Validation of a coupled, multiple response assessment for upper-division thermal physics

Katherine D. Rainey, Michael Vignal, and Bethany R. Wilcox
Department of Physics, University of Colorado, 390 UCB, Boulder, Colorado 80309, USA

(Received 1 February 2022; accepted 10 August 2022; published 12 September 2022)

Currently there are no assessment instruments available for upper-division thermal physics, though several introductory assessments are currently available. Notably missing from these introductory assessment are items targeting statistical mechanics. This leaves a gap in the content that can be assessed by upper-division thermal physics faculty. In this paper, we present a new, upper-division thermal physics assessment that explicitly addresses both classical thermodynamics and statistical mechanics: The Upper-level Statistical and Thermodynamics Evaluation for Physics (U-STEP). We piloted the U-STEP in multiple classes across various institutions during three semesters, and collected over 600 student responses in total. Here, we present multiple measures of validity and reliability for the U-STEP. We utilized classical test theory to determine item difficulties and discriminations, and differential item functioning to identify potential bias in assessment items that can manifest as performance gaps between different genders and races. The completed U-STEP will be the first upper-division thermal physics assessment available, and one of the first standardized physics assessments to explicitly address identification of possible item bias during the development process.

DOI: 10.1103/PhysRevPhysEducRes.18.020116

I. INTRODUCTION

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, and by researchers to guide curriculum development or instructional interventions. Despite the widespread presence of standardized assessments in most of the core physics content areas at both introductory and upper-division levels, limited assessments are available for thermal physics beyond the introductory level. Thermal physics, which includes both classical thermodynamics and statistical mechanics, is a core course required for attaining a physics bachelor's degree at most institutions. However, to date there are no *upper-division* thermal physics assessments. 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.

One challenge influencing the development of an upper-division thermal physics assessment is the varying nature of content foci across different thermal physics courses [1]. This presents difficulties when identifying which topics to focus on for developing a broadly usable assessment. However, by asking faculty about content priorities in their upper-division thermal physics courses in our prior work [1], we were able to narrow topics to include on an assessment. By doing this, the assessment becomes more likely to be useful for a wide range of courses by addressing content relevant for many instructors.

Additionally, capturing student *reasoning*, as opposed to solely knowledge of an answer to a multiple-choice question, is an important consideration for many upper-division courses and provides deeper insight into student understanding. Capturing student reasoning can better inform course transformations to improve student learning by providing more fine-grained insight into students' understanding. Free-response (FR) assessments can capture rich reasoning patterns; however, they are harder to score efficiently. One alternative to FR items that still solicits some student reasoning is coupled, multiple-response (CMR) items, which capture reasoning through multiple-response questions [2]. To date, we know of one CMR-based, upper-division, content-focused physics assessment—the CUE-CMR [3]. However, CMR-based assessments to address lab skills, such as the Physics Lab Inventory of Critical Thinking (PLIC) [4] and Modeling Assessment for Physics Laboratory Experiments (MAPLE) [5] are also available or under development. The Physics Inventory of Quantitative Literacy (PIQL) also utilizes a similar structure for items [6].

Another consideration when developing standardized assessments is the possible appearance of performance differences between different genders and races, which could be caused by bias within an item (which, for example, can be attributed to wording or context). Performance differences on assessments in physics are often used to inform pedagogical interventions [7], but if the differences in performance are attributable to bias instead of actual learning gaps, these efforts may be misinformed or ineffective. Analysis of gender- and race-based performance gaps to identify item bias are historically rarely done *during* the development process of physics assessments [8]. These types of analyses can aid in the identification of issues within assessment items that contribute to performance differences (i.e., bias). Identifying and addressing these issues can minimize performance gaps that could be attributed to something outside of meaningful differences in student performance. Thus, interventions based upon any remaining performance gaps would be better informed and based on true differences in performance on the assessment. A common way to identify possible contributions to performance gaps that could stem from the assessment items themselves is differential item functioning (DIF), which involves identifying assessment items for which otherwise similarly performing students perform differently.

In this paper, we present a new, upper-division thermal physics assessment—the Upper-level Statistical and Thermodynamics Evaluation for Physics (U-STEP)—which is composed of CMR items and multiple-choice items that have been examined to identify DIF. We present background on thermal physics, assessment, and approaches taken for validation (Sec. II); a description of the U-STEP (Sec. III); context and methods (Sec. IV); and results of our validation analyses (Sec. V–IX). We conclude with a discussion of future work (Sec. X), limitations, and conclusions (Sec. XI).

II. BACKGROUND

Thermal physics education research spans the space of many disciplines, including biology, chemistry, engineering, and physics, and is becoming increasingly present in the literature [9,10]. These studies have utilized various methods to investigate student conceptual understanding through isolated thermal physics problems, small quizzes, and larger-scale assessments. However, these studies primarily focus on introductory thermal physics topics. Investigations of upper-division thermal physics content (e.g., Refs. [11–13]) are comparatively less common than work at the introductory level.

Similarly, existing thermal physics assessments focus on introductory, classical thermodynamics topics such as heat, temperature, and thermodynamic laws [14–18]. To date, there are no statistical mechanics assessments on PhysPort, a centralized online repository of physics education research (PER)-based resources for physics faculty that is home to over 100 research-based assessments.[1] An assessment for upper-division thermal physics, such as the U-STEP, could facilitate comparisons between similar courses and assist in research on ways to improve student learning of thermal physics content 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 thermal physics. We also discuss the format of the U-STEP, as well as techniques used for validating the U-STEP and for addressing bias within assessment items.

A. Thermal physics content

Many of the existing studies into student understanding in thermal physics involve ideal gas systems, most of which revolve around student understanding and application of the ideal gas law in various contexts (e.g., Refs. [19,20]). Studies on student problem solving with nonideal gas systems, on the other hand, are essentially nonexistent in the literature. Studies investigating students' alternate conceptions[2] of heat and temperature are another very common theme throughout thermal physics literature and emerge at all levels, from K-12 through college-level courses [22–24]. One source of confusion in relation to heat and temperature is the colloquial use of the two terms, which are often used interchangeably [23,25], resulting in students often considering heat as a property of a system (e.g., state function) [24], like temperature, as opposed to a process-dependent quantity [25].

Much like heat, work is also commonly thought of by students as a state function as opposed to a process-dependent quantity [25,26], an issue likely compounded by the fact that many students are first introduced to work in the context of conservative forces. These issues cause challenges when reasoning about net work and net heat for cycles; the view of work and heat as state functions leads students to reason that each of these quantities are zero for cyclic processes [27]. 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 [28].

Investigations of student conceptions revolving around heat and cycles have also addressed entropy [29], a core concept for thermal physics that is often unfamiliar to students entering these classes [30,31]. Entropy can be viewed in terms of both classical thermodynamics and

[1]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."

[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 [21].

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 [32,33]. Entropy has historically been a challenging topic for thermal physics students, often being considered as a conserved quantity [31,34] or a measure of chaos or disorder[3] [35,36]. Misapplication of the second law of thermodynamics is also very common, including assertions that entropy must increase in all contexts [31] or confusion about whether to apply the second law to the *system* or the *universe* [30].

The studies described above have investigated student conceptual understanding of introductory thermodynamics content, mainly at the high school and introductory-college level [9]. Upper-division investigations are comparatively less common. One study bridging introductory and upper-division content was conducted by Meltzer [11]. In this study, Meltzer compared student use of diagrammatic representation (i.e., P-V diagrams), notation, and mathematical equations, as well as verbal explanations, and found that several alternate conceptions common at the introductory level continue with students to more advanced thermal physics courses, sometimes persisting after instruction.

Compared to thermodynamics, statistical mechanics studies are relatively rare. Some exceptions include investigations about student reasoning surrounding the Boltzmann factor [13] and Taylor expansions in statistical mechanics [12], 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 [32]. Others have considered instructional strategies, such as using statistical approaches to teach entropy [39]. Smith *et al.* created tutorials to investigate student reasoning using the Boltzmann factor to compare relative probabilities of states [13] and to address students' conceptions of entropy when approaching Carnot cycles [29]. A subset of questions from these tutorials was used to inform item development for the U-STEP.

The studies described above helped inform the development of items on the U-STEP both by suggesting important content areas to be included and by informing reasoning elements that reflect known student ideas.

[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.) [35], as well as concerns about students' use of the term without provision of their own definition of what is meant by *disorder* [36]. 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 [34]. Several scholars have suggested moving away from presentation of entropy as disorder [35,37,38].

B. Thermal physics assessment

Though there are six assessments categorized as "thermal or statistical" assessments on PhysPort, all of these assessments are categorized as being for "intro college" or "intermediate" levels (i.e., not upper division). Additionally, none of the four assessments that are readily available online[4] [14,16–18] cover statistical mechanics concepts, confirming the lack of, and need for, a broader, upper-division thermal physics assessment. Filling this gap would provide instructors with a more authentic tool to assess pedagogical changes to their undergraduate thermal physics courses as a whole.

We note that only five of the six thermal physics assessments on PhysPort were finalized and the sixth was recommended to not be used by the assessment developer. We could not access one of the five assessments because a request for access to the developer did not receive a response. 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) [40]. The TDT is not available on PhysPort, but its questions are available in Ref. [40]. Other existing thermal physics assessments address concepts such as specific thermodynamic laws and processes [18,40,41], as well as basic thermodynamic concepts such as heat and temperature [14–16], phase transformations [16], and thermal properties of materials [16,17].

