X. Ryan, and S. J. Pollock, "Development and uses of upper-division conceptual assessments," Physical Review Special Topics-Physics Education Research, vol. 11, no. 2, p. 020115, 2015.

[58] R. Hambleton and J. Rogers, "Item bias review," Practical Assessment, Research, and Evaluation, vol. 4, no. 1, p. 6, 1994.

[59] P. V. Engelhardt, "An introduction to classical test theory as applied to conceptual multiple-choice tests," Getting started in PER, vol. 2, no. 1, 2009.

[60] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, "Characterizing the gender gap in introductory physics," Physical Review Special Topics-Physics Education Research, vol. 5, no. 1, p. 010101, 2009.

[61] S. J. Pollock, N. D. Finkelstein, and L. E. Kost, "Reducing the gender gap in the physics classroom: How sufficient is interactive engagement?," Physical Review Special Topics-Physics Education Research, vol. 3, no. 1, p. 010107, 2007.

[62] M. Lorenzo, C. H. Crouch, and E. Mazur, "Reducing the gender gap in the physics classroom," American Journal of Physics, vol. 74, no. 2, pp. 118–122, 2006.

[63] G. Rasch, Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1980.

[64] R. Baierlein, Thermal Physics. Cambridge University Press, 1999.

[65] A. Carter, Classical and Statistical Thermodynamics. Prentice Hall, 2001.

[66] C. Kittel and H. Kroemer, Thermal Physics. W. H. Freeman & Co., 1980.

[67] F. Sears and G. Salinger, Thermodynamics, Kinetic Theory and Statistical Thermodynamics. Addison-Wesley Publishing Co., 1975.

[68] D. V. Schroeder, An Introduction to Thermal Physics. Addison Wesley, 1999.

[69] M. Zemansky and D. Dittman, Heat and Thermodynamics. McGraw-Hili, New York, NY, 1997.

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

[71] "Top Educators." https://www.aps.org/programs/education/statistics/topproducers.cfm, 2019.

[72] "2017 Physics Degree-Granting Minority-Serving Institutions." https://www.aps.org/programs/education/statistics/upload/List-HBCUs-BSIs-HSIs-2017.pdf, 2017.

[73] R. Ivie and K. Stowe, "Women in Physics, 2000," American Institute of Physics, no. R-430, 2000.

[74] PhysNet: The physics departments and documents network, "Physics Departments in the United States of America." http://de.physnet.net/PhysNet/us.html, 2013.

[75] "Carnegie classifications." http://carnegieclassifications.iu.edu/, 2018.

[76] F. Reif, Fundamentals of statistical and thermal physics. Waveland Press, 2009.

[77] S. E. Stemler, Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource: Content analysis. John Wiley & Sons, Inc., 2015.

[78] K. A. Ericsson and H. A. Simon, "Verbal reports as data," Psychological review, vol. 87, no. 3, p. 215, 1980.

[79] R. Jääskeläinen, "Think-aloud protocol," Handbook of translation studies, vol. 1, pp. 371–374, 2010.

[80] J. W. Cresswell, "Research Design: Qualitative, Quantitative, and Mixed Methods Approaches," pp. 183–213, SAGE Publications, Inc., 4th ed., 2014.

[81] B. R. Wilcox and S. J. Pollock, "Validation and analysis of the coupled multiple response colorado upper-division electrostatics diagnostic," Physical Review Special Topics-Physics Education Research, vol. 11, no. 2, p. 020130, 2015.

[82] M. W. Post, "What to do with "moderate" reliability and validity coefficients?," Archives of physical medicine and rehabilitation, vol. 97, no. 7, pp. 1051–1052, 2016.

[83] L. Ding and R. Beichner, "Approaches to data analysis of multiple-choice questions," Physical Review Special Topics-Physics Education Research, vol. 5, no. 2, p. 020103, 2009.

[84] C. Spearman, "The proof and measurement of association between two things.," vol. 15, no. 1, pp. 72–101, 1904.

[85] J. M. Cortina, "What is coefficient alpha? an examination of theory and applications.," Journal of Applied Psychology, vol. 78, no. 1, p. 98, 1993.

[86] F. Wilcoxon, "Individual comparisons by ranking methods," in Breakthroughs in statistics, pp. 196–202, Springer, 1992.

[87] R. Gutiérrez, "A "gap-gazing" fetish in mathematics education? problematizing research on the achievement gap," Journal for Research in Mathematics Education, pp. 357–364, 2008.

[88] X. An and Y.-F. Yung, "Item response theory: What it is and how you can use the irt procedure to apply it," SAS Institute Inc. SAS364-2014, vol. 10, no. 4, 2014.

[89] A. Sahin and D. Anil, "The effects of test length and sample size on item parameters in item response theory," 2017.

[90] M. R. Harwell and J. E. Janosky, "An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG," Applied Psychological Measurement, vol. 15, no. 3, pp. 279–291, 1991.

[91] R. P. Chalmers, "mirt: A multidimensional item response theory package for the r environment," Journal of Statistical Software, vol. 48, no. 6, pp. 1–29, 2012.

[92] A. Maydeu-Olivares and H. Joe, "Limited-and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework," Journal of the American Statistical Association, vol. 100, no. 471, pp. 1009–1020, 2005.

[93] Y. Xia and Y. Yang, "RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods," Behavior research methods, vol. 51, no. 1, pp. 409–428, 2019.

[94] M. W. Browne, R. Cudeck, K. A. Bollen, and J. S. Long, "Testing structural equation models," 1993.

[95] K. G. Jöreskog and D. Sörbom, LISREL 8: Structural equation modeling with the SIMPLIS command language. Scientific Software International, 1993.

[96] L.-t. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," Structural equation modeling: a multidisciplinary journal, vol. 6, no. 1, pp. 1–55, 1999.

[97] M. Orlando and D. Thissen, "Likelihood-based item-fit indices for dichotomous item response theory models," Applied Psychological Measurement, vol. 24, no. 1, pp. 50–64, 2000.

[98] J. Bruin, "newtest: command to compute new test @ONLINE," Feb. 2011.

[99] T. Kang and T. T. Chen, "Performance of the generalized s-x2 item fit index for polytomous irt models," Journal of Educational Measurement, vol. 45, no. 4, pp. 391–406, 2008.

[100] N. W. Scott, P. M. Fayers, N. K. Aaronson, A. Bottomley, A. de Graeff, M. Groenvold, C. Gundy, M. Koller, M. A. Petersen, M. A. Sprangers, et al., "A simulation study provided sample size guidance for differential item functioning (dif) studies using short scales," Journal of Clinical Epidemiology, vol. 62, no. 3, pp. 288–295, 2009.

[101] I. Paek and M. Wilson, "Formulating the rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with mantel–haenszel procedure in short test and small sample conditions," Educational and Psychological Measurement, vol. 71, no. 6, pp. 1023–1046, 2011.

[102] G. N. Masters, "A rasch model for partial credit scoring," Psychometrika, vol. 47, no. 2, pp. 149–174, 1982.

[103] D. Rizopoulos, "ltm: An r package for latent variable modeling and item response theory analyses," Journal of Statistical Software, vol. 17, no. 5, pp. 1–25, 2006.

[104] T. Kutscher, M. Eid, and C. Crayen, "Sample size requirements for applying mixed polytomous item response models: results of a monte carlo simulation study," Frontiers in Psychology, vol. 10, p. 2494, 2019.

[105] K. D. Rainey, A. P. Jambuge, J. T. Laverty, and B. R. Wilcox, "Developing coupled, multiple-response assessment items addressing scientific practices," in Physics Education Research Conference Proceedings, Physics Education Research Conference, (Virtual Conference), July 22-23 2020.

[106] L. Holman, D. Stuart-Fox, and C. E. Hauser, "The gender gap in science: How long until women are equally represented?," PLoS Biology, vol. 16, no. 4, p. e2004956, 2018.

[107] F. A. Hrabowski, "Boosting minorities in science," 2011.

[108] C. Kessel and D. J. Nelson, "Statistical trends in women's participation in science: Commentary on Valla and Ceci (2011)," Perspectives on Psychological Science, vol. 6, no. 2, pp. 147–149, 2011.

[109] United Nations Educational, Scientific and Cultural Organization, "Women, minorities, and persons with disabilities in science and engineering," 2015.

[110] E. Seymour and A.-B. Hunter, "Talking about leaving revisited," Talking About Leaving Revisited: Persistence, Relocation, and Loss in Undergraduate STEM Education, 2019.

[111] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, "Active learning increases student performance in science, engineering, and mathematics," Proceedings of the National Academy of Sciences, vol. 111, no. 23, pp. 8410–8415, 2014.

[112] M. Prince, "Does active learning work? a review of the research," Journal of Engineering Education, vol. 93, no. 3, pp. 223–231, 2004.

[113] K. M. Cooper, V. R. Downing, and S. E. Brownell, "The influence of active learning practices on student anxiety in large-enrollment college science classrooms," International Journal of STEM Education, vol. 5, no. 1, pp. 1–18, 2018.

[114] D. Allen and K. Tanner, "Infusing active learning into the large-enrollment biology class: seven strategies, from the simple to complex," Cell biology education, vol. 4, no. 4, pp. 262–268, 2005.

[115] A. Madsen, S. B. McKagan, and E. C. Sayre, "Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?," Physical Review Special Topics-Physics Education Research, vol. 9, no. 2, p. 020121, 2013.

[116] R. J. Beichner, J. M. Saul, D. S. Abbott, J. J. Morse, D. Deardorff, R. J. Allain, S. W. Bonham, M. H. Dancy, and J. S. Risley, "The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project," Research-based Reform of University Physics, vol. 1, no. 1, pp. 2–39, 2007.

[117] J. Roberts and R. Styron, "Student satisfaction and persistence: Factors vital to student retention," Research in Higher Education Journal, vol. 6, p. 1, 2010.

[118] M. Micari and P. Pazos, "Connecting to the professor: Impact of the student–faculty relationship in a highly challenging course," College Teaching, vol. 60, no. 2, pp. 41–47, 2012.

[119] J. N. Olson and J. A. Carter, "Caring and the college professor," in National Forum Journals: Focus on Colleges, Universities, and Schools, vol. 8, 2014.

[120] W. Buskist, J. Sikorski, T. Buckley, and B. K. Saville, "Elements of master teaching," The teaching of psychology: Essays in honor of Wilbert J. McKeachie and Charles L. Brewer, vol. 1, pp. 27–39, 2002.

[121] S. A. Meyers, "Do your students care whether you care about them?," College Teaching, vol. 57, no. 4, pp. 205–210, 2009.

[122] L. Muraskin, J. Lee, A. Wilner, and W. Swail, "Raising the graduation rates of low-income college students." The Pell Institute for the Study of Opportunity in Higher Education, 2004.

[123] T. A. Benson, A. L. Cohen, and W. Buskist, "Rapport: Its relation to student attitudes and behaviors toward teachers and classes," Teaching of Psychology, vol. 32, pp. 237–239, 2005.

[124] T. L. Strayhorn, "Factors influencing black males' preparation for college and success in stem majors: A mixed methods study.," Western Journal of Black Studies, vol. 39, no. 1, 2015.

[125] C. Goodenow, "Classroom belonging among early adolescent students: Relationships to motivation and achievement," The Journal of Early Adolescence, vol. 13, no. 1, pp. 21–43, 1993.

[126] C. Good, A. Rattan, and C. S. Dweck, "Why do women opt out? sense of belonging and women's representation in mathematics.," Journal of Personality and Social Psychology, vol. 102, no. 4, p. 700, 2012.

[127] D. R. Johnson, "Campus racial climate perceptions and overall sense of belonging among racially diverse women in stem majors," Journal of College Student Development, vol. 53, no. 2, pp. 336–346, 2012.

[128] G. M. Walton and G. L. Cohen, "A question of belonging: race, social fit, and achievement.," Journal of Personality and Social Psychology, vol. 92, no. 1, p. 82, 2007.

[129] M. C. Murphy, C. M. Steele, and J. J. Gross, "Signaling threat: How situational cues affect women in math, science, and engineering settings," Psychological Science, vol. 18, no. 10, pp. 879–885, 2007.

