Exploring student reasoning in statistical mechanics: Identifying challenges in problem-solving groups

Ebba Koerfer* and Bor Gregorcic

Department of Physics and Astronomy, Uppsala University, Box 516, SE-751 20 Uppsala, Sweden

(Received 6 November 2023; accepted 9 January 2024; published 12 February 2024)

Statistical mechanics has received limited attention in physics education research and remains a relatively underrepresented topic even in research on upper-division physics courses. The purpose of this study was to explore potential challenges that physics students encounter when they solve statistical mechanics problems in groups. Adopting a grounded approach, we video recorded and analyzed nine small student groups engaging in collaborative problem solving on the topic. The analysis involved iterative thematic coding, which gave rise to ten emergent categories of challenges. These were later divided into two broad groupings: *challenges with concepts* and *challenges with problem-solving strategies*. In the first grouping, we list seven identified categories related to the concepts of macrostates and microstates, distinguishable and indistinguishable particles, temperature, entropy, energy, equilibrium, heat bath, the Boltzmann distribution, and the partition function. In the second grouping, we list three categories related to the inappropriate application of common relations, difficulty managing tensions between calculated results and qualitative reasoning, and coming up with definitions of new and inconsistent concepts. Some of our findings are supported by existing research on the topic, and others are previously unreported. Based on our findings, we propose that future research should investigate the relations between the identified challenges on one hand, and students' epistemological framing, reasoning, and use of multiple representations on the other. Finally, we suggest that teachers should spend time engaging students in a conceptual discussion of the central ideas of statistical mechanics, motivating the choice and pointing out limitations of commonly used toy models, and linking course content to real-world phenomena.

DOI: 10.1103/PhysRevPhysEducRes.20.010105

I. INTRODUCTION

Statistical mechanics and thermodynamics are both parts of thermal physics, and their relevance covers phenomena from everyday life to applications across fields such as biology, chemistry, engineering, solid-state physics, and astrophysics [1]. Teaching a rich and diverse subject such as thermal physics is not a straightforward task. It can be approached in many different ways depending on the education level and intended area of application. At the level of introductory physics, or in disciplines such as earth science or engineering, thermodynamics is often sufficient and so less attention is paid to the more sophisticated content of statistical mechanics.

Research on student learning in thermal physics mainly concerns thermodynamics content at introductory levels, see the extensive review by Dreyfus *et al.* [2] from 2015 on teaching thermal physics in introductory physics, chemistry, and biology. These studies have focused on central concepts such as temperature, heat, ideal gases, heat engines, and the second law of thermodynamics. Entropy is another frequently explored topic and has long been considered a challenging concept for students. Many studies have looked at macroscopic perspectives of entropy in introductory courses. One common finding is a tendency among students to "over-apply" the second law of thermodynamics, for example, to argue that the entropy must increase regardless of the context, e.g., Christensen *et al.* [3]. Another recurring theme is students struggling to distinguish the system from its surroundings [3,4] and confusion regarding which of them the second law should be applied to [5].

Physics education research (PER) on topics at the upper-division level has gained more focus over the last two decades [6]. Such research can provide valuable insight into the intermediate stage of students' transition from novice to expert physicists [7]. Topics such as quantum mechanics and electromagnetism have primarily been investigated. Upper-division thermal physics, where students are typically introduced to statistical mechanics content, is however less explored (some examples include Refs. [4–6,8–19]). A few of these studies have investigated the transition from introductory to upper-division content, which is connected to

*ebba.koerfer@physics.uu.se

bridging the macroscopic perspective of thermodynamics to the microscopic perspective of statistical mechanics. Leinonen *et al.* [15] studied this in the context of the second law of thermodynamics, looking at how consistently upper-division students use the law at macroscopic and microscopic levels. One of their conclusions was that the macroscopic perspective of entropy is easier to understand than the statistical nature of entropy from the microscopic approach. Another study by Meltzer [8] observed learning patterns in an upper-division thermal physics course and found that several learning difficulties common among introductory students persisted with the upper-level students. Students' struggles with concepts in thermodynamics might carry on to the statistical mechanics context, where the descriptions of the concepts are perhaps even more difficult to grasp. Erceg *et al.* [16] investigated students' understanding of concepts related to the microscopic model of gas, as a part of developing the kinetic molecular theory of gases (KMTG) concept inventory. One of their findings indicates that similar student challenges appeared across subgroups from different teaching environments, such as different study programs, years of study, and curricula with a varying focus on thermodynamics.