Most standardized assessments in PER, including existing thermal physics assessments, have narrow scopes, honing in on a very specific subset of topics within a particular subdiscipline of physics. For example, the focus of the Force Concept Inventory (FCI) is forces, as opposed to "introductory mechanics." The goal of this approach is to increase the utility of the assessment; if content is covered on an assessment but not in a course, that assessment is unlikely to be used for that course. Thus, narrowing the scope of the assessment makes it more targeted and leads to increased likelihood of faculty use. This same approach of focusing on a narrower scope of canonical content, informed by a survey of thermal physics instructors [1], was used to develop the U-STEP as described in Sec. III.

C. Coupled, multiple response items

Most assessments utilized in PER are multiple-choice (MC) or free-response (FR) formats (see PhysPort). One MC thermal physics assessment, the TDT [40], 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

[4]The assessments accessible through PhysPort at the time of publication 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).

answer. The final version of the U-STEP 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 U-STEP is largely composed of coupled, multiple-response (CMR) items in addition to (relatively few) MC items. CMR items are composed of a MC question followed with a multiple-response question asking students to select one or more reasoning options (called reasoning elements) used to find the first answer [2]. Unlike traditional MC questions, CMR formats allow for scoring based on not only correctness but also consistency and completeness of reasoning. They also allow for partial credit as opposed to being solely dichotomous in their scoring. This is similar to the scoring mechanisms used for two-tiered assessments such as the TDT [40], which allows for a range of scores based on correctness and consistency between responses.

CMR formats are ideal for the U-STEP because they provide insight into student reasoning (much like FR items) while also allowing for online administration with streamlined scoring that can be automated. With an online administration comes more options for administering the assessment (e.g., it can be given outside of class so that faculty would not need to dedicate class time for it). Some of the items in the U-STEP are composed of a series of CMR questions.

D. Classical test theory

As mentioned in Sec. II B, many standardized physics assessments have narrow content scopes as a strategy to address content variability across institutions. Our prior work has shown this to be a relevant concern for upper-division thermal physics [1]; though there is some consistency across content covered in these courses at different universities, there is also a large amount of variability. Another motivation for keeping the content covered by an assessment relatively narrow is that one of the underpinning assumptions of classical test theory (CTT) is that the test be unidimensional, which is the idea that an assessment must focus on a single construct (e.g., focusing on "forces" or "motion," as opposed to "introductory physics"). While in practice no test is truly unidimensional, tests with narrow content focuses have historically been considered as more likely to satisfy the constraint of unidimensionality, though this assumption has not always been explicitly tested. There are a number of statistical tests that can determine whether an assessment is, in effect, unidimensional enough, and we will utilize these tests on U-STEP data in Sec. IX.

Conceptual assessments in discipline-based education research are commonly evaluated and validated using CTT. While other approaches to assessment validation exist, CTT provides a key baseline analysis for assessment validation regardless of other methods utilized. We used CTT as part of the validation process of the U-STEP 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 [42]. 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 pilot population tested is representative of the population of interest (e.g., upper-division physics students); this must be the case for the outcomes and measures of the assessment to hold in general for the full population. We note, however, that populations are rarely fully representative of all subpopulations, especially in physics, which is predominately White and male [43]. Thus, it is important with CTT to collect responses from a broad, representative sample of students to reduce the impact of the population-dependent statistics. This must be true in order to be confident that the test statistics output from CTT will hold for particular subgroups within the broader population of interest. Without a representative sample, outputs of the model, and therefore the test statistics themselves, are are likely to change when a different population is tested.

Additionally, since CTT requires unidimensionality, there are limits on the scope of content that can be included on an assessment. This poses a potential challenge for the U-STEP because thermal physics spans a large space ranging from classical thermodynamics to statistical mechanics. Thus, a thermal physics assessment that captures both of these areas may not be unidimensional enough, as it could test more than one construct. We present an investigation of unidimensionality of the U-STEP via exploratory factor analysis in Sec. IX.

E. Item bias and differential item functioning

Results of some CTT-validated assessments have found "achievement gaps" between certain groups of students (e.g., women and men [7,44–46]). For example, these studies show men tend to score higher than women on standardized physics assessments, such as the FCI. Studies on race-based performance differences are far less frequent in the physics education literature. Analysis of gender- (and race-) based performance gaps to identify bias have rarely been done *during* the development process of physics assessments. These types of analyses can aid in the identification of bias in assessment items which, if identified and rectified, could prevent artificially inflated performance differences.

Bias when referring to assessment items differs from the colloquial use of the term "bias," which may be intentional and/or based in prejudice. Instead, bias refers to characteristics of items that advantage or disadvantage certain groups.

Bias in items is characterized by a factor other than ability (such as gender, race, or socioeconomic status) impacting how likely one is to answer a question correctly [47]. Examples of factors that contribute to item bias could be the wording of an item prompt or unfamiliar contexts (e.g., using ice storms as a context could impact the scores of students in California, who would be unfamiliar with those storms). Appearance of bias in assessment items is problematic because it can mislead one in their conclusions or weaken inferences that can be made from results of the assessment. For example, if men outperform women on assessments, an unwary reader may conclude that men are more capable of performing well in physics than women. This is particularly concerning given the demographic composition of physics majors, which is predominately male (and White) [43], meaning it is likely that these assessments were validated in predominately White and male departments.

A common way to identify possible bias in items is differential item functioning (DIF). 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 possible 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 (i.e., item bias). 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, this warrants a closer investigation into the structure of that item.

Some researchers have investigated the FCI using DIF or similar techniques [8,47], 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. [48]), though some subsequent studies were unable to replicate those results [49]. Despite these contrasting results, it still suggests that bias can be identified and may be addressed or minimized during item development. To attempt to curb bias in the U-STEP, which would manifest as performance differences between groups that are attributable to something other than student ability, we conducted DIF analyses for two administrations of the U-STEP. This is discussed in Sec. VIII.

III. THE U-STEP

As described in previous work [1], the U-STEP development process began with a faculty content survey, which we used to identify content areas that faculty focus on in their upper-division thermal physics classes. While this survey did identify a number of consistent content areas taught by the vast majority of faculty, it also highlighted a significant amount of content variation. For the U-STEP, we focus on the areas of consistency, leaving development of an assessment instrument to target areas of variation for future work. After identifying the focal content areas, we developed free-response (FR) assessment items to target those topics, which we then piloted in one classroom in Fall 2019. Using responses to these items, we transformed the items into CMR and MC formats (as described in Ref. [1]). An example item (item 1), which addresses the concept of work, is presented in Fig. 1. As is typical of CMR items, this item is composed of one MC prompt followed by a MR prompt for reasoning.

The 15 items on the U-STEP vary in format: 6 items are MC (a single MC question or composed of multiple MC prompts); 6 items are CMR (a single CMR pair or multiple CMR pairs); and 3 items are CMR items with two or more MC prompts followed by a MR prompt. Different formats were used depending on AO being targeted and whether or not that AO, for example, includes an element of justification. The full set of items composing the U-STEP is presented in the Supplemental Material [50].

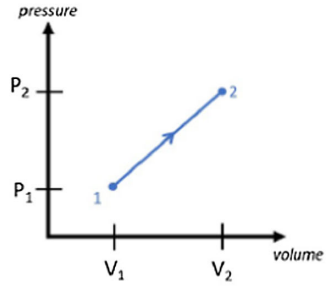
Based on results of the faculty content survey [1] we made the decision to make the U-STEP a post-test only assessment because several faculty indicated they expect their students to have little incoming exposure to thermal physics topics, such as entropy or statistical mechanics. Additionally, some jargon is unique to thermal physics and thus may be unfamiliar to students entering the course. In these cases, a pretest administration of the U-STEP would not produce meaningful measurements of student understanding of thermal physics content prior to taking the course. Additionally, making a different form of the U-STEP with accessible jargon and only familiar content (e.g., no entropy) limits straightforward comparisons, or determinations of gain, between pre- and postadministrations.

We conducted 13 student interviews and multiple pilot administrations while developing U-STEP items. After a FR pilot in Fall 2019, we piloted various forms of the CMR U-STEP during the Spring 2020, Fall 2020, and Spring 2021 semesters at multiple institutions. The FR pilot is an important step in the development of a CMR instrument to ensure that the reasoning elements are authentic and reflect actual student ideas. The assessment piloted in Spring 2020 was distributed in two versions in order to test all potential items without making the assessment too long. Based on the results of the Spring 2020 pilot, we removed some items and condensed the assessment into a single version that was piloted in Fall 2020 and Spring 2021 (see Sec. IV B). The versions between Fall 2020 and Spring 2021 were nearly identical, with only a few items undergoing small revisions between administrations. The student interviews and pilot administrations are described in detail in Sec. IV and Ref. [1].