[130] K. L. Lewis, J. G. Stout, S. J. Pollock, N. D. Finkelstein, and T. A. Ito, "Fitting in or opting out: A review of key social-psychological factors influencing a sense of belonging for women in physics," Physical Review Physics Education Research, vol. 12, no. 2, p. 020110, 2016.

[131] L. Espinosa, "Pipelines and pathways: Women of color in undergraduate stem majors and the college experiences that contribute to persistence," Harvard Educational Review, vol. 81, no. 2, pp. 209–241, 2011.

[132] J. Weidman, "Undergraduate socialization: A conceptual approach," Higher Education: Handbook of Theory and Research, vol. 5, no. 2, pp. 289–322, 1989.

[133] E. Wenger, Communities of practice: Learning, meaning, and identity. Cambridge University Press, 1999.

[134] J. Boaler, "The development of disciplinary relationships: Knowledge, practice and identity in mathematics classrooms," For the Learning of Mathematics, vol. 22, no. 1, pp. 42–47, 2002.

[135] Z. Hazari, P. M. Sadler, and G. Sonnert, "The science identity of college students: Exploring the intersection of gender, race, and ethnicity," Journal of College Science Teaching, vol. 42, no. 5, pp. 82–91, 2013.

[136] J. P. Gee, "Chapter 3: Identity as an analytic lens for research in education," Review of Research in Education, vol. 25, no. 1, pp. 99–125, 2000.

[137] S. Chen, K. R. Binning, K. J. Manke, S. T. Brady, E. M. McGreevy, L. Betancur, L. B. Limeri, and N. Kaufmann, "Am i a science person? a strong science identity bolsters minority students' sense of belonging and performance in college," Personality and Social Psychology Bulletin, pp. 1–14, 2020.

[138] J. W. Osborne and C. Walker, "Stereotype threat, identification with academics, and withdrawal from school: Why the most successful students of colour might be most likely to withdraw," Educational Psychology, vol. 26, no. 4, pp. 563–577, 2006.

[139] H. B. Carlone, "The cultural production of science in reform-based physics: Girls' access, participation, and resistance," Journal of Research in Science Teaching, vol. 41, no. 4, pp. 392–414, 2004.

[140] A. Bandura, "Self-efficacy," The Corsini encyclopedia of psychology, pp. 1–3, 1994.

[141] J. M. Blaney and J. G. Stout, "Examining the relationship between introductory computing course experiences, self-efficacy, and belonging among first-generation college women," in Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, pp. 69–74, 2017.

[142] S. Beyer, "Why are women underrepresented in computer science? Gender differences in stereotypes, self-efficacy, values, and interests and predictors of future CS course-taking and grades," Computer Science Education, vol. 24, no. 2-3, pp. 153–192, 2014.

[143] V. Sawtelle, A gender study investigating physics self-efficacy. PhD thesis, Florida International University, 2011.

[144] U. Tellhed, M. Bäckström, and F. Björklund, "Will I fit in and do well? The importance of social belongingness and self-efficacy for explaining gender differences in interest in STEM and HEED majors," Sex Roles, vol. 77, no. 1-2, pp. 86–96, 2017.

[145] S. Rugheimer, "Women in STEM resources," 2019.

[146] D. Dortch and C. Patel, "Black undergraduate women and their sense of belonging in stem at predominantly white institutions," NASPA Journal About Women in Higher Education, vol. 10, no. 2, pp. 202–215, 2017.

[147] C. Hill, C. Corbett, and A. St Rose, Why so few? Women in science, technology, engineering, and mathematics. ERIC, 2010.

[148] A. J. Fisher, R. Mendoza-Denton, C. Patt, I. Young, A. Eppig, R. L. Garrell, D. C. Rees, T. W. Nelson, and M. A. Richards, "Structure and belonging: Pathways to success for underrepresented minority and women PhD students in STEM fields," PLoS One, vol. 14, no. 1, p. e0209279, 2019.

[149] C. Funk and K. Parker, "Women and men in STEM often at odds over workplace equity," 2018.

[150] M. I. Norton and S. R. Sommers, "Whites see racism as a zero-sum game that they are now losing," Perspectives on Psychological Science, vol. 6, no. 3, pp. 215–218, 2011.

[151] L. R. Tropp and F. K. Barlow, "Making advantaged racial groups care about inequality: Intergroup contact as a route to psychological investment," Current Directions in Psychological Science, vol. 27, no. 3, pp. 194–199, 2018.

[152] F. Crosby, "The denial of personal discrimination," American behavioral scientist, vol. 27, no. 3, pp. 371–386, 1984.

[153] J. Jost and O. Hunyady, "The psychology of system justification and the palliative function of ideology," European Review of Social Psychology, vol. 13, no. 1, pp. 111–153, 2003.

[154] R. Mickelson, A. Parker, E. Stearns, S. Moller, and M. Dancy, "Family matters: Familial support and African American female success," Contemporary African American families: achievements, challenges, and empowerment strategies in the 21st century. New York: Routledge-Taylor & Francis, 2015.

[155] H. B. Carlone and A. Johnson, "Understanding the science experiences of successful women of color: Science identity as an analytic lens," Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, vol. 44, no. 8, pp. 1187–1218, 2007.

[156] CoNECD, Critiquing the "Underrepresented Minorities" Label: Disrupting Inequity, (Crystal City, Virginia), 2018.

[157] K. Crenshaw, "Mapping the margins: Identity politics, intersectionality, and violence against women," Stanford Law Review, vol. 43, no. 6, pp. 1241–1299, 1991.

[158] P. H. Collins, "Intersectionality's definitional dilemmas," Annual Review of Sociology, vol. 41, pp. 1–20, 2015.

[159] L. Bowleg, "When black+ lesbian+ woman≠ black lesbian woman: The methodological challenges of qualitative and quantitative intersectionality research," Sex Roles, vol. 59, no. 5-6, pp. 312–325, 2008.

[160] NCCJ, "What is privilege?." National Conference for Community and Justice, 2018.

[161] E. Bonilla-Silva, Racism without racists: Color-blind racism and the persistence of racial inequality in the United States. Rowman & Littlefield Publishers, 2006.

[162] C. W. Miller, B. M. Zwickl, J. R. Posselt, R. T. Silvestrini, and T. Hodapp, "Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion," Science Advances, vol. 5, no. 1, p. eaat7550, 2019.

[163] Syracuse University, "Impact of marginalization." http://counselingcenter.syr.edu/social-justice/impact-of-marginalization.html, 2018.

[164] S. Moller, N. Banerjee, M. C. Bottia, E. Stearns, R. A. Mickelson, M. Dancy, E. Wright, and L. Valentino, "Moving Latino/a students into STEM majors in college: The role of teachers and professional communities in secondary schools," Journal of Hispanic Higher Education, vol. 14, no. 1, pp. 3–33, 2015.

[165] E. Stearns, M. C. Bottia, J. Giersch, R. A. Mickelson, S. Moller, N. Jha, and M. Dancy, "Do relative advantages in STEM grades explain the gender gap in selection of a STEM major in college? A multimethod answer," American Educational Research Journal, vol. 57, no. 1, pp. 218–257, 2020.

[166] A. Strauss and J. M. Corbin, Grounded theory in practice. Sage, 1997.

[167] J. Ellis, M. L. Kelton, and C. Rasmussen, "Student perceptions of pedagogy and associated persistence in calculus," ZDM–Mathematics Education, vol. 46, no. 4, pp. 661–673, 2014.

[168] B. Gee and D. Peck, "The illusion of Asian success: Scant progress for minorities in cracking the glass ceiling from 2007–2015," Ascend: Pan-Asian Leaders, 2017.

[169] J. C. Williams, S. Li, R. Rincon, and P. Finn, "Climate control: Gender and racial bias in engineering?," Center for Worklife Law & Society of Women Engineers, 2016.

[170] C. M. Steele, "A threat in the air: How stereotypes shape intellectual identity and performance.," American Psychologist, vol. 52, no. 6, p. 613, 1997.

[171] P. Bourdieu, The logic of practice. Stanford University Press, 1990.

[172] A. Lareau, Unequal childhoods: Class, race, and family life. University of California Press, 2011.

[173] S. J. Ceci and W. M. Williams, "Women have substantial advantage in stem faculty hiring, except when competing against more-accomplished men," Frontiers in Psychology, vol. 6, p. 1532, 2015.

[174] S. Vogel and L. Schwabe, "Learning and memory under stress: implications for the classroom," NPJ Science of Learning, vol. 1, no. 1, pp. 1–10, 2016.

[175] M. Deziel, D. Olawo, L. Truchon, and L. Golab, "Analyzing the mental health of engineering students using classification and regression," in Educational Data Mining 2013, 2013.

[176] C. Saravanan and R. Wilks, "Medical students' experience of and reaction to stress: the role of depression and anxiety," The Scientific World Journal, vol. 2014, 2014.

[177] A. Kotak, "Depression in the scientific and technical workforce." https://blogs.sciencemag.org/sciencecareers/2007/10/depression-in-t.html, 2019.

[178] T. J. Atherton, R. S. Barthelemy, W. Deconinick, M. L. Falk, S. Garmon, E. Long, m. Plisch, E. H. Simmons, and K. Reeves, "LGBT Climate in Physics Report: Building an Inclusive Community in Physics," American Physical Society, 2016.

[179] D. Green, "Historically underserved students: What we know, what we still need to know," New Directions for Community Colleges, vol. 2006, no. 135, pp. 21–28, 2006.

[180] R. Meyer Monhardt, "Fair play in science education: Equal opportunities for minority students," The Clearing House: A Journal of Educational Strategies, Issues and Ideas, vol. 74, no. 1, pp. 18–22, 2000.

[181] M. F. Fox, G. Sonnert, and I. Nikiforova, "Successful programs for undergraduate women in science and engineering: Adapting versus adopting the institutional environment," Research in Higher Education, vol. 50, no. 4, pp. 333–353, 2009.

[182] L. Malcom and S. Malcom, "The double bind: The next generation," Harvard Educational Review, vol. 81, no. 2, pp. 162–172, 2011.

[183] M. Ong, J. M. Smith, and L. T. Ko, "Counterspaces for women of color in stem higher education: Marginal and central spaces for persistence and success," Journal of Research in Science Teaching, vol. 55, no. 2, pp. 206–245, 2011.

[184] S. Hurtado, K. Eagan, J. H. Pryor, H. Whang, and S. Tran, "Undergraduate teaching faculty: The 2010–2011 HERI faculty survey," Higher Education Research Institute: University of California, Los Angeles, 2012.

[185] N. Apkarian, D. Kirin, D. Bressoud, C. Rasmussen, S. Larsen, J. Ellis, D. Ensley, and E. Johnson, "Progress through calculus: Census survey technical report," Retrieved from Mathematical Association of America website: https://bit. ly/PtCCensusReport, 2017.

Appendix A

Other Core Topics Frequencies

The tables on the following pages present all other core topics and supporting topics presented on the faculty content survey discussed in Sec. 3.1, along with their frequencies. Frequencies of other core topics are calculated using the total number of respondents (N=75). Supporting topic frequencies are calculated using the number of respondents selecting that particular core topic.

Table A.1: Frequencies of "other core topics" and their supporting topics, with the exception of statistical mechanics (see. Table A.2). Frequencies of other core topics are calculated using the total number of respondents (N=75). Supporting topic frequencies are calculated using the number of respondents selecting that particular core topic.