The upper-division studies mentioned above have considered statistical mechanics content to varying degrees. Next, we provide an overview of studies that mainly focus on statistical mechanics, which are even rarer. Binary systems are a typical way to introduce some key concepts of statistical mechanics. This was considered by Mountcastle *et al.* [9] who found that students did not reliably reason that the relative uncertainties of binary outcomes decrease as the number of measurements increases. Loverude [12] also investigated student understanding of probability concepts in the context of simple systems. He found that students struggled to distinguish microstates from macrostates and apply mathematical relationships for the multiplicity. The same trend was later identified by Loverude in the context of two interacting Einstein solids [13], which was part of a series of tutorials developed in the study. In the context of a less simple system, Crossette *et al.* [17] also found confusion about macrostates and microstates among some of the interviewed graduate students.

Other studies have also focused on developing tutorial materials and assessment tools for teaching statistical mechanics. Recently, Rainey *et al.* [18,19] developed and validated the first upper-division thermal physics assessment that addresses both classical thermodynamics and statistical mechanics content, the U-STEP. Smith *et al.* developed tutorials and investigated student reasoning about the Boltzmann factor, an essential concept of statistical mechanics. This was considered in the context of students' understanding of Taylor series expansions [11], comparing relative probabilities of states [6], and its relation to the density of states function as expressions of multiplicity [10]. Over several years, these authors have consistently found that students often fail to recognize the situations in which it is justified to apply the Boltzmann factor and struggle to express its physical significance, after lecture instruction alone. However, their results indicate that after participating in the developed tutorial instruction [6], students were more likely to use the Boltzmann factor appropriately.

Entropy has been explored in several studies. As pointed out by Crossette *et al.* [17], entropy is a core concept of thermal physics and despite its precise statistical definition, it is subtle and difficult to understand. Loverude [14] and Crossette *et al.* [17] identified several resources that upper-division and graduate students used to conceptually describe entropy. These include relating entropy to disorder (a common connection for undergraduate students identified in Ref. [14]), temperature, mixing, or information. Relating entropy to the number of microstates, as stated in the statistical definition, has been reported as strongly preferred by students [5,17]. Language aspects and metaphors for entropy have also been considered from a thermodynamics perspective, for instance in Haglund *et al.* [20] and Haglund [21].

In his Ph.D. dissertation from 2022, Lo [22] presented findings from interviews with both undergraduate and graduate students. These interviews aimed to probe the students' understanding of central concepts from statistical mechanics such as basic probability, microstates, and macrostates, Einstein solids, the Boltzmann factor, the density of states, and the partition function. Lo concluded that the students generally could recognize the concepts, but only expressed a superficial understanding of them, and that their reasoning consisted of incomplete and/or loosely connected ideas rather than stable misunderstandings.

These existing studies on upper-division thermal physics agree that the research is limited in this area and that it is meaningful to expand it. Particularly, research on student learning in statistical mechanics, which is a challenging subject to learn due to its complexity and subtlety. Additionally, it relies heavily on mathematics. Such studies are also motivated by the general need for expanding the research on upper-division physics courses [23] and increasing the emphasis on epistemological skills [7] in teaching advanced courses. Gaining more insight into the thought process and reasoning of students, such as identifying challenging areas, is important for improving teaching and developing curriculum materials, to better guide students on their path to becoming expert physicists.