A. Scoring U-STEP items

Careful consideration was taken when developing the scoring scheme for the U-STEP. A key consideration was

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

FIG. 1. A CMR item included on the U-STEP addressing the concept of *work* (item 1). The item is composed of a multiple choice (MC) prompt, followed by a prompt for reasoning. The MC portion of the item asks only for direction of energy flow due to work, as opposed to the sign of work, to accommodate for different sign conventions used across different classes (i.e., based on defining work as done *on* or *by* the system). Note that "not determinable" was included here to account for the possibility that, for example, a student believes that the fact that this gas is nonideal means we cannot determine work from a P-V diagram.

how many points to assign the MC (one correct response) questions and MR (partial-credit possible, related to reasoning) questions. 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. The research team collectively decided to assign up to 3 points for MR prompts, and investigated the impacts of changing the weighting of the MC prompt (i.e., 2 points vs 3 points).

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

Scoring of the reasoning portion of the CMR items is based on both correctness and consistency. It is possible on some items for a student to answer the MC prompt incorrectly and still receive some small amount of credit if reasoning is consistent with their incorrect response. For example, item 11 on the U-STEP (see Supplemental Material [50]) relied on recall of the Boltzmann factor; students could receive partial credit if their reasoning was consistent with a response corresponding to assumption of

Scoring Scheme:

	A	B*	C	D
a	0	0	0	0
b	-3	-3	0	0
c	0.5	1.5	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

Example Responses:

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

FIG. 2. Scoring scheme (left) and application examples (right) for item 1 of the U-STEP (Fig. 1). Multiple choice (MC) options (A–D) are in the top row of the scoring scheme, while the leftmost 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. The * indicates the correct MC response. See Supplemental Material [50] to see other items and their scoring schemes.

an incorrect sign in the exponent of the Boltzmann factor. Most MC prompts, with the exception of two,[5] were scored dichotomously (i.e., either fully correct or fully incorrect, with students receiving either 0 or 2 points).

Scoring for MR prompts is more complex, as it must account for both 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, correct or consistent reasoning elements could be worth anywhere between +0.5 to +2 points, while incorrect or inconsistent reasoning elements could be worth −0.5 to −3 points (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 worth 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 in order for either to count towards the score. For reasoning, total scores were bounded to remain between 0 points and 3 points.[6] An example scoring scheme and its application is presented in Fig. 2. Note Student 2 in Fig. 2 received partial credit for consistency despite selecting an incorrect answer for the MC prompt. The scoring schemes used for each item are presented in Supplemental Material [50].

There is some level of subjectivity in the judgements made regarding how many points each reasoning selection is worth. Other individuals may have developed different scoring schemes than the ones we developed. This is true of all nondichotomously scored assessments include those with a free-response format, and, ultimately, consistency in the scoring scheme is what matters [51]. However, it is unlikely any significant differences would appear in deciding which selections are definitively correct or incorrect.

After analysis of both interviews and validation statistics (see Secs. IV C and VII B), we modified some scoring schemes before finalizing them. This was generally done because the initial scoring seemed too harsh, causing a disproportionate decrease in scores for many students or students from particular demographic groups. In other cases, it was determined that certain elements were true, even if not relevant to the problem, and thus we changed scoring for these statements such that the selection would neither add nor subtract from the score.

Because of the formats of different items, some items were worth up to 15 points (if composed of multiple CMR pairs) while others were worth only 2 points (i.e., pure MC). To avoid some items dominating the scoring of the assessment, the research team decided to weight each item equally for determining overall assessment scores. 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

[5]These two prompts were part of the same item and had the same response options (see item 14 in Supplemental Material [50]) which were composed of two parts—work and heat. A partially correct response would receive 0.5 points out of 2 points.

[6]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 the reasoning score could not subtract from their MC or overall score, but rather only from the reasoning portion on that particular item (e.g., Student 3 in Fig. 2).

presented in Fig. 2, 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 student 3 received a score of 0.4. Normalized scores were all added together then divided by 15 (the maximum score possible for the normalized U-STEP) to determine the overall score for each student. While we ultimately used the normalized scoring scheme, we also tested an unnormalized scoring scheme and did not see meaningful differences in the achieved statistical measures of validity of the assessment.

IV. CONTEXT

The U-STEP underwent several in-class pilot administrations and student interviews, information from which was used for finalization of items and validation. In the following sections we present the solicitation of participants (Sec. IV A), pilot administrations of the U-STEP (Sec. IV B), and student interviews (Sec. IV C). As discussed in Sec. IV D the development process of the U-STEP overlapped, in part, with the COVID-19 pandemic and discuss the impact this has on our results.

A. Respondent samples

The analyses done for the U-STEP development (e.g., DIF) required a large, diverse sample. To achieve this, we solicited participation from multiple institution types serving various student populations for both the faculty content survey and the pilot administrations. This provided us with a more representative sample of students and faculty. For example, our Fall 2020 sample was 20% women and 59% White, while the Spring 2021 sample was 25% women and only 50% White. This is a comparable representation of women in our sample compared to representation in the field of physics; it is also a lower representation of White students than that in physics more generally (and a closer reflection of the representation of nonHispanic White people in the U.S. population) [43].

B. In-class piloting of the U-STEP

The U-STEP was piloted in a FR beta-assessment version in Fall 2019 and three times in a CMR format in Spring 2020, Fall 2020, and Spring 2021. Table I presents information about participating courses for each pilot semester, including overall response rate (determined from the number of student who were given the assessment totalled across all classes) and average response rate (averaged across all classes). Class sizes across the pilot semesters ranged from two students to over 100 students, with an average and overall response rate greater than 70% for all semesters. *An Introduction to Thermal Physics* by Daniel V. Schroeder [52] was the most commonly used textbook for each pilot semester: Fall 2019 ($N = 1$, 100%), Spring 2020 ($N = 12$, 86%), Fall 2020 ($N = 8$, 44%), and Spring 2021 ($N = 9$, 50%). This aligns with results from the faculty content survey, where Schroeder's text was the most commonly cited text used by upper-division thermal physics faculty [1].

We piloted 20 FR items in-person at one institution in Fall 2019 through four assessment versions. Details of the format of these pilot assessments are discussed elsewhere [1]. 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 a preliminary assessment still under development is a a more significant burden than administering a CMR assessment outside of class.

Key piloting information is summarized in Table I. The Spring 2020 assessment was piloted in two CMR versions, each composed of 13 items. Because of the varying class sizes at our piloting sites, the number of institutions receiving each version was not equal. Instead, we distributed each version to a different number of institutions such that the number of students receiving each version was roughly equal. Thus, we had 5 distributions of one version (with $N = 125$ students) and 9 distributions of the other version (with $N = 123$ students). Unequal response rates for each institution led to the first version receiving more responses ($N = 106$ responses received) than the second ($N = 79$). We note instructors are encouraged to give credit for participation; however sometimes the incentive was not enough to encourage students to participate. Note that Spring 2020 was the semester when the COVID-19 pandemic hit and many institutions chose to switch to remote or online instruction styles

TABLE I. Information about the Fall 2019, Spring 2020, Fall 2020, and Spring 2021 pilot administrations of the U-STEP. In all cases, N here refers to the number of students enrolled rather than the number of responses. The Fall 2019 assessment versions were free-response, while the other versions were composed of CMR and MC items. Note the average response rate does not include classes with 0% response rate ($N = 1$ for Spring and Fall 2020).

		N_{students} per class			N_{students}	Response rate	
	$N_{\text{institutions}}$	Average	Minimum	Maximum	Total	Overall	Average
Fall 2019	1		not applicable		67	91%	N/A
Spring 2020	14	18	2	90	248	75%	78%
Fall 2020	10	23	8	86	227	75%	73%
Spring 2021	18	19	3	110	349	82%	91%

mid-semester. COVID-19 lockdowns and remote learning, as discussed in Sec. IV D, may have caused some burnout for students, contributing to lower response rates.

The Fall 2020 assessment was piloted in one version and composed of 15 items, representing the first composition of items that would eventually be included in the final U-STEP. Items not included on the final version were dropped based on low difficulty and discrimination values, concerning student response patterns, and/or to avoid redundancy. In one instance, we combined two items (items 3 and 21) because they addressed very similar content. Of the 10 piloting institutions, 9 received responses, contributing to a total of 170 responses; individual class response rates (not including the class with no responses) ranged from 30% ($N = 1$ class) to 100% ($N = 2$ classes). We piloted the same items in the Spring 2021 semester that were piloted in the Fall 2020 semester, with some items undergoing minor revisions between the administrations based on the Fall 2020 student responses and student interviews. All 18 of the Spring 2021 piloting classes received responses, contributing to a total of 285 responses; individual class response rates ranged from 64% ($N = 1$) to 100% ($N = 9$).

Students responses were filtered to identify any responses that could be considered invalid for some reason. For example, all responses that took less than 5 min from start to finish were dropped because 5 min is roughly the amount of time necessary to read skim through the exam in full—suggesting that these students did not take the diagnostic seriously. Additionally, any student who skipped or only partially responded to 5 or more of the items on the U-STEP were also dropped from the dataset. After these drops, the total number of responses retained for each semester was $N = 185$ for Spring 2020, $N = 164$ for Fall 2020, and $N = 277$ for Spring 2021. Overall response rates are summarized in Table I.

For each piloted assessment (and student interview, see Sec. IV C), students were asked to provide demographic information.[7] This was used for the analyses discussed in Sec. VIII, and to ensure we had a broad range of students and perspectives included during item development. Student demographics for each pilot administration can be found in the Supplemental Material [50]. Demographic information was collected at the end of the U-STEP on the last page of the assessment, to minimize possible impacts of stereotype threat [53].