Other Core (& *Supporting*) Topics	%	Other Core (& *Supporting*) Topics	%
Blackbody Radiation (N=62)	83	**Magnetism** (N=49)	65
Photon gases	86	*Curie temperature*	74
Planck's law	98	*Demagnetization*	41
Stefan-Boltzmann law	95	*Ferromagnetism*	57
Wien's law	79	*Paramagnetism*	88
Bosons (N=62)	83	**Phases** (N=60)	80
Bose-Einstein condensates	81	*Equations of state*	90
Boson gases	76	*Phase changes*	92
Chemical Reactions (N=41)	55	*Phase diagrams*	95
Chemical equilibrium	95	*Phase equilibirum*	68
Dalton's law	29	*Phase rule*	28
Mass action	42	*Vaporization & vapor pressure*	57
Reaction equilibrium	54	**Pressure Diagrams** (N=54)	72
Conduction, Convection, Radiation (N=38)	51	*PT diagrams*	82
Emission	53	*PV diagrams*	98
Thermal conductivity	74	*PVT diagrams*	39
Cooling Techniques (N=22)	29	**Processes** (N=68)	91
Diatomic Molecules/Gases (N=59)	79	*Adiabatic processes*	99
Degree of freedom	100	*Compressions*	97
Rotational modes	95	*Expansions*	94
Vibrational modes	88	*Isobaric (constant pressure) processes*	96
Diffusion (N=33)	44	*Isochoric (constant volume) processes*	93
Diffusive equilibrium	91	*Isothermal processes*	100
Effusion	15	*Reversible & irreversible processes*	93
Osmosis	21	**Pure Substances** (N=35)	47
Random walk	64	**Quantum Phenomena** (N=56)	75
Fermions (N=63)	84	*deBroglie wavelength*	64
Fermi energy	89	*Pauli exclusion*	80
Fermion gas	86	*Uncertainty*	54
Free electrons	81	*Quantization*	71
Fluids (N=15)	20	**Scaling** (N=52)	69
Hydrostatic system	40	**Semiconductors** (N=8)	11
Superfluids	40	**Solids** (N=41)	55
Viscosity	20	*Debye solids*	81
Kinetic Theory (N=54)	72	*Einstein solids*	95
Collisions & cross sections	46	*Phonons*	63
Equilibrium constant	33		
Mean free path	70		
Molecular velocity distribution	93		

Table A.2: Frequencies for the "other core topic" *statistical mechanics* and its supporting topics. The frequency for *statistical mechanics* is calculated using the total number of respondents (N=75). Supporting topic frequencies are calculated using the number of respondents selecting *statistical mechanics*.

Other Core (& *Supporting*) Topics	%
Statistical Mechanics (N=73)	97
Boltzmann factor	95
Degeneracy	86
(In)Distinguishable particles	96
Lagrangian multipliers	15
Microstates & macrostates	100
Multiplicity	90
Partition function	96
Probability	95
Classical distribution	93
Bose-Einstein distribution	89
Gaussian distribution	52
Fermi-Dirac distribution	90
Maxwell-Boltzmann distribution	90
Planck distribution	71

Appendix B

Assessment Objectives for the U-STEP

The following sections present assessment objectives (AOs) for 10 key topical areas identified from results of the content survey distributed to physics faculty who teach or have taught upper-division thermal physics. Note that not all thermal physics content areas are contained within these AOs, but they encompass important themes within these 10 topical areas. Topical areas are presented in alphabetical order. In some cases, there is overlap between learning goals (e.g., heat and the first law); in these cases the AO is placed in the section most strongly linked.

These AOs were used to inform development of items for the U-STEP, though not all were addressed with the assessment items. Note some AOs were used to inform item construction directly within the targeted content area, while some addressed AOs were identified post-hoc as also being addressed by items outside of the targeted content area. *AOs addressed by U-STEP items are indiciated.

B.1 Energy

Thermodynamic Potentials

Internal Energy

(1) *Students can use the fact that internal energy is a state function to reason about thermal systems.

(2) *Students can determine internal energy for an ideal gas from equipartition.

   (a) *Students can identify that internal energy doesn't change from initial to final states for a cycle.

Enthalpy; Gibbs and Helmholtz Free Energy

(3) Students can identify the physical meaning of different terms composing the thermodynamic potentials.

   (a) Students can identify that the TS term in the expressions for Gibbs and Helmholtz free energies accounts for heat from surroundings.

   (b) Students can identify that the PV term in the expressions for enthalpy and Gibbs free energy accounts for work done by surroundings.

(4) Students can articulate that depending on what macroscopic quantities are held fixed different thermodynamic potentials are minimized.

   (a) Students can articulate that enthalpy, Gibbs free energy, and Helmholtz free energy are used for non-isolated systems that can interact with their surroundings.

Equipartition

(5) *Students can articulate that changes in temperature cause changes in internal energy.

   (a) *Students can determine internal energy change given final and initial temperature and the number of degrees of freedom within a system.

(6) Students can articulate that the internal energy of an ideal gas is determined by the number of degrees of freedom available to atoms/molecules comprising the gas system.

(7) *Students can use equipartition to determine characteristics of an ideal gas system (e.g., whether a gas is monatomic or polyatomic based on number of degrees of freedom).

(8) *Students can identify under what conditions the equipartition theorem applies.

B.2 Engines & Refrigerators

(1) Students can explain the purpose of heat engines.

(2) Students can define efficiency and/or coefficient of performance.

(3) *Students can explain why heat engine efficiency can never be 100%.

   (a) *Students can articulate that creation of entropy is ultimately responsible for the limits on maximum possible efficiency of a heat engine.

(4) Students can describe how the temperatures of the hot and cold reservoirs impact the maximum possible efficiency.

(5) *Students can interpret PV and TS diagrams for a heat engine cycle.

   (a) Students can determine from the direction of an engine cycle whether the engine is a heat engine or refrigerator.

   (b) *Students can describe what happens in an engine for each leg of a cycle shown in a PV diagram.

   (c) *Students can describe what happens in an engine for each leg of a cycle shown in a TS diagram.

B.3 Entropy & the Second Law of Thermodynamics

(1) *Students can identify and apply the appropriate expression for entropy to solve problems based on the physical nature of a system. Example expressions include:

   (a) *Classical: $dS \geq dQ/T$ or $dS \geq CdT/T$ (equality only holds for reversible processes)

   (b) *Statistical: Boltzmann's equation $\rightarrow S{=}k_B \ln \Omega$

(2) *Students can determine if particular processes will result in change of entropy of a system or of the universe.

   (a) *Students can determine if entropy changes with volume.

   (b) Students can determine if entropy changes with particle number.

   (c) *Students can determine if entropy changes with energy and heat.

   (d) *Students can determine if entropy changes for mixtures of two different gases.

   (e) *Students can determine if entropy changes for particular cycles.

(3) *Students can calculate changes in entropy for particular processes.

   (a) *Students can calculate change in entropy when volume changes.

   (b) Students can calculate change in entropy when particle number changes.

   (c) Students can calculate change in entropy when energy within a system changes.

   (d) *Students can calculate change in entropy when two different gases are mixed.

   (e) Students can calculate change in entropy for particular cycles.

(4) *Students can use the second law of thermodynamics to determine if a process is physically possible.

(5) *Students can apply the second law of thermodynamics.

  (a) *Students can articulate why the second law cannot be applied to non-isolated systems.

  (b) Students can distinguish between reversible and irreversible processes using the second law.

  (c) Students can describe spontaneous processes using the second law.

    (i) Students can articulate the relationship between irreversible processes, spontaneity, and the second law.

  (d) *Students can articulate/identify when invoking the second law is the appropriate approach for a given situation.

  (e) *Students can identify assumptions that must be true in order to productively apply the second law.

(6) Students can articulate the second law of thermodynamics in terms of microstates and probabilities.

(7) *Students can use the fact that entropy is a state function to reason about thermal systems.

B.4 Equilibrium

(1) *Students can identify the exchanged quantities involved in different types of equilibrium.

  (a) *Students can articulate that systems exchange **energy** when approaching **thermal** equilibrium.

  (b) *Students can articulate that systems exchange **volume** when approaching **mechanical** equilibrium.

  (c) *Students can articulate that systems exchange **particles** when approaching **diffusive** equilibrium.

(2) *Students can articulate the what quantities become equal for different types of equilibrium.

  (a) *Students can articulate that in **thermal** equilibrium **temperatures** are equal.

  (b) *Students can articulate that in **mechanical** equilibrium **pressures** or **forces** are equal.

  (c) *Students can articulate that in **diffusive** equilibrium **chemical potentials** are equal.

(3) *Students can recognize that interacting systems that have sat in contact for a *very long time* are in equilibrium.

B.5 The First Law of Thermodynamics

(1) Students can compare and contrast properties of heat and work.

(2) Students can articulate the physical significance of the first law of thermodynamics.

   (a) Students can articulate from the first law of thermodynamics that heat and work are both forms of energy transfer.

   (b) Students can articulate that the first law of thermodynamics accounts for all energy exchanged (e.g. energy is conserved) during a thermodynamic process.

(3) *Students can use the fact that heat and work can change internal energy when reasoning about thermal systems.

   (a) *Students can articulate and calculate how heat entering/leaving a system contributes to changes in internal energy.

   (b) *Students can articulate and calculate how work done on/by a system contributes to changes in internal energy.

   (c) *Students can articulate that doing work or having heat transfer does not necessitate that internal energy changes.

(4) *Students can explain and determine how work and heat contribute to changes in internal energy for different processes.

(5) *Students can determine characteristics of work and heat using the first law of thermodynamics and PV diagrams.

   (a) *Students can invoke the first law of thermodynamics to make conclusions about the sign of heat given information about internal energy changes and a PV diagram.

   (b) *Students can invoke the first law of thermodynamics to determine the magnitude of heat and direction of heat flow using a PV diagram for an ideal gas.

B.6 Gases

(1) Students can use the ideal gas law to relate changes in pressure, volume, and temperature.

   (a) Students can determine how one variable changes when changes in the other two variables are specified.

   (b) Students can identify that information on one or two variables only does not provide enough information to determine changes in the others.

(2) *Students can determine the number of degrees of freedom for different systems.

   (a) Students can articulate where degrees of freedom come from.

   (b) Students can articulate that some degrees of freedom get “frozen out” at certain temperatures.

   (c) Students can determine the number of degrees of freedom for monatomic and diatomic ideal gases.

(3) Students can use a PV diagram and the ideal gas law to determine temperature in terms of given variables.

  (a) Students can use a PV diagram to determine the change in temperature for a process using the ideal gas law.

  (b) Students can compare quantities using a PV diagram of an isothermal process using the ideal gas law.

(4) *Students can identify/articulate what assumptions must be true for a gas to be considered an ideal gas.

  (a) Students can articulate the assumptions that must be made to treat a gas as an ideal gas.

  (b) Students can determine when treating a real gas as an ideal gas is appropriate.

B.7 Heat

(1) Students can articulate that adding heat to a system doesn't always raise temperature of the system and use this to solve problems.

  (a) Students can provide or recognize examples where processes add heat but do not change temperature.

(2) *Students can articulate that heat is a flow of energy, caused by a temperature difference.

  (a) *Students can articulate why heat has no meaning at a single state.

    (i) *Students can articulate that since heat flow between two states depends on the path between those states, it is not a state variable.

(3) *Students can articulate that heat flows spontaneously from higher temperature systems to lower temperature systems and use this to solve problems.

  (a) Students can explain why heat flows spontaneously in a specific direction (e.g., maximizing entropy).

(4) *Students can use the definition of heat capacity to solve problems.

  (a) Students can determine the heat capacity at constant volume C_V.

  (b) Students can determine the heat capacity at constant pressure C_P.

(5) *Students can determine when heat enters or leaves a system...

  (a) *given physical constraints of a process.

  (b) *using a TS diagram.

  (c) *using a PV diagram.

B.8 Statistical Mechanics

(1) *Students can distinguish between microstates and macrostates.

(2) *Students can determine the number of **micro**states for a system.

  (a) Students can use combinatorics to calculate the number of microstates.

(3) *Students can determine the number of **macro**states for a system.

(4) Students can articulate the fundamental assumption of statistical mechanics and its significance.

(5) Students can manipulate large numbers.

  (a) Students can identify when a term or factor is small enough to be ignored relative to other relevant terms in an expression.

  (b) Students can apply logarithms to make very large numbers manageable.

    (i) Students can articulate the utility of logarithms when working with large and very large numbers.

(6) Students can distinguish between Bose-Einstein and Fermi-Dirac distributions on a graph.

(7) Students can determine the entropy and Helmholtz free energy of a system using the system's partition function.

(8) Students can determine the partition function for systems with degenerate states.

(9) Students can match terms in a system's partition function with a state's particular energy.

(10) *Students can determine and compare probabilities of states.

  (a) *Students can determine and compare probabilities of simple systems using counting.