Most of these current studies within statistical mechanics have considered data collection contexts such as individual student interviews, written answers to questions, and questionnaires, but rarely problem-solving groups [24]. Since student reasoning in statistical mechanics remains a relatively uncharted topic in PER, especially in collaborative settings, we aimed to explore it further. We chose to do so without limiting ourselves to specific research questions or theories beforehand. Inspired by a grounded approach, we instead ask the broad question: *What challenges do upper-division physics students face when solving statistical mechanics problems in groups?* Note that we reported on findings from our preceding pilot study in Ref. [25] and presented a selection of those initial findings at the GIREP 2022 conference [26]. After additional data collection and several iterations of analysis, we formulated the final categories of challenges that emerged from the data. In the final part of the paper, we discuss our study, how our findings relate to previous research, and how it serves as a departure point for formulating research questions for future research.

II. PHYSICS BACKGROUND

Statistical physics deals with finding a description of the physical properties of large, macroscopic systems, typically containing a number of particles in the order of magnitude of Avogadro's number, 10^{23}. Describing such systems from a microscopic perspective essentially means solving the equations of motion for all components, if it is a classical system. This is practically impossible and not preferable. Moreover, for a quantum system, the wave function for the entire system would need to be specified and evolved in time using the appropriate Hamiltonian, which is even more intimidating. Two different approaches have been explored historically to deal with such systems in a feasible way. First came the development of classical thermodynamics in the first half of the 19th century. It was centered on the laws of thermodynamics, which are based on and justified by experiments on macroscopic systems. The second approach started with Maxwell's kinetic theory of gases, which instead aimed to derive the macroscopic laws and all macroscopic properties of a system from its microscopic properties. It is referred to as statistical mechanics. Next, we provide an overview of some relevant fundamental concepts of statistical mechanics, based on Refs. [1,27].

A. Macrostates and microstates, entropy, and temperature

Generally, a macrostate is a state of a system specified by macroscopic observables. For instance, given an isolated system in equilibrium, the energy E, the volume V, and the number of particles N are fixed, then (E, V, N) specifies a macrostate of the system in equilibrium. A system in nonequilibrium requires further specification of additional macroscopic quantities, labeled α, which essentially depend on what can be observed. Macrostates are then specified by (E, V, N, α) where α can refer to many additional variables $(\alpha_1, \alpha_2, \ldots)$.

Microstates are instead states of a system defined by complete microscopic descriptions. This can, for example, correspond to specifying positions and velocities of all its components. Specifying a microstate completely for large systems is usually unreasonable. Determining the number of microstates for a particular macrostate, in other words, the number of microstates corresponding to the same E, V, N, α, is, however, often sufficient. This is defined as the statistical weight (or multiplicity) of a macrostate, often denoted $\Omega(E, V, N, \alpha)$. In classical mechanics, the microstates form a continuum and "counting" them generally involves integrations over phase space. However, in quantum mechanics, the microstates form a discrete set for a finite system, which means that they can be counted and summed over explicitly [28].

A fundamental assumption in statistical mechanics is that for an isolated system, in a certain macrostate, all microstates corresponding to that macrostate are equally probable. This is the postulate of equal probability. The equilibrium postulate further states that for an isolated system, described by some fixed macroscopic parameters (E, V, N, etc.) and some variable parameters α, at equilibrium α will take values that maximize the statistical weight. Thus the equilibrium state is the most probable since it has the most number of microstates. Thermodynamics gives another perspective, that the equilibrium state corresponds to the maximum value of entropy. From this, a "bridge equation" was proposed, connecting the macroscopic quantity entropy of a system to its microscopic description. The Boltzmann formulation of this fundamental relation is given by

$$S = k_B \ln(\Omega), \tag{1}$$

where S is the entropy of the system in a particular macrostate with statistical weight Ω and k_B is Boltzmann's constant. There are many ways to conceptualize entropy. One example is as a measure of the amount of information that is lost by specifying the state of the system using macroscopic observables.