C. Student interviews

We conducted student interviews throughout item development and revision ($N = 13$). After CMR items drafting, prior to being piloted in Spring 2020, we conducted 5 interviews in a pencil-and-paper format, 2 and 3, respectively, for the two version drafts that we piloted in classes later that semester.[8] Additionally, prior to the Fall 2020 pilot, we conducted 5 remote interviews via Zoom using a single CMR version. To facilitate these interviews, we utilized the same online platform used for distributing the pilot assessment (i.e., Qualtrics) and Zoom screen sharing. All interviews were audio and video recorded. Video recording was included to capture any hand motions or facial expressions that might help with interpreting students' interaction with the exam or thought process.

After the Fall 2020 pilot, we conducted 3 additional remote validation interviews. Thus, each item underwent 12–13 validation interviews before finalization.[9] During interviews, participants were prompted to work through the U-STEP while talking aloud through their reasoning. During this portion, the interviewer did not interact except to remind participants to think aloud or to answer logistical questions asked by participants. After finishing the U-STEP, the interviewer would ask follow-up questions based on students' reasoning and selections, and would also talk interested students through the correct answers. All interviewed students provided demographic information at the end of the interview. Most interviews were conducted with men ($N = 10$) and White students ($N = 10$). Note that though 10 men and 10 White students completed interviews, these groups did not fully overlap, and thus fewer than 10 White men were interviewed. Interviewed students came from four different institutions, with 10 students from one institution and 1 student from each of the remaining 3.

D. A note on the COVID-19 pandemic

All semesters of data collection with CMR versions of the assessment overlapped with the COVID-19 pandemic. Thus, many courses were taught at least partially remotely or online during all or part of those semesters. This had little impact on the administration of the assessment, as the intention for the U-STEP was always for it to be administered in an online format. However, the disruption caused by the pandemic on university instruction likely had impacts on who was enrolled in physics courses as well as who completed the assessment during this time period. This, along with the hybrid or remote nature of courses used for piloting, may have impacts on the use and interpretation of the U-STEP for in-person classes; these types of impacts can be revisited and investigated once in-person instruction resumes. Additionally, the fluctuation in course format caused by the pandemic had implications on

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

[8]Interviews happened the week before most campuses closed due to COVID-19. Thus, fewer interviews were conducted than originally intended.

[9]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 part of preliminary work.

the analyses we could perform to investigate the reliability of the U-STEP, as discussed in the following section.

V. RESULTS: RELIABILITY

Reliability refers to the extent to which a measurement remains the same when done repeatedly, or the stability of a measurement over time [54]. There are several ways to confirm an assessment's reliability, including determination of internal consistency measures, time-test correlations, test-retest comparisons, and partial sample comparisons [42].

We measured the internal consistency for the U-STEP by determining Cronbach's alpha. It is recommended this coefficient has a value of $\alpha \geq 0.7$ to be considered acceptable for group comparisons [55]. The two Spring 2020 versions yielded values of $\alpha = 0.71$ and $\alpha = 0.69$; the value for the Fall 2020 pilot was $\alpha = 0.78$ and the value for the Spring 2021 pilot was $\alpha = 0.82$. This indicates the final version of the U-STEP is reliable at the level of group measurement as needed for this type of assessment.

We also conducted time-test correlations for each pilot semester, which look at the relation between a student's score on the assessment and the amount of time they spent completing it. High time-test correlations indicate that time constraints or time expenditure are having a significant impact on students' performance. As conceptual assessments are not intended to measure students ability to answer questions quickly, low time-test correlations are indicative of a more reliable instrument. Timing information was collected automatically by the online platform; however, this timing information represents the amount of time the student spent with the link open, not necessarily how long they actually spent working on the items and, thus, typically overestimated the amount of time a student spent. Making no corrections for this, time-test correlation coefficients for the Spring 2020 versions were -0.22 and -0.05, and dropped to -0.01 and -0.03 for the Fall 2020 and Spring 2021 semesters, respectively.

Perhaps a more realistic measure of the relation between time spent and score comes from narrowing our focus to the subset of students whose overall time spent is consistent with what we might expect in a classroom environment. To do this, we recalculated time-test correlations for students who completed the assessment in no more than 50 min and no less than 10 min (70% of responses fell into this range). This time frame was selected because the assessment is designed to be completed within a 50 min class period, and it takes approximately 10 min to read through the assessment in its entirety without trying to answer the questions—any student completing the assessment in less than 10 min is unlikely to be thinking seriously about the items. The time-test correlation coefficients for this subset of the students in the Spring 2020 versions were 0.03 and 0.38, and then 0.17 and 0.08 for the Fall 2020 and Spring 2021 semesters, respectively. This indicates assessment scores were not significantly influenced by the amount of time spent on completing the assessment, providing evidence for the time-test reliability of the U-STEP.

Test-retest comparisons involve administering a test on two separate occasions under similar testing conditions with no intervening learning opportunities [42]. They provide insight into the extent to which individual students' scores are stable. One common proxy for test-retest reliability is the examination of average scores for the course as a whole across multiple semesters of a single course where the student population and instruction are consistent. Unfortunately, disruptions due to the COVID-19 pandemic made test-retest examinations for the U-STEP unfeasible. A test-retest comparison would require a steady university experience to compare between semesters; COVID-19 disrupted the classroom experiences of students, which could in turn impact assessment performance. The Spring 2020 semester was taught partially in person at most universities, while the Fall 2020 and Spring 2021 semesters were largely taught remotely or in a hybrid format. These differences, combined with the fact that many thermal physics courses are taught during only one semester make a robust test-retest comparison unfeasible with the current dataset and will instead be the subject of future work.

We did conduct a correlation between the Fall 2020 and Spring 2021 semesters for the one institution that participated both semesters (taught by the same instructor). The average scores on the U-STEP for these two semesters were 58% and 60%, respectively. We also conducted correlations between average item scores each semester, and found correlation coefficients of 0.94 (when considering all items) and 0.96 (when considering all items except item 4, which underwent a significant revision, see Fig. 3). We note these semesters were the only semesters with similar testing conditions (i.e., the two semesters taught with hybrid-remote instruction) and had similar overall course averages (86% and 90% for the fall and spring, respectively). While these high correlations provide preliminary evidence for test-retest reliability of the U-STEP, these results should be interpreted with caution. These findings come from only a single institution and this institution has a well-documented fall-spring effect due to selection effects between on- and off-semester courses.

Partial-sample comparisons allow one to correct for selection affects caused by response rate less than 100%. The most common proxy for partial-sample reliability is achieved by comparing the pre- and postscores of both the matched an un-matched student population (as is done in Ref. [56]), thus providing a sense of whether differential attrition of, for example, lower performing students has a significant impact on the overall course average. Since the U-STEP is only administered as a post-test, pre-post partial-sample comparisons are not possible. However, as differential attrition is expected in most educational environments, we encourage any instructor or researcher to

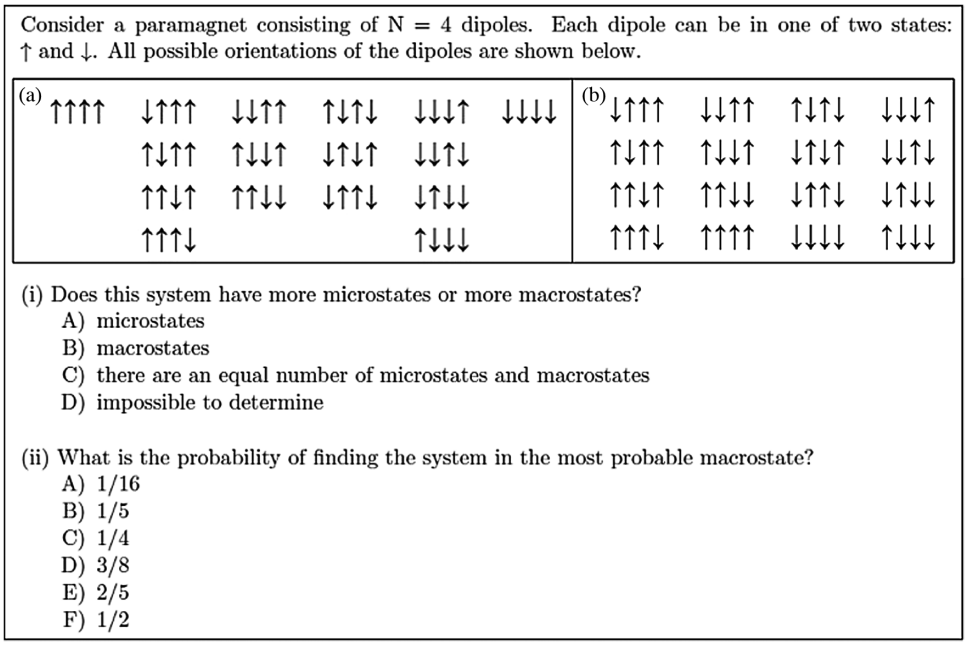


FIG. 3. Item 4 of the U-STEP: (a) The figure presented with the item in Spring 2020 and Fall 2020 and (b) the figure presented in Spring 2021. When the item used (a), a large fraction of students chose C for (ii), which aligns with the longest column being the most probable. After the figure changed from (a) to (b), more students selected D for (ii), increasing both the item difficulty and discrimination drastically (see Table II).

carefully consider the potential impacts of partial-sample effects when interpreting their students' performance on research-based assessments. This is particularly important when unforeseen disruptions occur that have disproportional impacts on particular subgroups of students (e.g., the COVID-19 pandemic).