  (b) *Students can determine and compare probabilities for a thermodynamic system using the Boltzmann factor.

  (c) *Students can determine and compare probabilities for a thermodynamic system after an energy shift using the Boltzmann factor.

(11) Students can distinguish between large and small systems.

  (a) Students can articulate the thermodynamic limit and when it is applicable.

  (b) Students can use graphical representations of states and probabilities to make conclusions about the physical nature of a system (e.g., multiplicity graphs or energy diagrams).

B.9 Temperature

(1) Students can articulate that objects in thermal equilibrium have the same temperature.

   (a) Students can identify that two objects that have been in thermal contact "for a very long time" must have the same temperature.

(2) *Students can articulate differences between heat and temperature.

   (a) *Students can articulate that temperature differences cause heat flow.

   (b) *Students can articulate that temperature is a property of a system and heat is not.

(3) Students can use basic kinetic theory to relate temperature and particle speeds.

(4) Students can determine the temperature of a system given entropy as a function of internal energy or internal energy as a function of entropy.

(5) Students can compare temperatures at two points using an entropy vs. internal energy graph.

(6) *Students can identify and interpret an isotherm on a PV diagram for an ideal gas.

(7) Students can identify the appropriate temperature scale (e.g. Kelvin, Celsius, Fahrenheit) for different applications and articulate why they are appropriate.

B.10 Work

(1) *Students can interpret and use PV diagrams to determine work.

   (a) *Students can determine the magnitude of work done on/by a system using a PV diagram.

   (b) *Students can determine whether work contributes to energy entering or leaving a system using a PV diagram.

   (c) Students can compare magnitudes of work done on/by a system by comparing areas under a curve on a PV diagram.

   (d) Students can determine *how* internal energy changes using a PV diagram showing an *adiabatic* process.

(2) Students can articulate why work has no meaning at a single state.

   (a) Students can articulate that since the amount of work done during a process depends on the specifics of that process, it is not a state variable.

Appendix C

U-STEP Pilot Demographics

Table C.1: Gender demographics for U-STEP pilot administrations. N-values are presented.

Gender	Fall 2019 (N=61)	Spring 2020 (N=169)	Fall 2020 (N=164)
Man	45	109	122
Woman	13	44	32
Non-binary/Other	3	4	4
Prefer not to Answer	0	12	6

Table C.2: Racial demographics for U-STEP pilot administrations. N-values are presented.

Race	Fall 2019 (N=61)	Spring 2020 (N=169)	Fall 2020 (N=164)
Asian	9	38	27
Underrepresented Minority	6	28	30
White	46	83	97
Prefer not to Answer	0	20	10

Appendix D

The U-STEP: The Full Assessment

Please complete the following diagnostic to the best of your ability. Your instructor will determine how the diagnostic contributes to your course grade. Your score will be determined by completion of the diagnostic as opposed to correctness. Your individual responses will not be provided to your instructor.

Some questions are multiple-choice, multiple-response, or a combination of both. When prompted, select **all** responses **that support your choice**.

Below are variables and their associated definitions. Unless otherwise specified, assume these meanings.

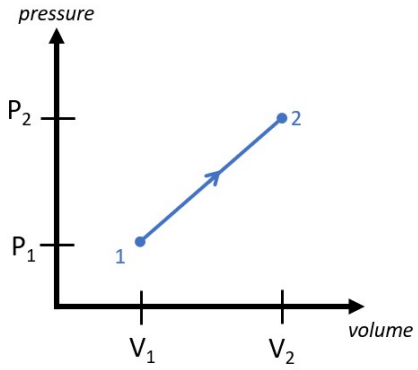
h – Planck's constant (6.63×10^{-34} J·s, 4.14×10^{-15} eV·s)
k_B – Boltzmann's constant (1.38×10^{-23} J/K, 8.62×10^{-5} eV/K)
m – mass
N – particle number
P – pressure
Q – heat
T – temperature
U – internal energy
V – volume
W – work

Please give your best effort. **Do not use your textbook or notes for reference.** Complete this diagnostic on your own in one sitting. Do **not** collaborate with others to complete this diagnostic.

If you have no idea how to answer a question, leave it blank. **Do not guess.**

Item 1

A **non-ideal gas** system undergoes a process taking it from point 1 to point 2, as indicated on the diagram below.



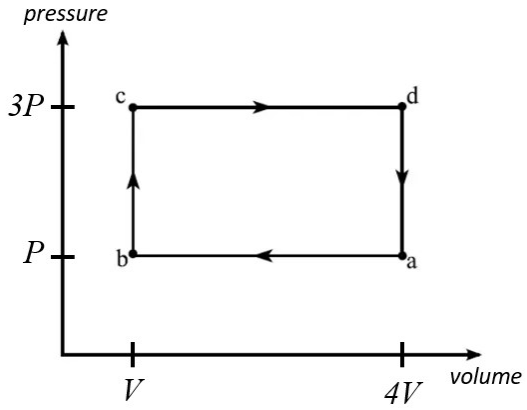
Did work contribute to energy entering or leaving the system?

Work caused energy to...
A) enter the system
B) leave the system
C) remain unchanged
D) not determinable

because... *(select all that support your response above)*
a) pressure increases
b) pressure decreases
c) volume increases
d) volume decreases
e) internal energy increases
f) internal energy decreases
g) the change in internal energy is unknown
h) the magnitude of work is directly proportional to the area under the curve for all gas systems
i) the magnitude of work is not directly proportional to the area under the curve for non-ideal gases

Item 2

Consider a **non-ideal** gas that undergoes the cycle below: a → b → c → d → a.



(i) The change in internal energy over one full cycle is...
- A) 0
- B) positive
- C) negative
- D) not enough information

because... *(select all that support your response above)*
- a) internal energy is a state function (i.e., process independent)
- b) temperature is a state function (i.e., process independent)
- c) the gas underwent one complete cycle
- d) the net work done on the system is zero
- e) the net work done on the system is non-zero
- f) no net heat flowed into or out of the system
- g) net heat flowed into or out of the system
- h) we would need to know how internal energy depends on pressure and volume

(ii) The magnitude of the net work done on the system is...
- A) 0
- B) 3PV
- C) 6PV
- D) 8PV
- E) 11PV
- F) not enough information

(iii) Does net heat enter or leave the system?
- A) net heat enters the system
- B) net heat leaves the system
- C) net heat does not enter or leave the system
- D) it depends
- E) not enough information

Item 3

Consider the following statement:

A thermodynamic system has a certain amount of heat,
just like it has a certain temperature, pressure, and volume

This statement is...
A) true
B) false

because... *(select all that support your response above)*
a) the amount of heat contained in a system can be calculated from the system's temperature, pressure, and volume
b) the amount of heat contained in a system can be calculated from changes in the system's temperature, pressure, and volume
c) the amount of heat contained in a system can be calculated from a system's heat capacity and temperature
d) heat is a quantity exchanged between systems
e) heat is a flow of thermal energy
f) heat is a scalar, like temperature, pressure, and volume
g) heat is not a state function *(i.e., heat is not process independent)*

Item 4

Consider a paramagnet consisting of N = 4 dipoles. Each dipole can be in one of two states: ↑ and ↓. All possible orientations of the dipoles are shown below.

↓↑↑↑	↓↓↑↑	↑↓↑↓	↓↓↓↑
↑↓↑↑	↑↓↓↑	↓↑↓↑	↓↓↑↓
↑↑↓↑	↑↑↓↓	↓↑↑↓	↓↑↓↓
↑↑↑↓	↑↑↑↑	↓↓↓↓	↑↓↓↓

(i) Does this system have more microstates or more macrostates?

A) microstates
B) macrostates
C) there are an equal number of microstates and macrostates
D) impossible to determine

(ii) What is the probability of finding the system in the most probable macrostate?

A) 1/16
B) 1/5
C) 1/4
D) 3/8
E) 2/5
F) 1/2

Item 5

Consider two interacting systems A and B. If System A has 4 microstates and System B has 2 microstates, what is the total number of microstates of the combined system (which includes both A and B)?

A) 1
B) 6
C) 7
D) 8
E) 15
F) 20

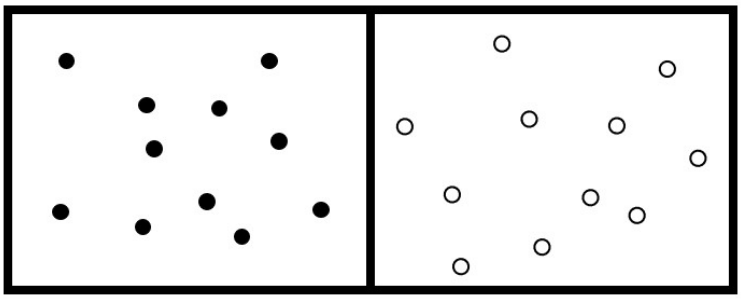
Item 6

Consider two systems in different types of contact moving towards equilibrium. For each type of contact, determine what quantity is exchanged between the two systems and which quantity becomes equal in equilibrium. Fill out the table using the word bank below. Ensure your responses are spelled correctly.

Type of contact:	**Thermal contact**	**Mechanical contact**	**Diffusive contact**
Exchanged Quantity			
Equalized quantity after a very long time:			

Item 7

Consider a system composed of a box of volume V which is divided in half by a removable partition. One half of the container holds an ideal gas composed of N Helium atoms and the other holds a gas composed of N Argon atoms. The gases have been sitting for a very long time.



(i) When the partition is removed, the entropy of the system...
- A) increases
- B) decreases
- C) remains the same

because... *(select all that support your response above)*
- a) two species of particles are mixing
- b) the volume of each gas species increases
- c) the volume of the system remains the same
- d) the number of accessible states increases
- e) the number of accessible states remains the same
- f) the number of particles increases
- g) the number of particles remains the same
- h) it is required to do so by the second law of thermodynamics

(ii) When the partition is removed in the situation above, the magnitude of the change in entropy takes the form:

$$|\Delta S| = \alpha N k_B$$

Given the Sackur-Tetrode equation

$$S = N k_B \left\{ ln \left[\frac{V}{N} \left(\frac{4\pi m}{3h^2} \frac{U}{N} \right)^{3/2} \right] + \frac{5}{2} \right\}, \tag{D.1}$$

what is the value of α?
- A) 0 (i.e., the entropy does not change)
- B) 2
- C) 2ln2
- D) ln2
- E) not enough information

Item 8

A heat engine is a thermodynamic cycle designed to transform heat into useable work. Consider one complete cycle of a heat engine operating between two thermal reservoirs. This heat engine operates using a working substance that expands and compresses during each cycle.

(i) A Carnot engine is a heat engine for which each step of the cycle is a perfectly reversible process. As a result of one complete cycle of a Carnot engine, the entropy of the universe will...

A) increase
B) decrease
C) remain the same
D) not determinable with the given information

because... *(select all that support your response above)*

a) the second law of thermodynamics does not allow entropy to decrease
b) reversible processes increase entropy
c) reversible processes decrease entropy
d) reversible processes do not change entropy
e) Carnot cycles are not realistic

(ii) Consider an engine that is *more* efficient than a Carnot engine. As a result of one complete cycle of this new engine, the entropy of the universe will...