Temperature is a challenging concept in many contexts and it has, along with the term "heat," been the focus of many studies across the disciplines of science, technology, engineering, and mathematics [2]. We will briefly summarize how temperature can be described in physics literature and how entropy is related to temperature. A common, simple way to think about temperature is through its operational definition; temperature is what you measure with a thermometer. A more fundamental way to define temperature is two objects in thermal contact have the same temperature if there is no spontaneous net transfer of energy between them, that is, they are in thermal equilibrium. However, the statistical approach yields a very useful theoretical definition of temperature:

$$\frac{1}{T} = \left(\frac{\partial S}{\partial U}\right)_{N,V}. \tag{2}$$

Here temperature is the inverse of the partial derivative of the entropy with respect to the energy of a system, taken with the system's volume and number of particles held fixed. An illustrative example for deriving this result is two weakly coupled Einstein solids, for instance, Ref. [1]. It also provides a way to understand the relationship between energy, temperature, and entropy. If the first object has a steeper S vs U graph than the second object, it will gain energy such that its entropy is increased, while the second object loses some energy and its entropy is slightly decreased. The total entropy increases, so the process happens spontaneously. We relate the first object to a lower temperature (the steeper slope) and the second object to a higher temperature (the shallower slope) since energy is transferred from an object with a higher temperature to an object with a lower temperature.

B. The Boltzmann distribution

Here we briefly describe the context and most relevant properties associated with the Boltzmann distribution, sometimes called the canonical distribution. It can be considered "*the most useful formula in all of statistical mechanics*" (p. 223) [1]. A complete derivation can be found in most textbooks on statistical mechanics, such as [1,27,29,30].

Consider a system in thermal equilibrium with a much larger system, a "heat bath" or "heat reservoir," with a fixed temperature T. The systems can only exchange energy in heating processes. Now consider the states $|i\rangle$ of the small system, with energies E_i. The Boltzmann distribution gives the probability that the system is in a particular energy state $|i\rangle$:

$$p_i = \frac{1}{Z} e^{-\beta E_i}, \tag{3}$$

where $\beta = 1/k_B T$ and Z is a normalizing factor. The term $e^{-\beta E_i}$ is the relative probability, called the Boltzmann factor. Z is referred to as the *partition function* and it is given by the sum of Boltzmann factors over all states:

$$Z = \sum_i e^{-\beta E_i} = \sum_{E_i} g(E_i) e^{-\beta E_i}, \tag{4}$$

to ensure that $\sum_i p_i = 1$. Equivalently, it is given by a sum over all energy levels with corresponding degeneracy $g(E_i)$. The partition function is very useful for calculating various properties of the system. For example, the average energy of the system is given by

$$\langle E \rangle = \frac{1}{Z} \sum_i E_i e^{-\beta E_i} = -\frac{\partial}{\partial \beta} \ln Z. \tag{5}$$

Note that if one considers many copies of the small system, all weakly interacting by energy exchange only, then for one of these copies the rest of them act as a heat bath. This is called a canonical ensemble, and from it, many useful results can be derived in a similar way [27].

Certain systems are commonly used as examples in teaching statistical mechanics, such as the fairly realistic and mathematically simple Einstein solid model. One example of displaying the utility of the Boltzmann distribution is a system of a one-dimensional oscillator in thermal contact with a large Einstein solid [1], acting as the heat bath. Another common example is the two-state paramagnet, which also allows simple calculations. As Moore and Schroeder [31] pointed out, it can be a beneficial pedagogical example that pushes students toward a more general understanding of temperature. The two-state paramagnet turned out to be relevant to our findings, which we comment on in Secs. III B and IV. In the following section, we comment on the notion of distinguishable and indistinguishable particles, an important and subtle aspect of many tasks that students face in statistical mechanics.