VI. RESULTS: FACE VALIDITY

Validity refers to a measure of the extent to which a test or assessment measures what is says it does, has implication for the interpretations that can be inferred from produced scores [42]. Face validity refers to the extent to which experts and test takers view content on a test as relevant to, and appropriate for, the targeted context (e.g., upper-division thermal physics) [57]. Here, we discuss two types of face validity: content validity and construct validity.

A. Content validity

Content validity refers to how well the assessment covers the targeted content domain (e.g., thermal physics) [42]. Content validity is typically established early in the development process, while other types of validity (see Secs. VI B and VII A), are established after the assessment has been piloted. Content validity can be addressed by several routes, including expert input and review.

The first approach to ensuring the content validity for the U-STEP was to solicit faculty input from the beginning of item development. This was done through the faculty content survey, which identified content commonly covered in upper-division thermal physics (see Ref. [1]). This process ensured that the topics within the U-STEP would address key topics within the domain of thermal physics that were valued by the majority of instructors. Results from the survey were used to inform writing of assessment objectives (AOs), which we 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 aided in revisions and finalization of the AOs used to develop items. Examples of AO feedback included experts expressing that some AOs were unclear or confusing, often with suggestions for rewording particular AOs. For example, one set of AOs referred to *objects* exchanging energy; incorporating feedback led to this term being changed to *systems*. Other feedback was related to relevance in the thermal physics course they usually teach. For example, one expressed concern over the AOs focusing on free energies and enthalpy, saying those are rarely focused on in their course. This aligned with results from the faculty survey [1], and thus no items on the final U-STEP covered this content.

B. Construct validity

Construct validity is associated with ensuring that the assessment is accurately targeting the characteristics it claims to measure in order to ensure valid interpretations

of the results [42]. To establish construct validity, throughout item development and revision, we conducted student interviews, as described in Sec. IV C. These interviews were designed to verify that item prompts and response options were interpreted by students as intended; when this was not the case, we revised the items 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 we defined all variables in the preceding prompt). Additionally, we analyzed interviews to ensure students' selections aligned with their articulated reasoning. In only a few instances[10] did students' choices not align with their reasoning. Often this was due to reading the response options too quickly or not remembering the definition of a term (e.g., state function) or entity (e.g., the Boltzmann factor). We made a small number of revisions to items after the interviews; these revisions happened between each set of interviews (as opposed to after all 13 were conducted).

VII. RESULTS: STATISTICAL VALIDITY

Other measures of the validity of an assessment are produced via statistical analysis of student responses from pilot administrations. This includes establishing *criterion validity*, which involve comparing assessment performance to other relevant measures [42], as well as determining item difficulty and discrimination values using CTT. In the following sections, we present two criterion analyses conducted at one institution for the Fall 2020 and Spring 2021 pilot administrations of the U-STEP. We also present validation statistics determined via CTT using our full set of respondents for each piloting semester.

A. Criterion validity

For this study, overall assessment averages were compared to students' average exam scores and final course grades for a subset of respondents ($N = 76$ and $N = 45$ for the Fall 2020 and Spring 2021 semesters, respectively). This analysis was done with a single institution, as we only had access to grade data at one institution each semester. The course this sample was pulled from was taught by the same instructor both semesters in a hybrid-remote format.

We calculated Pearson correlation coefficients comparing each students' overall assessment score with their average for all course exams, as well as their final course grade. Correlations between students' average exam scores and achieved assessment score was 0.43 for Fall 2020 and 0.60 for Spring 2021. The correlation between students' final course grade and assessment score was 0.36 in Fall 2020 and 0.59 in Spring 2021. Some consider the range of 0.4–0.7 to be an acceptable range for validity correlations, which is consistent with the thresholds used historically in PER, though others have argued coefficients within that range are too low to account for a sufficient amount of variance between scores [58]. Our Spring 2021 correlations fell within the range of 0.4–0.7, while those from Fall 2020 straddled the lower bounds of this range.

There are several factors that could have affected the above correlations. The semester these data were collected was during a hybrid remote or in-person 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; this differs from the U-STEP, for which students were asked to not access any resources. Additionally, for all exams with exception of the final, students were able to do test corrections to improve their exam scores. It is of note that the instructor of the course had more experience teaching remotely and writing take-home exams in the spring semester; it is possible that the instructor was able to write more effective exams in the spring due to more practice. This would cause the exam scores, and thus final course grade, to more accurately reflect student knowledge and understanding, which is consistent with the increased correlations observed for that semester.

Exams were also weighted less heavily during these semesters than is typical for in-person iterations of the considered course. These course modifications may have changed the nature of what exams were testing, and, therefore, may not represent as appropriate a 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. However, these results are a promising, preliminary indication of the criterion validity of the U-STEP, and future work will expand these analyses to additional semesters and institutions.

B. Classical test theory

We also calculated standard test validation statistics using CTT. This process helped 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 difficulty. Item difficulty is a measure of how difficult an item is to answer correctly, and is reported as the average score on the item [59]. This means that higher difficulty values actually represent 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 [42,59].

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

Higher discrimination values indicate better differentiation between high and low performers.

Table II presents item difficulty (b) and discrimination (a) measures for the Spring 2020, Fall 2020, and Spring 2021 pilot administrations of the preliminary and final U-STEP items.[11] 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, we found item difficulties by averaging all scores achieved by individuals for each item; our difficulty values range from 0 to 1.

For the Spring 2020 pilot, item difficulties ranged from 0.14 to 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 0.31 to 0.76. The Spring 2021 pilot (same version as Fall 2020, with minor revisions) difficulties ranged from 0.29 to 0.76. The literature suggests ideal difficulties lie with the range of 0.30–0.90 [59]; only one item piloted in Spring 2021 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 and Spring 2021, the overall difficulties were 0.52 and 0.54, respectively. This number is consistent with overall difficulties observed in other upper-division conceptual assessments [51] and suggest that the U-STEP is a challenging instrument for this population of undergraduate physics majors.

Discrimination values for items were determined using a Spearman correlation between item scores and average score on the rest of the assessment [60]. A Spearman correlation was chosen due to the nonnormal distributions of item scores. For the Spring 2020 pilot, item discrimination values ranged from 0.14 to 0.51. For the Fall 2020 pilot, item discriminations ranged from 0.21 to 0.56. For the Spring 2021 pilot, item discriminations ranged from 0.22 to 0.63. The literature suggests these values should lie above 0.30 [59]. One item in the Spring 2021 pilot fell below this threshold with $a = 0.22$ (item 1). The item with the highest discrimination value was the same for the Fall 2020 and Spring 2021 pilots, as was true for the item with the lowest discrimination value.

The item with the lowest discrimination (item 1) was also the most difficult ($b = 0.31$ and 0.29 for the Fall 2020 and Spring 2021 administrations, respectively). This item is shown in Fig. 1. An analysis of student work showed most students selected "work caused energy to enter the system" (option A). This indicates student difficulties with recognizing the direction of energy flow due to work, and possibly a misapplication of concepts to nonideal gases. Many students indicated internal energy increased—a statement that is true for ideal gases but not necessarily for a nonideal gas. We have opted to retain this question despite its lower difficulty and discrimination values as we believe student responses here reflect an important difficulty with nonideal gases that has implications for instruction in upper-division thermal physics classes.

We found overall assessment discrimination, as measured by Ferguson's delta, to be 0.987 and 0.974 for the two Spring 2020 versions, 0.992 for the Fall 2020 administration of a single version, and 0.996 for the Spring 2021 administration. These meet the desired requirement of $\delta \geq 0.9$ [59]. Note the unnormalized scoring scheme was used to determine Ferguson's delta as this statistical test requires discrete scoring bins.

Six items included on the Spring 2020 pilot were ultimately dropped. This was due largely to concerns that the exam would be too long if all items were included. Items were selected for removal either because they did not reach the standard CTT thresholds for difficulty and/or discrimination (e.g., items 17, 19, and 20), or to balance out the content coverage of the exam with its length (e.g., item 18). We made one significant item revision based on our CTT analyses. As can be seen in Table II, item difficulty for item 4 was 0.54 in Spring 2020 and 0.51 in Fall 2020, and rose to 0.70 in Spring 2021. This change in difficulty (in addition to improved discrimination) resulted after revision to the figure provided in the item. Item 4, presented in Fig. 3, asks about the probability of most probable macrostate for a set of $N = 4$ dipoles. As can seen in Fig. 3, the provided figure of the arrangement of dipoles was changed between the Fall 2020 [Fig. 3(a)] and Spring 2021 [Fig. 3(b)] administrations.

We suspected the length of the columns was causing students to select the most common distractor, $P = 1/4$. With the arrangement in Fig. 3(a), students who know the highest macrostate probability aligns with the highest number of microstates in that state may have been inclined to say the most probable macrostate has 4 microstates, because that is the length of the longest two columns, leading them to conclude $P = 4/16 = 1/4$. For the Spring 2021 administration, we chose to use a new figure [Fig. 3(b)], which moved the leftmost (all $\uparrow$) and rightmost (all $\downarrow$) states to the bottom of the $2\uparrow - 2\downarrow$ columns. We believe this led to less reliance on the figure to determine the most probable state and instead relying on physical intuition or direct state counting. This lead to more students concluding that $2\uparrow - 2\downarrow$ would be the most probable state, with a correct probability of $P = 6/16 = 3/8$.