A) increase
B) decrease
C) remain the same
D) not determinable with the given information

because... *(select all that support your response above)*

a) the second law of thermodynamics does not allow entropy to decrease
b) the efficiency does not depend on change in entropy
c) efficiency is dependent only on heat flows and reservoir temperatures
d) this more efficient engine would have a higher entropy change than the Carnot engine
e) this more efficient engine would have a lower entropy change than the Carnot engine
f) Carnot efficiency assumes positive entropy change in the universe
g) Carnot efficiency assumes zero entropy change in the universe
h) Carnot efficiency assumes negative entropy change in the universe
i) Carnot engines are the least efficient engine
j) Carnot engines are the most efficient engine

{*continued on next page*}

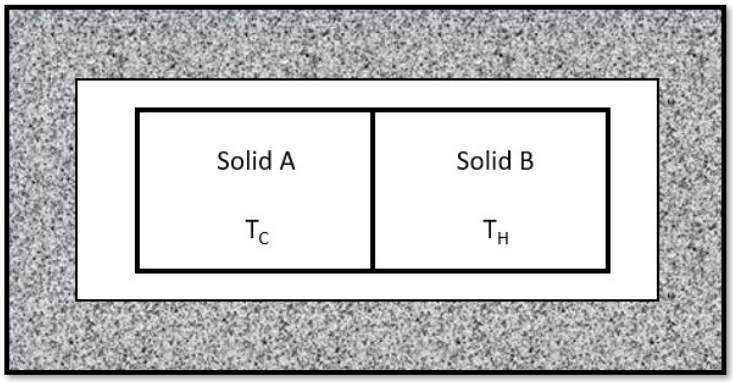
Item 8 *(continued)*

What do the laws of thermodynamics imply about this more efficient engine? *(select all that apply)*

a) One of the thermal reservoirs is at a negative temperature
b) The difference between the thermal reservoir temperatures is very large
c) The difference between the thermal reservoir temperatures is very small
d) The engine is physically possible but would be difficult to achieve in the real world
e) The engine is not physically possible because it is prohibited by the second law
f) The engine is not physically possible because it is prohibited by conservation of energy

Item 9

You perform an experiment involving two solids (A and B) which are initially at temperatures T_C and T_H, respectively (with $T_C < T_H$). You place them in thermal contact with each other inside of an insulated case to isolate them from everything else. Neglect thermal expansion, stresses, etc.



After the objects have come to thermal equilibrium, what can be said about the change in entropy for Solid A (ΔS_A), Solid B (ΔS_B), and the combined system of the two solids?

(i) The entropy for Solid A...
- A) increases
- B) decreases
- C) remains the same

because... *(select all that support your response above)*
- a) heat flows from Solid A to Solid B
- b) heat flows from Solid B to Solid A
- c) change in entropy is proportional to heat flow
- d) change in entropy is inversely proportional to change in temperature
- e) the temperature of Solid A increased
- f) the temperature of Solid A decreased
- g) the second law of thermodynamics does not allow entropy to decrease

{*continued on next page*}

Item 9 *(continued)*

(ii) The entropy for Solid B...
- A) increases
- B) decreases
- C) remains the same

because... *(select all that support your response above)*
- a) heat flows from Solid A to Solid B
- b) heat flows from Solid B to Solid A
- c) change in entropy is proportional to heat flow
- d) change in entropy is inversely proportional to change in temperature
- e) the temperature of Solid B increased
- f) the temperature of Solid B decreased
- g) the second law of thermodynamics does not allow entropy to decrease

(iii) The entropy of the combined system of the two solids...
- A) increases
- B) decreases
- C) remains the same

because... *(select all that support your response above)*
- a) $|\Delta S_A| < |\Delta S_B|$
- b) $|\Delta S_A| > |\Delta S_B|$
- c) $|\Delta S_A| = |\Delta S_B|$
- d) $\Delta S_A > 0$ and $\Delta S_B < 0$
- e) $\Delta S_A < 0$ and $\Delta S_B > 0$
- f) ΔS_A and ΔS_B have the same sign
- g) change in entropy is proportional to temperature
- h) change in entropy is inversely proportional to temperature
- i) the second law of thermodynamics does not allow entropy to decrease
- j) the system is isolated

Item 10

The heat capacity at constant volume C_V of an ideal gas in a sealed container is found to be

$$C_V{=}3Nk_B \ ,$$

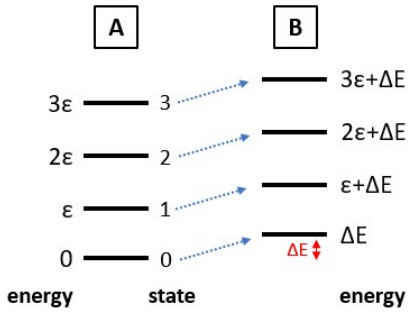
where N is the number of particles and k_B is Boltzmann's constant.

How many degrees of freedom do the particles composing the gas have?

A) 1
B) 3
C) 5
D) 6
E) Not enough information

Item 11

A particle is in configuration A, with the energy of each available state indicated on the left. The system undergoes a process such that the energy of the states is shifted by $\Delta E > 0$ (referred to as configuration B). (See figure below.)



(i) How does the partition function change due to the energy shift?

The partition function...
A) increases
B) decreases
C) remains the same

because... *(select all that support your response above)*
a) there are the same number of states before and after the energy shift
b) the partition function does not depend on the value of the energy level
c) each term in the partition function gains an $e^{(\Delta E/k_B T)}$ factor
d) each term in the partition function gains an $e^{(-\Delta E/k_B T)}$ factor
e) the states in configuration B would have more energy than the states in configuration A
f) a particle in configuration B would have more energy than a particle in configuration A

{*continued on next page*}

Item 11 *(continued)*

(ii) How does the probability of the particle being in state 2, $P(2)$, change due to the energy shift?

$P(2)$ ...
A) increases
B) decreases
C) remains the same

because... *(select all that support your response above)*
a) the particle is more excited
b) the Boltzmann factor increased due to the energy shift
c) the Boltzmann factor decreased due to the energy shift
d) the Boltzmann factor doesn't depend on energy
e) the partition function increased due to the energy shift
f) the partition function decreased due to the energy shift
g) the partition function doesn't depend on energy
h) the Boltzmann factor and partition function changed by the same factor
i) the Boltzmann factor changed by a greater factor than the partition function
j) the partition function changed by a greater factor than the Boltzmann factor
k) there are the same number of possible states before and after the energy shift

(iii) How does the ratio of probabilities of being in states 1 and 2 compare between A and B? That is, how does $P_A(1)/P_A(2)$ compare to $P_B(1)/P_B(2)$?

The ratio $P_A(1)/P_A(2)$ is...
A) less than
B) greater than
C) equal to

$P_B(1)/P_B(2)$ because... (select all that support your response above)
a) probability increases with increasing energy
b) probability decreases with increasing energy
c) state 2 is more probable in configuration A than configuration B
d) state 2 is more probable in configuration B than configuration A
e) each probability changes by the same factor going from configuration A to B
f) the differences between energy levels are the same in both systems
g) the ground state energy changes the same amount as other states' energies

Item 12

Consider a system with one accessible state.

What can be said about the system's entropy?

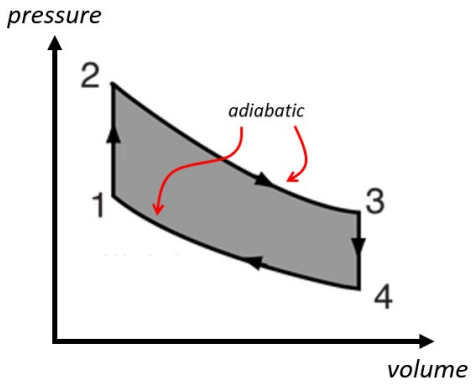
A) The entropy is 1
B) The entropy is much less than 1
C) The entropy is zero
D) The entropy is a non-zero constant value
E) The entropy changes with time
F) No statements can be made about the system's entropy

What can be said about the system's partition function?

A) The partition function is equal to 1
B) The partition function is greater than 1
C) The partition function is between 0 and 1
D) The partition function is composed of one Boltzmann factor
E) The partition function is determined using Boltzmann's equation
F) No statements can be made about the system's partition function

Item 13

Consider the system of a heat engine (not necessarily using an ideal gas) taken through one cycle ($1 \to 2 \to 3 \to 4 \to 1$) shown in the pressure vs. volume diagram below. Legs $2 \to 3$ and $4 \to 1$ are adiabatic processes. Assume the working substance does not undergo a phase change.



(i) On which leg(s) of the cycle does energy **enter** the system as heat?

Heat enters the system on leg(s)...
A) $1 \to 2$
B) $2 \to 3$
C) $3 \to 4$
D) $4 \to 1$

because on this leg (or these legs)... *(select all that support your response above)*
a) an adiabatic process occurred, so heat must have entered
b) pressure increases
c) pressure decreases
d) internal energy increases
e) internal energy decreases
f) volume is held constant
g) temperature is directly proportional to pressure
h) heat must enter when temperature increases
i) heat must enter when temperature decreases

{*continued on next page*}

Item 13 *(continued)*

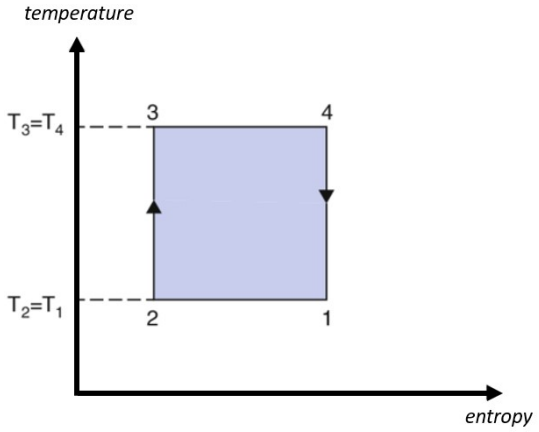
(ii) On which leg(s) of the cycle does energy **leave** the system as heat?

Heat leaves the system on leg(s)...
A) $1 \rightarrow 2$
B) $2 \rightarrow 3$
C) $3 \rightarrow 4$
D) $4 \rightarrow 1$

because on this leg (or these legs)... *(select all that support your response above)*
a) an adiabatic process occurred, so heat must have left
b) pressure increases
c) pressure decreases
d) internal energy increases
e) internal energy decreases
f) volume is held constant
g) temperature is directly proportional to pressure
h) heat must leave when temperature increases
i) heat must leave when temperature decreases

Item 14

Consider the temperature (T) vs. entropy (S) diagram below for a single cycle ($1 \to 2 \to 3 \to 4 \to 1$) of an ideal gas, as shown below.



(i) Describe what happens on the following legs of the cycle with respect to energy going into or out of the system as heat and work.

On leg $1 \to 2$...

A) heat flows in, work causes energy to flow in
B) heat flows in, work causes energy to flow out
C) heat flows in, work does not cause energy flow
D) heat flows out, work causes energy to flow in
E) heat flows out, work causes energy to flow out
F) heat flows out, work does not cause energy flow
G) no heat flows, work causes energy to flow in
H) no heat flows, work causes energy to flow out
I) no heat flows, work does not cause energy flow

On leg $3 \to 4$...

A) heat flows in, work causes energy to flow in
B) heat flows in, work causes energy to flow out
C) heat flows in, work does not cause energy flow
D) heat flows out, work causes energy to flow in
E) heat flows out, work causes energy to flow out
F) heat flows out, work does not cause energy flow
G) no heat flows, work causes energy to flow in
H) no heat flows, work causes energy to flow out
I) no heat flows, work does not cause energy flow

{*continued on next page*}

Item 14 *(continued)*

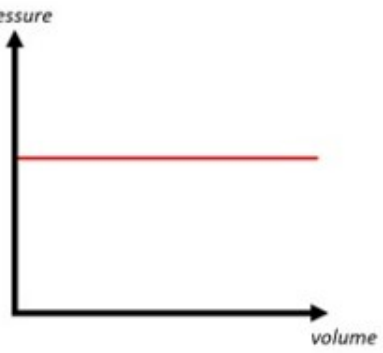
(ii) The shaded area within the cycle represents...
- A) change in internal energy (ΔU)
- B) work (W)
- C) heat (Q)

because... *(select all that support your response above)*
- a) $\Delta U = \int TdS$
- b) $Q = \int TdS$
- c) the magnitude of work is equal to the area under a curve
- d) the first law of thermodynamics relates W and Q to ΔU
- e) Q=0 over the full cycle
- f) W=0 over the full cycle

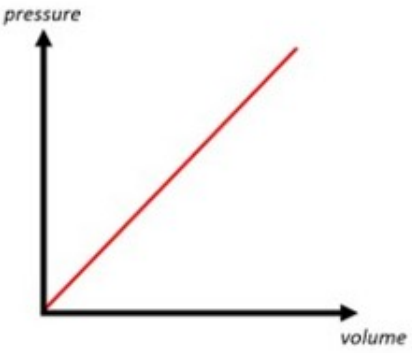
Item 15

Which of the following most closely resembles an isothermal process for an ideal gas?