C. Distinguishing between particles

Statistical mechanics relies on probabilistic arguments. Therefore, it is important to clarify whether the studied particles are assumed to be distinguishable or not. Counting microstates of a system in a certain macrostate will not be the equivalent for distinguishable and indistinguishable particles [27]. In classical physics, identical particles can sometimes be considered distinguishable because they can, in principle, be labeled and their trajectories followed separately at each point in time. In quantum mechanics, it is not possible to label identical particles, nor track them simultaneously in space and time, thus they are truly indistinguishable [32]. Exchanging two such particles will not yield a new state of the system since it looks identical to the first and there is no measurable difference.

However, in statistical mechanics, it is common to introduce a way to distinguish particles within certain models, particularly when modeling solid-state systems. Identical particles can be assumed to be fixed, apart from small vibrations, in lattice sites which allows the sites to be labeled. Thus, the particles are considered distinguishable. It is important to note that the distinctness comes from the labels of the lattice sites, which are separated in space, and not the particles themselves (which still are identical and indistinguishable when unconfined) [27].

III. METHODOLOGY

The overall design and methods of this study are based on a grounded approach, similar to the starting points of, e.g., grounded theory [33], phenomenography [34], and thematic analysis [35]. A grounded approach is suitable for our study since it concerns exploring something novel and complex; the still rather scarcely studied topic of student reasoning in statistical mechanics. We ask the broad question "What is going on here?", with respect to the challenges that students might face when solving statistical mechanics problems in groups. By avoiding asking specific research questions *a priori*, we reduce the risk of asking unfruitful questions. Both the findings and the methodological strategies emerged from the empirical data.

This required a flexible study design, and the analysis was developed by conceptualizing the data rather than imposing a predetermined theoretical framework on them. Note that this study has not attempted to theorize but could inform such projects in the future. The data analysis process is described in Sec. III C.

We acknowledge the influence that our perspectives as researchers had on the research process and outcome. Both conscious and unconscious. Such as our constructivist epistemological views and physics disciplinary perspectives. Awareness of the prior studies in this area also influenced the design and possibly the analysis of our study, despite efforts to observe the data through an open and broad lens. The aim of this exploratory study was to characterize challenging aspects for the students, through categories that emerged from the data. The focus was to find areas of student thinking and reasoning patterns that would be interesting to explore further.

We begin by describing the research context in terms of the data collection, ethical considerations, trustworthiness, and interview participants. Next, we summarize the interview questions, comment on the design of the problems, describe how they were distributed to the student groups, and provide suggestions for how each problem can be solved. Finally, we describe our process of data analysis and how it evolved during the study.

A. Data collection and interview participants

We collected data by video and audio recording groups of two to three students as they were given a set of problems to solve. In the following section, we describe how we designed these problems and how they can be solved. The participants were student volunteers from a statistical mechanics course at a large Swedish university, mainly third (final) year students from the bachelor's program in physics and a few engineering students. We collected data in two phases, during two subsequent course rounds. The syllabus was the same for both phases, but the course lecturer had changed. The course follows the book *Statistical Physics* by Mandl [27] and consists of traditional lectures as well as separate tutorial sessions. The study was disconnected from the course, both in the sense that recording sessions took place outside the regular schedule and that participating had no effect on students' course grades. All groups were interviewed after the key topics, those relevant to the interview questions, had been covered in the lectures. However, we note that since the sessions were spread out across a time interval during the course, the last groups had more time to process the content and attend more lectures. The sessions in the first phase were held from mid-November to mid-December, while the sessions in the following year were held early to mid-December.

The student volunteers were given an information sheet, stating, for example, that their identity would be protected and that they can withdraw their consent at any time.

TABLE I. Groups and student pseudonyms (phase 1). Note that the superscript "a" indicates transcripts translated from Swedish.

Group A	Group B	Group C	Group D[a]
Alex	Ben	Cole	Dana
Alice	Bill	Clara	Derek
Adam	Beth	Chris	Dylan

All participants signed a consent form prior to the interviews. The students received a cinema ticket as compensation for participating. The consent form, data processing, and data storage were conducted in accordance with the General Data Protection Regulation [(GDPR) (EU) 2016/679]. Ethical approval from the Swedish Ethical Review Authority was not required since no sensitive personal data were collected.