VIII. RESULTS: DIFFERENTIAL ITEM FUNCTIONING

As mentioned in Sec. II E, DIF involves statistical comparisons of subgroups within a population to investigate the extent to which an individual item may be indicating different abilities for students with similar scores on the assessment overall. This can be used to identify bias within items (i.e., characteristics of items that disadvantage

[11]Note these analyses used the normalized scoring scheme discussed in Sec. III A.

TABLE II. CTT analysis results—difficulties (b) and discriminations (a)—for the Spring 2020, Fall 2020, and Spring 2021 pilot administrations of the U-STEP items. N values for the Spring 2020 administration change due to the different versions of the assessment piloted and varied number of institutions receiving each version. (Each version in the Spring was composed of a set of 6 anchor items and 7 secondary items, which differed based on version.) N values for the Fall 2020 and Spring 2021 pilots were constant ($N = 164$ and $N = 277$, respectively) due to only a single version being piloted.

			Spring 2020			Fall 2020 ($N = 164$)		Spring 2021 ($N = 277$)	
Item	Topical area	Item focus	N	*b*	*a*	*b*	*a*	*b*	*a*
1 (CMR)	Work	Work from P-V diagram for nonideal gas	99	0.28	0.14	0.31	0.24	0.29	0.22
2 (CMR)	First law	Internal energy change, work, and heat from P-V diagram	169	0.47	0.23[a]	0.52	0.50	0.50	0.43
3 (CMR)	Heat	Differences between heat and temperature	not applicable	not applicable	not applicable	0.45	0.40	0.50	0.49
4[b] (CMR)	Stat. mech.	Micro or macrostates, probability (P)	169	0.54	0.35[a]	0.51	0.27	0.70	0.44
5 (MC)	Stat. mech.	Multiplicity for interacting systems	99	0.75	0.40	0.76	0.25	0.76	0.40
6 (MR)	Equilibrium	Thermal, mechanical, diffusive equilibrium	99	0.72	0.41	0.63	0.52	0.62	0.53
7 (CMR)	Entropy	Mixing different monatomic gases	70	0.58	0.45	0.66	0.33	0.69	0.47
8 (CMR)	Entropy	Carnot engine and entropy	99	0.51	0.38	0.51	0.47	0.54	0.53
9 (CMR)	Entropy	Heat flow between solids	70	0.49	0.43	0.51	0.45	0.57	0.52
10 (MC)	Energy	Relation between heat capacity and degrees of freedom	70	0.60	0.48	0.57	0.30	0.65	0.44
11 (CMR)	Stat. mech.	Boltzmann factors, Z^c, and P with energy shift	169	0.47	0.44[a]	0.42	0.55	0.39	0.63
12 (MC)	Stat. mech.	Entropy and Z^c for single state	99	0.61	0.42	0.51	0.50	0.44	0.56
13 (CMR)	Engines	Heat and work for legs of cycle on P-V diagram	99	0.40	0.22	0.46	0.40	0.45	0.32
14 (CMR)	Engines	Entropy-temperature diagram, heat, and work	70	0.23	0.26	0.31	0.36	0.36	0.47
15 (MR)	Temperature	Graph of isotherm for ideal gas	70	0.59	0.36	0.68	0.30	0.65	0.35
16 (CMR)	Temperature	Differences between heat and temperature	169	0.65	0.30[a]	not applicable	not applicable	not applicable	not applicable
17 (CMR)	Stat. mech.	Fundamental assumption of stat. mech. [52]	169	0.52	0.13[a]	not applicable	not applicable	not applicable	not applicable
18 (CMR)	Heat	Heat from P-V diagram	169	0.40	0.30[a]	not applicable	not applicable	not applicable	not applicable
19 (MC)	Energy	When equipartition holds	99	0.19	0.33	not applicable	not applicable	not applicable	not applicable
20 (MC)	Stat. mech.	Probability of degenerate state	70	0.14	0.20	not applicable	not applicable	not applicable	not applicable
21 (CMR)	Heat	Differences between heat and temperature	70	0.51	0.51	not applicable	not applicable	not applicable	not applicable

[a]The discrimination from anchor items for the Spring 2020 pilot are presented as averages across the two versions; all items with this symbol were anchor items that appeared in both versions piloted in Spring 2020.

[b]The figure provided in item 4 was revised between the Fall 2020 and Spring 2021 administrations (see Fig. 3).

[c]Partition function.

one group over another) that can lead to artificial performance differences. Some DIF analyses of physics assessments, such as those described here, have been conducted previously (e.g., Ref. [8]), but these were typically done *after* the assessment has been formalized.

Often differences in performance based on gender and/or race are identified using averages of all students within each considered population. DIF instead only compares the performance between students who have similar overall assessment scores. The 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. We did this analysis for the Spring 2020, Fall 2020, and Spring 2021 pilot administrations separately to inform changes to the items in real time, but will focus on the Fall 2020 and Spring 2020 results here. Additionally, since all items but one underwent only minor revision, we will combine these two semesters in order to increase statistical power given differential representation across gender and race in our sample.

To account for different averages for different courses, we first converted students overall assessment scores into z scores based on the individual course averages and standard deviations. We then sorted students into performance levels by ranking students by overall assessment z score, from highest scores to lowest. Then, the bottom and top 25th percentiles (i.e., the 25% of students with the lowest overall z scores and 25% with the highest overall z scores) were selected. We split these two groups by gender or race to compare scores across subgroups. 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, in the combined Fall 2020–Spring 2021 analysis, the top and bottom 25th percentiles were each composed of 110 students; the top 25th percentile contained 19 women and 83 men.[12]

The purpose of splitting students into percentile groups is to allow for comparison of students with similar overall performance. If we instead focused 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 group overall legitimately are showing lower performance due to, for example, differential access to resources or prior experience.

The analyses presented here focus on two genders (men and women) and three racial categories [Asian, White, and underrepresented minority (URM)]. These groupings were largely informed by demographic representation in science, technology, engineering, and mathematics (STEM) [43] as well as in our dataset. Men are overrepresented compared to women in STEM. Similarly, Asian and White students are overrepresented in STEM compared to URM students. While Asian and White students are both overrepresented, they are split into separate categories in this analysis due to their potentially distinct racial experience. This is an important distinction as race influences the experiences students encounter when pursuing STEM degrees. For example, Asian students may encounter racism when pursuing STEM degrees that White students do not [61]. The URM category was not separated into different racial categories due to low N. For example, our samples consisted of only $N = 17$ Black students across all pilot administrations. We note students who selected prefer not to answer on the demographic form are not included in the DIF analyses, though may have been in the upper or lower 25th percentiles. Similarly, nonbinary students are not included in the gender analysis due to low N, though nonbinary students may have been in the upper or lower 25th percentiles.

As indicated above, DIF analyses were conducted throughout the piloting process. Two items were identified during these analyses as showing potential bias. For the Spring 2020 pilot, only one item (item 16 in Table II) 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 [62]; we detected no statistically significant differences between men and women or Asian and URM students. This item was combined with item 21 to produce a single CMR item (item 3) for later pilots. Additionally, Item 13 showed statistically significant evidence of gender differences after the Fall 2020 pilot. Item 13 relies on recall of the term *adiabatic* and is one of the only items on the U-STEP that requires students to remember the definition of a more technical piece of jargon.[13] As such, we wanted to examine whether this might be the source of the differential performance. Analyses of response patterns to the item found similar frequencies of selection of the distractor requiring knowledge of this term between compared groups showing significant differences in item scores. We saw an average of 6.5% of men and 10% of women select this distractor in Fall 2020, and 12% of White and 16% of URM students select it in Spring 2021. As these frequencies are similar across comparison groups, the distractor is likely not the source of the performance differences. 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, provision of this definition would make the problem trivial (see Supplemental Material [50]).

[12]Others in this percentile were nonbinary or did not report demographic information.

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

For the purpose of identifying items that might be problematic in the final version of the U-STEP, we used a standard threshold for statistical significance of $p < 0.05$ (t test). Given the number of comparisons being made here (15 items each with 8 comparison tests each) this threshold is very liberal and is likely to result in false positives. To offset this, we were looking for evidence of consistent patterns across both ability levels (upper and lower 25th percentile) when identifying items that might demonstrate bias. Overall, 7 items on the U-STEP showed performance differences that were statistically significant at the 0.05 level in one or more of the gender or race comparisons (with effect sizes varying from 0.4 to 0.9). Of these items, one (Item 5 in Table II) had a statistically significant result in two racial comparisons with White students in the lower 25th percentile outperforming both Asian and URM students in the lower 25th percentile. Since this pattern was not replicated in the upper 25th percentile, no changes were made to this item as a result of this analysis. The other 6 items had only one statistically significant comparison and thus did not show consistent evidence of differential performance; these items also did not undergo modification. DIF analyses of the U-STEP items will continue as we aggregate larger datasets over time; however, these findings suggest that any DIF present in the U-STEP is smaller than can be detected with the current dataset.