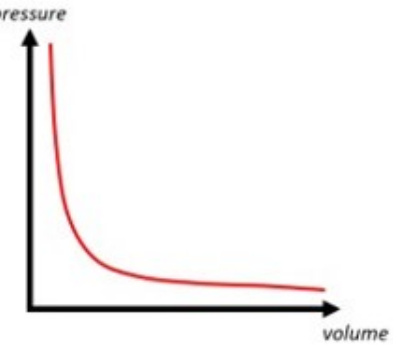
A)



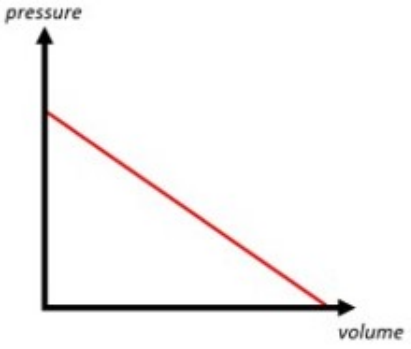
B)



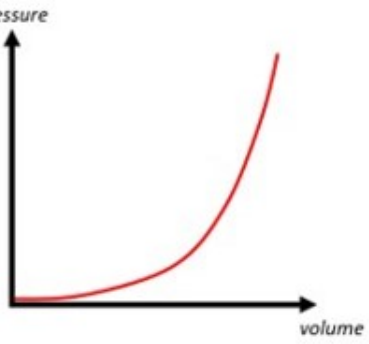
C)



D)



E)



— End of the U-STEP —

Items Not Included on the Final U-STEP

The following 6 items were included on one or both versions of the assessment piloted in Spring 2020 and will not be included in the final U-STEP. They are presented here because they are referenced in Table 4.2.

Item 16

Consider the following statement:

Any object at a given temperature contains a certain amount of heat.

This statement is...

A) true
B) false

because... *(select all that support your response above)*

a) heat is a quantity exchanged between systems
b) it is impossible to know exactly how much heat an object contains
c) heat is a flow of thermal energy
d) heat and temperature are the same thing
e) heat is not a state function
f) the amount of heat in an object determines its temperature
g) it does not make sense to talk about heat as a quantity that can be contained
h) other: _____

Item 17

During a class discussion about a system composed of weakly coupled simple harmonic oscillators, one student says:

> *Even if the system is isolated and in thermal equilibrium, some microstates within the system are more probable than others because each oscillator has access to multiple energy states (e.g. can hold different amounts of energy).*

Do you agree or disagree?

A) agree
B) disagree

Why? *(select one)*

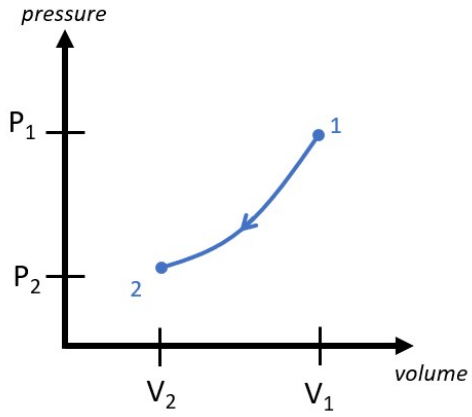
A) the Boltzmann factors are different for different oscillators
B) the Boltzmann factors are the same for each oscillator
C) all microstates are equally probable
D) all microstates are not equally probable
E) the oscillators hold different amounts of energy
F) each oscillator must hold the same amount of energy

I know this because... *(select all that support your response above)*

a) the system is isolated
b) the system is in thermal equilibrium
c) probability only depends on temperature
d) the oscillators are modeled as identical
e) otherwise this would violate equipartition
f) some states are not accessible
g) other: _____

Item 18

A monatomic ideal gas compresses according to the pressure (P) versus volume (V) diagram below.



To achieve this change, does heat flow into or flow out of the system, or is no heat required?

Heat flow is...

A) into the system
B) out of the system
C) zero
D) impossible to determine

because... *(select all that support your response above)*

a) temperature increased
b) temperature decreased
c) temperature didn't change
d) $\Delta U<0$
e) $\Delta U>0$
f) $\Delta U = 0$
g) energy flows into the system as work
h) energy flows out of the system as work
i) according to the first law, heat and work determine internal energy change
j) a PV diagram only provides information about work, not heat
k) other: _____

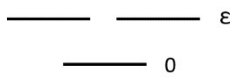
Item 19

Which of the following **must** be true in order for the equipartition theorem to be used for a given system? *(select all that apply)*

a) The equipartition theorem always holds, regardless of system conditions
b) The system is isolated
c) The system is in the "thermodynamic limit" (large N)
d) The system is in thermal equilibrium
e) $k_B T$ for the system is much, much larger than atomic energy spacing
f) The system undergoes a slow process
g) The system is a gas
h) The system is a solid
i) All system microstates are equally likely
j) All system macrostates are equally likely
k) There are some quadratic degrees of freedom in the energy function
l) There are only quadratic degrees of freedom in the energy function
m) There are no inter-particle interactions
n) Other: _____

Item 20

Consider a small system in equilibrium with a heat bath at temperature T. The system is isolated such that no particles can enter or leave the system. The ground state has zero energy and both excited states have energy ϵ.



What is the probability of finding a particle in the ground state when $T = \epsilon/k_B$, where k_B is the Boltzmann constant?

A) 1/3
B) 2/3
C) $\frac{1}{1+e/2}$
D) $\frac{1}{1+2/e}$
E) $\frac{1}{1+e}$
F) $\frac{1}{1+2e}$

Item 21

Consider the following statement:

Heat is a property of a system, like temperature, pressure, and volume.

This statement is...

A) true
B) false

because... *(select all that support your response above)*

a) heat is a property that is determined by a system's temperature, pressure, and volume
b) heat is a property that can be changed by changing temperature, pressure, and volume
c) a system contains a finite amount of heat, like pressure or volume
d) heat is a quantity exchanged between systems
e) heat is a flow of thermal energy
f) heat is a scalar, like temperature, pressure, and volume
g) heat is not a state function
h) other: _____

Appendix E

Item Scoring for the U-STEP

The following tables display the scoring schemes for the items presented in Appendix D. Multiple-choice (MC) options are in *capital* letters; multiple-response (MR) options are in *lowercase* letters. Correct MC selections are indicated (*). Matrices of scores correspond to coupled, multiple response MC-MR pairs and are labeled when multiple pairs are present. (See Appendix D for reference.) An illustration of how to apply these scoring schemes can be found in Figure 4.1. All items in Table 4.2 are presented; note items 16-21 will not be included in the final U-STEP.

Item 1: Work & PV Diagram

	A	B*	C	D
a	0	0	0	0
b	-3	-3	0	0
c	0	0	0	0
d	-3	-3	0	0
e	-3	0	0	0
f	-3	-3	0	0
g	0	0	0	0
h	0.5	1.5	0	0
i	-3	-3	0	0
	A	B*	C	D
a, c	0.5	1.5	0	0

Item 2: 1st Law & PV Diagram

(i)	A*	B	C	D
a	2	0	0	0
b	-1.5	0	0	0
c	2	0	0	0
d	-3	0	0	0
e	0	0	0	0
f	-3	0	0	0
g	0	0	0	0
h	-3	0	0	0.5

(ii)

A	0
B	0
C*	2
D	0
E	0
F	0

(iii)

A*	2
B	0
C	0
D	0
E	0

Item 3: Heat (see Fig.3.7)

	A	B*
a	0	-5
b	0	-5
c	0	-5
d	0	1
e	0	1
f	0	0
g	0	3

Item 4: Statistical Mechanics Micro/Macrostates & Probability

(i)

A*	2
B	0
C	0
D	0

(ii)

A	0
B	0
C	0
D*	2
E	0
F	0

Item 5: Statistical Mechanics Multiplicity

A	0
B	0
C	0
D*	2
E	0
F	0

Item 6: Equilibrium—Thermal, Mechanical, & Diffusive

heat	2
volume	2
particles	2
temperature	2
pressure	2
chemical potential	2

Item 7: Entropy & Mixing Gases

(i)	A*	B	C
a	2	0	-2
b	1	0	-2
c	0	0	0.5
d	1	0	-2
e	-3	0	-2
f	-3	0	-2
g	0	0	0.5
h	0	0	0

	A*	B	C
c, g	0	0	-0.5

(iii)	
A	0
B	0
C*	2
D	0
E	0

Item 8: Entropy & Carnot Engines

(i)	A	B	C*	D
a	0.5	0	1	0
b	0	0	-3	0
c	0	0	-3	0
d	0	0	3	0
e	0	0	0	0

(ii-1)	A	B*	C	D
a	0.5	-3	0	0
b	0	-3	-2	-2
c	0	-3	0.5	0.5
d	0	-3	-2	-2
e	0	1.5	0	0
f	0	-3	-2	-2
g	0	1.5	0	0
h	0	-3	-2	-2
i	0	-3	-2	-2
j	0	0.5	0	0

(ii-2)	A	B*	C	D
a	0	0	0	0
b	0	-3	0	0
c	0	-3	0	0
d	0	-3	0	0
e	0	3	0	0
f	0	-2	0	0

Item 9: Entropy & Heat Flow Between Solids

Note for (iii): The top row indicates the MC response patterns across (i), (ii), and (iii).

(i)	A*	B	C
a	-3	0	0
b	1.5	-3	0
c	1.5	0	0
d	-1.5	-3	0
e	0	0	0
f	-3	-3	0
g	-1	-3	0
	1	2	3
a, c	0	0.5	0

(ii)	A	B*	C
a	0	-3	0
b	0.5	1.5	0
c	0	1.5	0
d	-3	-1.5	0
e	-3	-3	0
f	0	0	0
g	-1	-3	0
	1	2	3
a, b	-3	0	0
a, c	0.5	0	0

(iii)	A, A, A	A, B, A*	A, B, C	B, A, A	B, A, C
a	0	-3	-3	0	-3
b	0	1	-3	-3	-3
c	0	-3	0	-3	0
d	-3	1	0	-3	-3
e	-3	-3	-3	0	0
f	0.5	-3	-3	-3	-3
g	-3	-3	-3	0	-3
h	0	1	-1	-1	-1
i	0	0	0	0	0
j	0	0	0	0	0
	A, A, A	A, B, A*	A, B, C	B, A, A	B, A, C
i, j	0.5	0.5	0.5	0.5	0.5
f, i, j	-0.5				
c, d			0.5		
c, d, i, j			-0.5		
a, e				0.5	
a, e, i, j				-0.5	
c, e					0.5
c, e, i, j					-0.5

Item 10: Energy & Degrees of Freedom

A	0
B	0
C	0
D*	2
E	0

Item 11: Statistical Mechanics (see Fig. 3.9)

(i)

	A	B*	C
a	-3	-1	0.5
b	-3	-3	-3
c	0.5	-3	-3
d	-3	3	-3
e	0.5	1	-3
f	0	0	-3

(ii)

	A	B	C*
a	0	-3	-3
b	-3	-3	-3
c	-3	0.5	1
d	-3	-3	-3
e	-3	-3	-3
f	0.5	-3	1
g	-3	-3	-3
h	-3	-3	1
i	0	-3	-3
j	-3	0	-3
k	-3	-3	0

(iii)

	A	B	C*
a	-3	-3	-3
b	-3	-3	-3
c	0.5	-3	-3
d	-3	0.5	-3
e	-3	-3	0.5
f	0	0	3
g	0	0	0.5

Item 12: Statistical Mechanics, Entropy, & the Parittion Function)

(i)

A	0
B	0
C*	2
D	0
E	0
F	0

(ii)

A	0
B	0
C	0
D*	2
E	0
F	0

Item 13: Engines, Heat, & Work

(i)	A*	B	C	D
a	-3	0	-3	0
b	0.5	0	-3	0
c	-3	0	0	0
d	0	0	-3	0
e	-3	0	-3	0
f	0.5	0	0	0
g	-1.5	0	-1.5	0
h	-3	0	-3	0
i	-3	0	-3	0
j	0	0	0	0

	A*	B	C	D
b, f	2	0	0	0
c, f	0	0	0.5	0

(ii)	A	B	C*	D
a	-3	0	-3	0
b	0	0	-3	0
c	-3	0	0.5	0
d	-3	0	-3	0
e	-3	0	0	0
f	0	0	0.5	0
g	-1.5	0	-1.5	0
h	-3	0	-3	0
i	-3	0	-3	0
j	0	0	0	0

	A	B	C*	D
b, f	0.5	0	0	0
c, f	0	0	2	0

Item 14: Engines & Entropy-Temperature Diagram

(i)	
A	0.5
B	0
C	0
D*	2
E	0.5
F	0.5
G	0.5
H	0
I	0

(ii)	
A	0.5
B*	2
C	0.5
D	0
E	0.5
F	0
G	0
H	0.5
I	0

(iii)	A	B	C*	D
a	0	0	-2	0
b	-1	0	3	0
c	-1	0	-3	0
d	0	0	0	0
e	-1	0	-3	0
f	0	0	-3	0

	A	B	C*	D
a, e	0.5	0	0	0

Item 15: Isotherm of an Ideal Gas

A	0
B*	2
C	0
D	0
E	0

The following scoring schemes are for items not included in the finalized U-STEP but were piloted in Spring 2020.