The student participants were divided into groups of two to three peers. Groups and corresponding student pseudonyms, to protect the students' anonymity, are shown in Tables I and II. All sessions lasted approximately 1.5 h (with a break in the middle for the groups in phase 1) in a room where the students had access to the problem sheets, pens, and a whiteboard. Before recording started, the students were encouraged to express their thoughts verbally, in a collaborative think-aloud manner, during the upcoming session and to utilize the whiteboard if they wished. They were told that the most important point is to express their reasoning, not to present a certain answer. Moreover, they are free to ask questions at any point before, during, and after the session. The role of the researcher (E. K.) was to facilitate the session, observe, and take notes. Care was taken to observe as unobtrusively as possible, while also moderating the session by asking occasional follow-up questions, e.g., if the students got stuck or encouraging the students to elaborate if the line of reasoning was unclear.

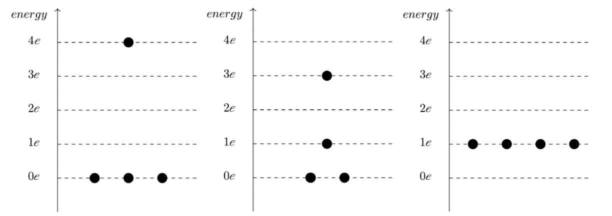
We chose to interview the students in groups for several reasons. One advantage is that students can verbalize their thoughts in a more natural way in groups with peers compared to individual, one-on-one interviews with a researcher. In group interviews, students can articulate their reasoning and keep their focus on the task at the same time, rather than explaining their thoughts to an interviewer (which can influence their thinking significantly) [36]. A successful group interview can create a safer

TABLE II. Groups and student pseudonyms (phase 2). Note that the superscript "a" indicates transcripts translated from Swedish.

Group E[a]	Group F	Group G[a]	Group H	Group I[a]
Eric	Fran	Gabe	Hans	Ike
Elias	Filip	Grace	Henry	Isak
Ethan	Fiona	⋯	⋯	Ivan

Consider 4 identical atoms in a solid, they can vibrate around their fixed points but not move around (**distinguishable**). The energy levels are 0, e, $2e$, $3e$ ... and the **total energy of the system is $4e$**.

(a) Rank the probabilities of finding the system in each of the configurations below.



(b) Are the depicted configurations macrostates or microstates of the system?
(c) Based on your answer to (b), would you change your ranking in (a)?
(d) What would the "entropy" of the system mean in this case, and how does it relate to the configurations?
(e) Would your answer change if the identical atoms were indistinguishable?

FIG. 1. The four atoms problem, is the first problem given to student groups in phase 1. The subquestions (b)–(e) were hidden from the students until they finished their discussion of part (a), hoping to observe their initial approach to (a) without steering them in a certain direction.

environment for students to express their thoughts, despite being recorded in an *in vitro* setting, and reduces the risk of the students feeling interrogated. Students tend to ask each other to explain in other words or elaborate their reasoning, which also serves as a valuable form of internal member checking among the participants. Note that the benefits that group interviews have for research data collection can also be seen as increasing the possibilities of learning [37]. This was also an important consideration for organizing the activities in the form of collaborative group work. Moreover, we note that in our case, the groups consisted of participants who knew each other. Additionally, the groups with exclusively Swedish-speaking students were given the opportunity to speak in English or Swedish, depending on what they felt more comfortable with. See table notes at the end of the caption of Tables I and II. Some of these groups chose English since the course was given in English.

B. Interview questions

Student groups from the first phase were initially asked to "*Explain the meaning of macrostates and microstates of a system.*", an explicit prompt to probe their general ideas about it. Second, they received what we will call the "four atoms" problem, see Fig. 1. This problem was inspired by tutorial material from the course, as well as the "strings question" in Crossette *et al.* [17]. Third, the groups were given the "three-state system" problem, see Fig. 2. This task was taken from a previous exam in the course, but we added a question in part (b): "*What is the difference between this situation and the one in (a)?*" and a prompt to justify their method. This problem was meant to probe the students' ability to identify under which circumstances the Boltzmann distribution is applicable and observe how they utilize it. Finally, they were explicitly asked to reflect

Consider a system of 2 non-interacting lattice sites, each of which can be in one of three states with energy $\epsilon = +1, 0, -1$ respectively.