IX. RESULTS: EXPLORATORY FACTOR ANALYSIS

As discussed previously, one implicit assumption of CTT as well as many other testing theories is that the instrument is unidimensional, i.e., all items target a single underlying construct. To examine the dimensionality of the U-STEP, we conducted an exploratory factor analysis (EFA) using data from the Spring 2021 pilot using the fa() program in R. The resulting scree plot is given in Fig. 4 and shows a single clear elbow with a single factor sitting well above the others. One method for identifying the number of significant factors suggests that only factors with eigenvalues greater than 1.0 are considered significant [63,64]. Figure 4 shows only one factor with an eigenvalue greater than 1.0, with all other eigenvalues falling well below this threshold. This suggests one dominant factor across our items. To confirm this finding, we also utilized a parallel analysis, which compares the eigenvalues of the U-STEP data with those generated from random data [65]; these analyses confirmed the findings from the scree plot in Fig. 4, suggesting a single dominant factor.

We also investigated individual item loadings for our factor analysis with 1 factor. Typically, recommendations within the social sciences suggest that items should have factor loading values between 0.4 and 0.7 to be considered as loading significantly onto a particular factor [66]; we found 12 of our 15 items fell within that range for our one-factor solution.[14] It is also recommended to drop items with loading factors less than 0.3 [64]; we found loading factors for all items were above 0.3 [66], with the exception of item 1 which was very close to this threshold (0.29). As discussed in Sec. VII B, we believe student responses to this item reflect an important student difficulty that has implications for instruction and thus have opted to retain this item in the final version of the U-STEP.

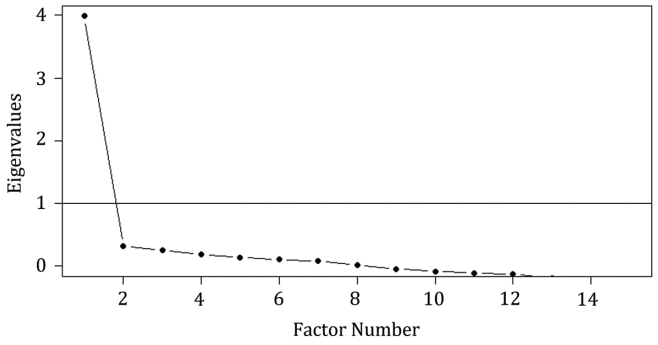


FIG. 4. Scree plot from EFA for the U-STEP showing a clear elbow after the first factor. A horizontal line is shown for an eigenvalue of 1, as it is suggested to only keep factors with eigenvalues greater than 1.

Based on these outputs, we ran a confirmatory factor analysis (CFA) assuming only 1 factor using the cfa() function in R to investigate the hypothesis that one loading factor was sufficient for the data set. We found the Tucker-Lewis index of factoring reliability to be 0.96; it is recommended this value lie above 0.95 to be considered a "very good" fit [67]. Additionally, the composite reliability was 0.84, which is above the minimum threshold of 0.8 [68]. We also found a root mean square error approximation (RMSEA) index of 0.03; it is recommended this value lie below 0.05 for a very good fit [67].

Our analysis points to one dominant factor, or construct, to which nearly all items can be connected. The topics covered by the U-STEP items include entropy, engines, energy, equilibrium, the first law, heat, statistical mechanics, temperature, and work. While an expert physics might anticipate that these topics are distinct, our findings suggests that this distinction is not reflected in the patterns in students responses on the U-STEP; rather, these analyses suggest that the U-STEP is statistically unidimensional. This finding supports the validity of the CTT analyses in the previous section and has implications for future work (see Sec. X). Additionally, this finding indicates that any attempt to separate out the items of the U-STEP into distinct subscales targeting topical areas considered distinct by experts is not statistically supported by this analysis.

[14]For the interested reader, specific factor loadings for items 1–15, respectively, were 0.29, 0.48, 0.56, 0.47, 0.44, 0.60, 0.51, 0.59, 0.56, 0.46, 0.70, 0.62, 0.34, 0.56, 0.38.

One important limitation of this factor analysis is that the U-STEP has both CMR and MC item formats, introducing the possibility of method effects. Method effects occur when items of different formats have different factor structures. Given the small number of pure MC questions ($N = 3$) and to provide some test for method effects, we removed the three MC questions and re-ran the confirmatory factor analysis with a single factor on just the remaining CMR items. Fit measures for the CMR items alone all remained above standard thresholds for good fit. However, an important limitation of this analysis is that it does not eliminate the possibility of method effects with the remaining 3 MC questions.

X. FUTURE WORK

Though we completed multiple validation studies, as presented in this paper, there are multiple areas for potential continued analyses of the U-STEP including ongoing analysis using various item response theory (IRT) models [69]. This technique has advantages over traditional methods because of the assumptions of the models we will use (variations of the Rasch model), which mathematically separates item difficulty from student ability.[15] As a result of this, test statistics derived from IRT are less dependent on student populations than CTT (given the pilot sample is fairly representative with a large enough N). IRT also provides another method for conducting DIF analyses to identify item bias.

Once we complete validation analyses of the U-STEP (through both CTT and IRT), we plan to develop infrastructure for distributing the U-STEP online and encoded scoring scheme (to allow for streamlined scoring). This process would include developing and implementing automated methods for reporting student performance to instructors with appropriate comparison data to help instructors to interpret their students performance. We also plan widely publicize the U-STEP to instructors and make it available on PhysPort.

Additionally, the faculty content survey discussed in Sec. III [1] has contributed to the development and validation of an assessment of scientific practices associated with upper-division thermal physics, and our IRT analysis of the U-STEP will help inform the validation of this new assessment. This assessment, the Thermal And Statistical Physics Assessment (TaSPA), will provide actionable feedback to instructors to help them enact pedagogical and curricular changes [70,71].

[15]Note "ability" is historically used term in IRT literature and 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.

XI. CONCLUSIONS

Here, we present the development and statistical validation (via classical test theory) of a new, upper-division thermal physics assessment: the Upper-level Evaluation for Thermodynamics and Statistical Physics. Based on three semesters of data collection including more than 600 student responses, we provide evidence for the reliability and statistical validity of the U-STEP for a diverse population of upper-division physics students. In particular, we found most U-STEP items fell within recommended difficulty and discrimination values, and that student response patterns in interviews and pilot test indicated they are consistently and accurately interpreting the prompts. Additionally, we saw no items showing consistent evidence of differential item functioning and thus potentially containing item bias. Exploratory factor analysis provided evidence that the U-STEP is largely unidimensional. This assessment will serve as the first *upper-division* thermal physics assessment in PER, and the first PER-based assessment to include statistical mechanics content.

Our study does have limitations including being conducted largely during the ongoing COVID-19 pandemic. COVID-19 limited some of the analyses we could conduct, such as comparing performance on the U-STEP across semesters. However, these studies could be conducted in the future, when institutions shift back to in-person instruction. Additionally, piloting the FR versions at a single, large research institution, as well as conducting interviews with students from largely that same institution, does limit the type of student responses and insights we could incorporate while developing items. However, we made efforts to solicit interview participants from multiple institutions. Additionally, we included a large set of institutions and student populations for online piloting of the U-STEP and in this we included an "other" box to capture reasoning patterns that may not have been originally captured with the FR versions or in student interviews.

The U-STEP is the first assessment available for evaluating both upper-division thermodynamics and statistical mechanics content understanding with CMR items. As such, it has the potential to serve as an important tool toward improving instruction and student learning in this core undergraduate physics course. It also represents one of the first assessments for which the possibility of bias in the form of differential item functioning was explicitly addressed during the development process. Following the recommendations of others (e.g., Ref. [8]), we encourage other assessment developers to similarly attend to bias in developing new assessments. Such efforts are critical to crafting instruments that provide fair and reliable measures of student learning.

ACKNOWLEDGMENTS

We thank the instructors who piloted the U-STEP in their courses, the students who took it, and the students who participated in interviews. We also thank the content experts who provided feedback on the assessment objectives that guided item development. This work was supported by funding from the Center for STEM Learning and the Department of Physics at the University of Colorado Boulder, as well as the National Science Foundation DUE Grant No. 2013332.

[1] K. D. Rainey, M. Vignal, and B. R. Wilcox, Designing upper-division thermal physics assessment items informed by faculty perspectives of key content coverage, Phys. Rev. Phys. Educ. Res. **16**, 020113 (2020).

[2] B. R. Wilcox and S. J. Pollock, Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020124 (2014).

[3] B. R. Wilcox and S. J. Pollock, Multiple-choice assessment for upper-division electricity and magnetism, in *Proceedings of PER Conf 2013, Portland, Oregon*, 10.1119/perc.2013.pr.079.

[4] 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, Phys. Rev. Phys. Educ. Res. **15**, 010135 (2019).

[5] B. Pollard, M. F. J. Fox, L. Ríos, and H. J. Lewandowski, Creating a coupled multiple response assessment for modeling in lab courses, in Proceedings of PER Conf 2020, virtual conference, 10.1119/perc.2020.pr.Pollard.

[6] T. I. Smith, P. Eaton, S. W. Brahmia, A. Olsho, A. Boudreaux, and C. Zimmerman, Toward a valid instrument for measuring physics quantitative literacy, in Proceedings of PER Conf 2020, virtual conference, 10.1119/perc.2020.pr.Smith_T.

[7] M. Lorenzo, C. H. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).

[8] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010103 (2018).

[9] 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, Am. J. Phys. **83**, 5 (2015).