Item 16: Temperature (see Fig. 3.6)

	A	B*
a	0	1
b	0	-3
c	0	1
d	0.5	-3
e	0	3
f	0.5	-3
g	0	2
h	0	0

Item 17: The Fundamental Assumption of Statistical Mechanics

	A	B*
1	0.5	0
2	0	0
3	0	2
4	0	0
5	0	0
6	0	0

(i)	(A) Agree						(B*) Disagree					
(ii)	A, a	A, b	A, c	A, d	A, e	A, f	B, a	B, b	B, c	B, d	B, e	B, f
a	0	0	0	0	0	0	0	0	1.5	0	0	0
b	0	0	0	0	0	0	0	0	1.5	0	0	0
c	0	0	0	0	0	0	0	0	-1	0	0	0
d	0	0	0	0	0	0	0	0.5	0.5	0	0	0.5
e	0	0	0	0	0	0	0	0	-5	0	0	0
f	0	0	0	0	0	0	0.5	0	-5	0	0	0
g	0	0	0	0	0	0	0	0	0	0	0	0
	A, a	A, b	A, c	A, d	A, e	A, f	B, a	B, b	B, c	B, d	B, e	B, f
b, e	0	0	0	0	0	0	0	0	0	0	0	0.5
b, d, e	0	0	0	0	0	0	0	0	0	0	0	-0.5

Item 18: Heat & PV-Diagram

	A	B*	C	D
a	0	-3	0	-3
b	0	0.5	0	0
c	0	-3	0	-3
d	0	0.5	0	0
e	0	-3	0	-3
f	0	-3	0	-3
g	0	1	0	-3
h	0	-1	0	0
i	0	1	0	0
j	0	-2	0	0
k	0	0	0	0

	A	B*	C	D
b, d	0	0	0	0.5
b, d, h, i	0	0	0	0.5

Item 19: Energy & Equipartition

a	b	c	d	e	f	g	h	i	j	k	l	m	n
-3	1	0	1	1	-2	-1	-1	0	-3	-1	1	0	0

Item 20: Statistical Mechanics & Probability of Degenerate States

A	0
B	0
C	0
D*	2
E	0
F	0

Item 21: Heat (see Fig. 3.6)

	A	B*
a	0	-5
b	0	-5
c	0	-5
d	0	1
e	0	1
f	0	-5
g	0	3
h	0	0

Appendix F

Differential Item Functioning Results

The following tables present results from differential item functioning (DIF) analyses for gender and race from the Fall 2020 pilot administration. Performance differences between men and women (Table F.1); Asian and underrepresented minority (URM) students (Table F.2); Asian and White students (Table F.3); and White and URM students (Table F.4) are presented for the upper- and lower-25th percentiles. Statistical significance and effect size are also presented.

Table F.1: DIF results for the Fall 2020 pilot based on gender, comparing **women** and **men**. Differences in item averages are presented (Δ_{avg} = average(women) - average(men)), along with significance of the difference determined by a Mann-Whitney test (p) and effect size for both the top and bottom 25th percentiles. Positive Δ_{avg} indicate women performed better on the item; negative values indicate men performed better. Statistically significant values are idicated in **bold***. The upper 25th percentile was composed of 5 women and 34 men. The lower 25th percentile was composed of 12 women and 25 men.

	Upper 25th Percentile			Lower 25th Percentile		
Item	Δ_{avg}	p	effect size	Δ_{avg}	p	effect size
1	-0.12	0.37	-0.42	-0.06	0.19	-0.31
2	0.10	0.46	0.44	-0.14	**0.04***	-0.76
3	0.00	0.69	-0.01	-0.03	0.79	-0.11
4	-0.08	0.52	-0.32	-0.21	0.09	-0.61
5	0.06	0.62	0.26	0.10	0.57	0.20
6	0.02	0.87	0.10	-0.03	0.80	-0.09
7	-0.08	0.41	-0.44	-0.08	0.42	-0.27
8	-0.24	0.06	-0.93	0.09	0.33	0.37
9	-0.17	0.35	-0.88	-0.02	0.78	-0.09
10	-0.08	0.64	-0.24	0.01	0.97	0.02
11	0.02	0.75	0.10	-0.03	0.37	-0.14
12	-0.11	0.63	-0.34	0.09	0.33	0.28
13	0.30	**0.01***	1.38	-0.19	**<0.01***	-0.95
14	0.26	0.16	0.70	0.03	0.97	0.14
15	0.06	0.062	0.26	-0.19	0.28	-0.39

Table F.2: DIF results for the Fall 2020 pilot based on race, comparing **Asian** and **underrepresented minority (URM)** students. Differences in item averages are presented (Δ_{avg} = average(Asian) - average(URM)), along with significance of the difference determined by a Mann-Whitney test (p) and effect size for both the top and bottom 25th percentiles. Positive Δ_{avg} indicate Asian students performed better on the item; negative values indicate URM students performed better. Statistically significant values are idicated in **bold***. The upper 25th percentile was composed of 8 Asian students and 4 URM students. The lower 25th percentile was composed of 7 Asian students and 9 URM students.

	Upper 25th Percentile			Lower 25th Percentile		
Item	Δ_{avg}	p	effect size	Δ_{avg}	p	effect size
1	0.13	0.60	0.36	0.09	0.44	0.46
2	-0.03	0.73	-0.18	0.07	0.52	0.41
3	0.03	0.28	0.08	-0.03	0.86	-0.12
4	0.19	0.21	0.87	-0.07	0.78	-0.18
5	0.25	0.22	0.91	0.02	1.00	0.03
6	0.02	0.79	0.10	0.06	0.35	0.22
7	-0.16	0.11	-1.18	-0.18	0.14	-0.74
8	-0.08	0.61	-0.36	-0.10	0.27	-0.50
9	0.05	0.60	0.35	0.06	0.59	0.27
10	0.13	0.69	0.31	-0.19	0.44	-0.42
11	0.16	0.31	0.56	0.06	0.87	0.38
12	0.00	0.92	0.00	-0.13	0.50	-0.42
13	0.19	0.29	0.76	-0.07	0.70	-0.28
14	-0.08	0.85	-0.24	0.12	**0.04***	1.13
15	0.00	N/A	N/A	0.21	0.43	0.43

Table F.3: DIF results for the Fall 2020 pilot based on race, comparing **Asian** and **White** students. Differences in item averages are presented (Δ_{avg} = average(Asian) - average(White)), along with significance of the difference determined by a Mann-Whitney test (p) and effect size for both the top and bottom 25th percentiles. Positive Δ_{avg} indicate Asian students performed better on the item; negative values indicate White students performed better. No statistically significant differences appeared in this analysis. The upper 25th percentile was composed of 8 Asian students and 29 White students. The lower 25th percentile was composed of 7 Asian students and 16 White students.

	Upper 25th Percentile			Lower 25th Percentile		
Item	Δ_{avg}	p	effect size	Δ_{avg}	p	effect size
1	0.07	0.59	0.21	0.05	0.48	0.22
2	0.07	0.67	0.32	0.00	1.00	-0.01
3	0.05	0.31	0.16	0.02	0.77	0.08
4	0.00	1.00	-0.01	0.08	0.60	0.27
5	0.03	0.65	0.21	0.01	1.00	0.02
6	-0.05	0.67	-0.25	-0.01	0.89	-0.03
7	0.01	0.95	0.04	-0.09	0.44	-0.36
8	-0.04	0.66	-0.14	-0.16	0.10	-0.60
9	0.05	0.67	0.25	0.07	0.48	0.27
10	-0.02	0.89	-0.07	-0.04	0.84	-0.11
11	-0.04	0.66	-0.18	0.07	0.43	0.32
12	-0.06	0.70	-0.19	-0.04	1.00	-0.14
13	0.04	0.68	0.19	-0.05	0.53	-0.25
14	0.21	0.16	0.57	0.06	0.06	0.29
15	0.10	0.37	0.37	0.05	0.84	0.11

Table F.4: DIF results for the Fall 2020 pilot based on race, comparing **White** and **under-represented minority (URM)** students. Differences in item averages are presented (Δ_{avg} = average(White) - avgerage(URM)), along with significance of the difference determined by a Mann-Whitney test (p) and effect size for both the top and bottom 25th percentiles. Positive Δ_{avg} indicate White students performed better on the item; negative values indicate URM students performed better. No statistically significant differences appeared in this analysis. The upper 25th percentile was composed of 29 White students and 4 URM students. The lower 25th percentile was composed of 16 White students and 9 URM students.

	Upper 25th Percentile			Lower 25th Percentile		
Item	Δ_{avg}	p	effect size	Δ_{avg}	p	effect size
1	0.06	0.80	0.19	0.04	0.95	0.21
2	-0.10	0.50	-0.41	0.07	0.40	0.38
3	-0.03	0.68	-0.09	-0.05	0.61	-0.19
4	0.19	0.15	0.81	-0.16	0.37	-0.44
5	0.22	0.11	0.92	0.01	1.00	0.01
6	0.07	0.38	0.38	0.07	0.27	0.26
7	-0.17	0.12	-0.81	-0.09	0.44	-0.31
8	-0.04	1.00	-0.15	0.07	0.67	0.25
9	0.00	0.68	0.01	-0.01	0.98	-0.05
10	0.15	0.44	0.44	-0.15	0.44	-0.33
11	0.20	0.17	0.85	0.00	0.42	-0.02
12	0.06	0.57	0.20	-0.09	0.43	-0.25
13	0.15	0.30	0.64	-0.02	0.80	-0.08
14	-0.30	0.14	-0.82	0.06	0.95	0.34
15	-0.10	0.54	-0.35	0.15	0.46	0.32

Appendix G

Item Response Theory Statistics

The tables on the following pages present IRT analysis results, as discussed in Ch. 4. Table G.1 presents Rasch analysis results for the three thresholds considered—40%, 50%, and 60%—including difficulties and fit statistics for each item. The remaining tables present fits statistics for the three thresholds when misift items are removed.

Table G.1: Difficulty values (b) and fit statistics for IRT analysis using three thresholds for dichotomizing data. Lower S-X^2 values indicate better fit of data to the IRT model. Root mean square error of approximation (RMSEA) values are indicated as ^{E}excellent (RMSEA<0.01) where applicable. Statistically significant fit statistics (determined by p) at the 0.05 level are indicated in **bold***. Large S-X^2 values with statistically significant p indicated model misfit (i.e., the model does not fit the data well).

	40% Thresholds				50% Threshold				60% Threshold			
Item	b	S-X^2	RMSEA	p	b	S-X^2	RMSEA	p	b	S-X^2	RMSEA	p
1	0.51	8.47	0.019	0.389	1.84	14.72	0.082	**0.040***	1.83	17.73	0.086	**0.023***
2	-0.80	13.60	0.066	0.093	0.15	9.86	0.038	0.275	0.27	5.60	0.000^{E}	0.691
3	-0.70	7.91	0.000^{E}	0.443	0.09	13.02	0.062	0.111	0.09	8.06	0.007^{E}	0.428
4	-1.90	8.17	0.012	0.417	-1.89	7.52	0.000^{E}	0.482	1.69	10.16	0.041	0.254
5	-1.41	10.80	0.046	0.214	-1.40	7.79	0.000^{E}	0.454	-1.40	18.48	0.090	**0.018***
6	-1.45	14.21	0.069	0.076	-1.44	12.13	0.056	0.145	-0.33	9.16	0.030	0.329
7	-2.05	2.76	0.000^{E}	0.948	-1.25	15.17	0.085	**0.034***	-0.60	3.70	0.000^{E}	0.814
8	-0.09	4.43	0.000^{E}	0.729	0.15	8.15	0.011	0.419	0.39	6.34	0.000^{E}	0.609
9	-1.00	7.34	0.000^{E}	0.501	-0.18	8.59	0.000^{E}	0.476	0.18	10.09	0.040	0.259
10	-0.33	6.49	0.000^{E}	0.593	-0.33	12.81	0.061	0.119	-0.33	6.86	0.000^{E}	0.552
11	0.06	13.75	0.077	0.056	0.48	7.79	0.000^{E}	0.454	0.86	6.49	0.000^{E}	0.592
12	-1.11	9.11	0.029	0.333	-1.10	5.46	0.000^{E}	0.707	0.96	5.78	0.000^{E}	0.672
13	-1.18	12.04	0.056	0.149	0.36	16.20	0.079	**0.040***	1.32	13.58	0.056	0.138
14	0.58	18.19	0.088	**0.020***	0.97	15.26	0.097	**0.018***	1.32	12.71	0.050	0.176
15	-0.90	8.35	0.016	0.400	-0.90	7.48	0.000^{E}	0.486	-0.89	3.22	0.000^{E}	0.864

Table G.2: Rasch analysis fit statistics for the **40% threshold** assessment with item 14 removed. (Original fit statistics for full assessment shown for reference.) Root mean square error of approximation (RMSEA) values are indicated as $^{\text{E}}$excellent (RMSEA<0.01) where applicable. Statistically significant fit statistics (determined by p) at the 0.05 level are indicated in **bold***. Large S-X^2 values with statistically significant p indicated model misfit (i.e., the model does not fit the data well). Note: only items misfit in the full assessment are removed for this analysis.

	No Items Removed			Item 14 Removed		
Item	S-X^2	RMSEA	p	S-X^2	RMSEA	p
1	8.47	0.019	0.389	13.37	0.075	0.064
2	13.60	0.066	0.093	10.98	0.059	0.139
3	7.91	0.000^{E}	0.443	9.30	0.045	0.232
4	8.17	0.012	0.417	10.94	0.059	0.141
5	10.80	0.046	0.214	13.37	0.064	0.100
6	14.21	0.069	0.076	16.63	0.081	**0.034***
7	2.76	0.000^{E}	0.948	4.04	0.000^{E}	0.775
8	4.43	0.000^{E}	0.729	11.25	0.050	0.188
9	7.34	0.000^{E}	0.501	7.26	0.000^{E}	0.509
10	6.49	0.000^{E}	0.593	8.95	0.027	0.347
11	13.75	0.077	0.056	12.69	0.071	0.080
12	9.11	0.029	0.333	3.75	0.000^{E}	0.879
13	12.04	0.056	0.149	10.25	0.042	0.248
14	18.19	0.088	**0.020***	removed		
15	8.35	0.016	0.400	6.40	0.000^{E}	0.603

Table G.3: Rasch analysis fit statistics for the **50% threshold** assessment with items 1 and 7 individually removed. (Original fit statistics for full assessment shown for reference.) Root mean square error of approximation (RMSEA) values are indicated as $^{\text{E}}$excellent (RMSEA<0.01) where applicable. Statistically significant fit statistics (determined by p) at the 0.05 level are indicated in **bold***. Large S-X^2 values with statistically significant p indicated model misfit (i.e., the model does not fit the data well).

	No Items Removed			Item 1 Removed			Item 7 Removed		
Item	S-X^2	RMSEA	p	S-X^2	RMSEA	p	S-X^2	RMSEA	p
1	14.72	0.082	**0.040***	removed			12.25	0.080	0.057
2	9.86	0.038	0.275	7.77	0.000^{E}	0.456	8.43	0.018	0.393
3	13.02	0.062	0.111	13.16	0.063	0.106	4.53	0.000^{E}	0.806
4	7.52	0.000^{E}	0.482	4.39	0.000^{E}	0.734	8.50	0.036	0.291
5	7.79	0.000^{E}	0.454	5.50	0.000^{E}	0.704	7.59	0.000^{E}	0.474
6	12.13	0.056	0.145	9.46	0.033	0.305	8.25	0.014	0.409
7	15.17	0.085	**0.034***	15.60	0.076	**0.048***	removed		
8	8.15	0.011	0.419	7.62	0.000^{E}	0.471	9.29	0.031	0.318
9	8.59	0.000^{E}	0.476	7.70	0.000^{E}	0.463	8.85	0.025	0.355
10	12.81	0.061	0.119	10.17	0.041	0.254	7.43	0.019	0.386
11	7.79	0.000^{E}	0.454	5.82	0.000^{E}	0.667	5.45	0.000^{E}	0.606
12	5.46	0.000^{E}	0.707	4.52	0.000^{E}	0.718	6.62	0.000^{E}	0.578
13	16.20	0.079	**0.040***	20.98	0.100	**0.007***	18.16	0.088	**0.020***
14	15.26	0.097	**0.018***	10.17	0.041	0.253	14.90	0.083	**0.037***
15	7.48	0.000^{E}	0.486	6.11	0.000^{E}	0.527	6.16	0.000^{E}	0.629

Table G.4: Rasch analysis fit statistics for the **50% threshold** assessment with items 13 and 14 individually removed. (Original fit statistics for full assessment shown for reference.) Root mean square error of approximation (RMSEA) values are indicated as $^{\text{E}}$excellent (RMSEA<0.01) where applicable. Statistically significant fit statistics (determined by p) at the 0.05 level are indicated in **bold***. Large S-X^2 values with statistically significant p indicated model misfit (i.e., the model does not fit the data well).

	No Items Removed			Item 13 Removed			Item 14 Removed		
Item	S-X^2	RMSEA	p	S-X^2	RMSEA	p	S-X^2	RMSEA	p
1	14.72	0.082	**0.040***	13.41	0.087	**0.037***	11.94	0.078	0.063
2	9.86	0.038	0.275	9.79	0.049	0.201	5.25	0.000^{E}	0.630
3	13.02	0.062	0.111	8.04	0.006^{E}	0.429	17.17	0.094	**0.016***
4	7.52	0.000^{E}	0.482	9.62	0.048	0.211	14.73	0.072	0.065
5	7.79	0.000^{E}	0.454	7.13	0.000^{E}	0.523	4.40	0.000^{E}	0.820
6	12.13	0.056	0.145	8.82	0.025	0.358	11.56	0.052	0.172
7	15.17	0.085	**0.034***	13.85	0.067	0.086	9.05	0.042	0.249
8	8.15	0.011	0.419	8.81	0.040	0.267	4.04	0.000^{E}	0.775
9	8.59	0.000^{E}	0.476	9.43	0.033	0.307	8.85	0.025	0.355
10	12.81	0.061	0.119	16.80	0.082	**0.032***	8.33	0.016	0.402
11	7.79	0.000^{E}	0.454	7.61	0.023	0.368	5.05	0.000^{E}	0.653
12	5.46	0.000^{E}	0.707	8.44	0.018	0.391	8.03	0.030	0.330
13	16.20	0.079	**0.040***	removed			17.18	0.094	**0.016***
14	15.26	0.097	**0.018***	9.08	0.043	0.247	removed		
15	7.48	0.000^{E}	0.486	7.74	0.025	0.356	5.51	0.000^{E}	0.598

Table G.5: Rasch analysis fit statistics for the **50% threshold** assessment with items 1, 7, 13, and 14 removed simultaneously. (Original fit statistics for full assessment shown for reference.) Root mean square error of approximation (RMSEA) values are indicated as $^{\text{E}}$excellent (RMSEA<0.01) where applicable. Statistically significant fit statistics (determined by p) at the 0.05 level are indicated in **bold***. Large S-X^2 values with statistically significant p indicated model misfit (i.e., the model does not fit the data well).

	No Items Removed			Items 1, 7, 13, & 14 Removed		
Item	S-X^2	RMSEA	p	S-X^2	RMSEA	p
1	14.72	0.082	**0.040***	removed		
2	9.86	0.038	0.275	5.41	0.000^{E}	0.493
3	13.02	0.062	0.111	2.90	0.000^{E}	0.822
4	7.52	0.000^{E}	0.482	8.42	0.035	0.297
5	7.79	0.000^{E}	0.454	18.19	0.112	**0.006***
6	12.13	0.056	0.145	8.45	0.050	0.207
7	15.17	0.085	**0.034***	removed		
8	8.15	0.011	0.419	2.72	0.000^{E}	0.843
9	8.59	0.000^{E}	0.476	9.41	0.046	0.225
10	12.81	0.061	0.119	9.14	0.043	0.243
11	7.79	0.000^{E}	0.454	5.79	0.000^{E}	0.447
12	5.46	0.000^{E}	0.707	9.62	0.061	0.142
13	16.20	0.079	**0.040***	removed		
14	15.26	0.097	**0.018***	removed		
15	7.48	0.000^{E}	0.486	9.48	0.060	0.148

Table G.6: Rasch analysis fit statistics for the **60% threshold** assessment with items 1 and 5 removed. (Original fit statistics for full assessment shown for reference.) Root mean square error of approximation (RMSEA) values are indicated as ^{E}excellent (RMSEA<0.01) where applicable. Statistically significant fit statistics (determined by p) at the 0.05 level are indicated in **bold***. Large S-X^2 values with statistically significant p indicated model misfit (i.e., the model does not fit the data well). Note: only items misfit in the full assessment are removed for this analysis.

	No Items Removed			Item 1 Removed			Item 5 Removed			Items 1 & 5 Removed		
Item	S-X^2	RMSEA	p	S-X^2	RMSEA	p	S-X^2	RMSEA	p	S-X^2	RMSEA	p
1	17.73	0.086	**0.023***	removed			9.77	0.037	0.281	removed		
2	5.60	0.000^{E}	0.691	7.31	0.000^{E}	0.504	3.87	0.000^{E}	0.795	8.85	0.040	0.264
3	8.06	0.007^{E}	0.428	5.85	0.000^{E}	0.664	2.74	0.000^{E}	0.908	1.90	0.000^{E}	0.965
4	10.16	0.041	0.254	19.22	0.103	**0.008***	13.33	0.064	0.101	11.80	0.054	0.161
5	18.48	0.090	**0.018***	21.50	0.113	**0.003***	removed			removed		
6	9.16	0.030	0.329	8.61	0.038	0.282	10.04	0.052	0.186	11.92	0.078	0.064
7	3.70	0.000^{E}	0.814	5.77	0.000^{E}	0.567	4.08	0.000^{E}	0.771	7.07	0.008^{E}	0.421
8	6.34	0.000^{E}	0.609	9.70	0.036	0.287	1.85	0.000^{E}	0.968	4.32	0.000^{E}	0.743
9	10.09	0.040	0.259	11.14	0.049	0.194	10.66	0.057	0.154	13.51	0.076	0.061
10	6.86	0.000^{E}	0.552	8.23	0.033	0.313	6.33	0.000^{E}	0.502	6.57	0.024	0.363
11	6.49	0.000^{E}	0.592	7.69	0.025	0.361	6.66	0.000^{E}	0.574	9.20	0.044	0.238
12	5.78	0.000^{E}	0.672	3.11	0.000^{E}	0.875	5.90	0.000^{E}	0.658	2.95	0.000^{E}	0.937
13	13.58	0.056	0.138	22.81	0.107	**0.004***	10.94	0.036	0.280	20.85	0.090	**0.013***
14	12.71	0.050	0.176	9.22	0.031	0.324	11.70	0.043	0.231	8.35	0.000^{E}	0.499
15	3.22	0.000^{E}	0.864	1.80	0.000^{E}	0.970	5.53	0.000^{E}	0.596	3.57	0.000^{E}	0.828