[10] J. W. Clark, J. R. Thompson, and D. B. Mountcastle, Comparing student conceptual understanding of thermodynamics in physics and engineering, AIP Conf. Proc. **1513**, 102 (2013).

[11] D. E. Meltzer, Observations of general learning patterns in an upper-level thermal physics course, AIP Conf. Proc. **1179**, 31 (2009).

[12] T. I. Smith, J. R. Thompson, and D. B. Mountcastle, Student understanding of Taylor series expansions in statistical mechanics, Phys. Rev. ST Phys. Educ. Res. **9**, 020110 (2013).

[13] T. I. Smith, D. B. Mountcastle, and J. R. Thompson, Student understanding of the Boltzmann factor, Phys. Rev. ST Phys. Educ. Res. **11**, 020123 (2015).

[14] S. Yeo and M. Zadnik, Introductory thermal concept evaluation: Assessing students' understanding, Phys. Teach. **39**, 496 (2001).

[15] H.-E. Chu, D. F. Treagust, S. Yeo, and M. Zadnik, Evaluation of students' understanding of thermal concepts in everyday contexts, Int. J. Sci. Educ. **34**, 1509 (2012).

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

[17] P. Wattanakasiwich, P. Taleab, M. D. Sharma, and I. D. Johnston, Construction and implementation of a conceptual survey in thermodynamics, Int. J. Innovation Sci. Math. Educ. **21**, 29 (2013).

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

[19] 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, Am. J. Phys. **73**, 1055 (2005).

[20] 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, Am. J. Phys. **73**, 1064 (2005).

[21] 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, Chem. Educ. Res. Pract. **15**, 320 (2014).

[22] P. G. Jasien and G. E. Oberem, Understanding of elementary concepts in heat and temperature among college students and K-12 teachers, J. Chem. Educ. **79**, 889 (2002).

[23] M. Sözbilir, A review of selected literature on students' misconceptions of heat and temperature, Boğaziçi University J. Educ. **20**, 25 (2003).

[24] A. A. Alwan, Misconception of heat and temperature among physics students, Proc. Social Behav. Sci. **12**, 600 (2011).

[25] 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, Phys. Rev. ST Phys. Educ. Res. **9**, 020112 (2013).

[26] D. E. Meltzer, Investigation of students' reasoning regarding heat, work, and the first law of thermodynamics in an introductory calculus-based general physics course, Am. J. Phys. **72**, 1432 (2004).

[27] D. E. Meltzer, Investigation of student reasoning regarding concepts in thermal physics, APS Forum on Education, 4 (2005).
[28] 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, Am. J. Phys. **70**, 137 (2002).
[29] T. I. Smith, W. M. Christensen, D. B. Mountcastle, and J. R. Thompson, Identifying student difficulties with entropy, heat engines, and the Carnot cycle, Phys. Rev. ST Phys. Educ. Res. **11**, 020116 (2015).
[30] B. R. Bucy, J. R. Thompson, and D. B. Mountcastle, What is entropy? Advanced undergraduate performance comparing ideal gas processes, AIP Conf. Proc. **818**, 81 (2006).
[31] W. M. Christensen, D. E. Meltzer, and C. Ogilvie, Student ideas regarding entropy and the second law of thermodynamics in an introductory physics course, Am. J. Phys. **77**, 907 (2009).
[32] R. Leinonen, M. A. Asikainen, and P. E. Hirvonen, Grasping the second law of thermodynamics at university: The consistency of macroscopic and microscopic explanations, Phys. Rev. ST Phys. Educ. Res. **11**, 020122 (2015).
[33] N. Crossette, M. Vignal, and B. R. Wilcox, Investigating graduate student reasoning on a conceptual entropy questionnaire, Phys Rev. Phys. Educ. Res. **17**, 020119 (2021).
[34] M. Loverude, Identifying student resources in reasoning about entropy and the approach to thermal equilibrium, Phys. Rev. ST Phys. Educ. Res. **11**, 020118 (2015).
[35] M. Sözbilir, What students' understand from entropy?: A review of selected literature, J. Baltic Sci. Educ. **2**, 21 (2003).
[36] E. M. Carson and J. R. Watson, Undergraduate students' understandings of entropy and Gibbs free energy, University Chem. Educ. **6**, 4 (2002).
[37] F. L. Lambert, Disorder-a cracked crutch for supporting entropy discussions, J. Chem. Educ. **79**, 187 (2002).
[38] R. Wei, W. Reed, J. Hu, and C. Xu, Energy spreading or disorder? understanding entropy from the perspective of energy, *Teaching and Learning of Energy in K–12 Education* (Springer, New York, 2014), pp. 317–335.
[39] T. A. Moore and D. V. Schroeder, A different approach to introducing statistical mechanics, Am. J. Phys. **65**, 26 (1997).
[40] C. Kamcharean and P. Wattanakasiwich, Development and application of thermodynamics diagnostic test to survey students' understanding in thermal physics, Int. J. Innovation Sci. Math. Educ. **24**, 14 (2016).
[41] B. Brown and C. Singh, Development and validation of a conceptual survey instrument to evaluate students' understanding of thermodynamics, Phys. Rev. Phys. Educ. Res. **17**, 010104 (2021).
[42] P. V. Engelhardt, An introduction to classical test theory as applied to conceptual multiple-choice tests, Getting started in PER **2** (2009).
[43] National Science Foundation, Women, minorities, and persons with disabilities in science and engineering (National Science Foundation, Washington, DC, 2017).
[44] 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** (2002) pp. 1–14, https://physics.indiana.edu/~hake/PERC2002h-Hake.pdf.
[45] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).
[46] S. J. Pollock, N. D. Finkelstein, and L. E. Kost, Reducing the gender gap in the physics classroom: How sufficient is interactive engagement?, Phys. Rev. ST Phys. Educ. Res. **3**, 010107 (2007).
[47] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis, Gender bias in the Force Concept Inventory?, AIP Conf. Proc. **1413**, 171 (2012).
[48] L. McCullough, Gender, context, and physics assessment, J. Int. Women's Studies **5**, 20 (2004).
[49] 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?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).
[50] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.18.020116 for contains the full U-STEP instrument, scoring rubrics for all U-STEP items, and demographic breakdowns for the pilot student populations.
[51] 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, Phys. Rev. ST Phys. Educ. Res. **11**, 020115 (2015).
[52] D. V. Schroeder, *An Introduction to Thermal Physics* (Addison Wesley, Reading, MA, 1999).
[53] C. M. Steele, A threat in the air: How stereotypes shape intellectual identity and performance, Am. Psychol. **52**, 613 (1997).
[54] J. Kirk, M. L. Miller, and M. L. Miller, *Reliability and Validity in Qualitative Research* (Sage, Newbury Park, CA, 1986), Vol. 1.
[55] J. M. Cortina, What is coefficient alpha? An examination of theory and applications., J. Appl. Psychol. **78**, 98 (1993).
[56] B. R. Wilcox and H. J. Lewandowski, Students' epistemologies about experimental physics: Validating the Colorado Learning Attitudes about Science Survey for experimental physics, Phys. Rev. Phys. Educ. Res. **12**, 010123 (2016).
[57] R. R. Holden, Face validity, The Corsini encyclopedia of psychology **4**, 1 (2010).
[58] M. W. Post, What to do with "*moderate*" reliability and validity coefficients?, Archives Phys. Med. Rehab. **97**, 1051 (2016).
[59] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).
[60] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. **15**, 72 (1904).
[61] E. O. McGee, B. K. Thakore, and S. S. LaBlance, The burden of being "*model*": Racialized experiences of Asian STEM college students, J. Diversity Higher Educ. **10**, 253 (2017).
[62] F. Wilcoxon, Individual comparisons by ranking methods, *Breakthroughs in Statistics* (Springer, New York, 1992), pp. 196–202.

[63] J. W. Osborne and A. Costello, Getting the most from your analysis, Pan **12**, 131 (2009).

[64] A. L. Comrey, Common methodological problems in factor analytic studies, J. Consulting Clinical Psychol. **46**, 648 (1978).

[65] J. C. Hayton, D. G. Allen, and V. Scarpello, Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis, Organ. Res. Meth. **7**, 191 (2004).

[66] A. Field, *Discovering Statistics Using IBM SPSS Statistics* (Sage, Newbury Park, CA, 2013).

[67] L.-t. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Structural Eq. Model. Multidisc. J. **6**, 1 (1999).

[68] R. G. Netemeyer, W. O. Bearden, and S. Sharma, *Scaling Procedures: Issues and Applications* (Sage Publications, Newbury Park, CA, 2003).

[69] B. R. Wilcox, K. Rainey, and M. Vignal, Methods for utilizing item response theory with coupled multiple-response assessments, presented at PER Conf 2022, Grand Rapids, MI (to be published).

[70] A. P. Jambuge, K. D. Rainey, A. Sirnoorkar, J. T. Laverty, and B. R. Wilcox, Assessment feedback: A tool to promote scientific practices in upper-division, in *Physics Education Research Conference 2020* (2020), pp. 234–239.

[71] K. D. Rainey, A. P. Jambuge, J. T. Laverty, and B. R. Wilcox, Developing coupled, multiple-response assessment items addressing scientific practices, in Proceedings of PER Conf 2020, virtual conference, 10.1119/perc.2020.pr.Rainey.