(a) Assuming that each state can be one of the three states with equal probability, compute the average values of U and U^2 (where U is the total energy of the system).

(b) Let the system be in contact with a heat bath at temperature T. What is the difference between this situation and the one in (a)? Calculate the average values of U and U^2 and justify your method. Comment on low and high temperature limits.

FIG. 2. The three-state system problem, the second problem given to student groups in both phase 1 and phase 2.

on it: "*Explain your understanding of the Boltzmann distribution (where does it come from, why is it useful, and in which situations can it be applied?)*".

In phase 2, the student groups were first given one version of the "small paramagnet" problem, see Fig. 4, followed by the other version. Groups F and H were first given the version with Fig. 4(a), after completing the task, they received the version with Fig. 4(b). Groups E, G, and I obtained the versions in the opposite order. This problem was adapted from item 4 of the U-STEP [19], but we removed the multiple-choice options to promote less confined discussions. In their validation study, Rainey *et al.* [19] reported that many students chose the incorrect answer $P = 1/4$ for question (ii) when Fig. 4(a) was used but did so significantly less with Fig. 4(b). They suspected that the distractor was the length of the longest column in Fig. 4(a), with four microstates, which led to the conclusion that $P = 4/16 = 1/4$. Figure 4(b) was believed to lead students to rely less on the figure and more on physical intuition or direct state counting. We chose to let our student groups discuss both versions, to gain deeper insight into the reasoning process, and how they interpret the two different representations. The final problem given to the groups in phase 2 was the three-state system problem. Exactly the same as during the first phase, see Fig. 2. Many interesting themes emerged from the initial analysis (in phase 1) of the

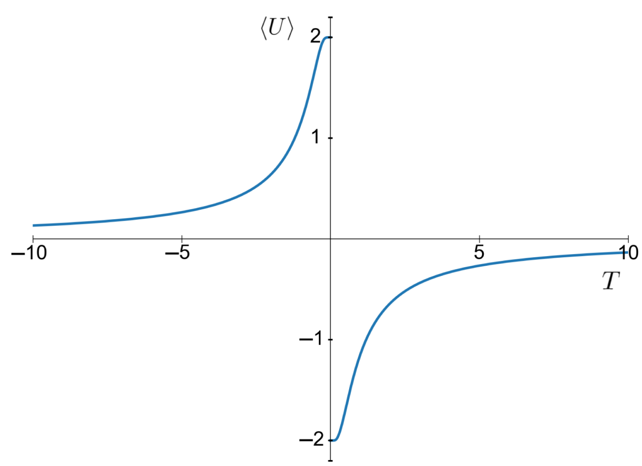
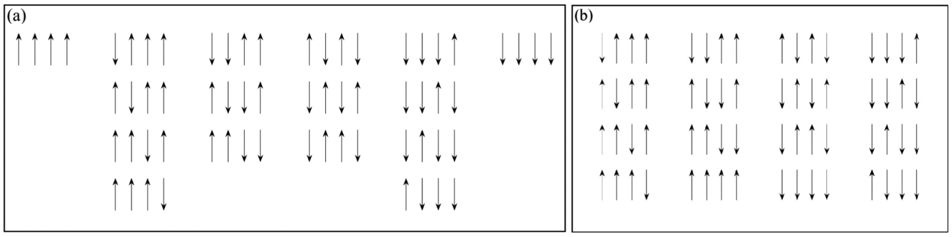


FIG. 3. Graphical visualization of a part of the solution to question (b) in the three-state system problem. Average total energy U as a function of temperature T, according to Eq. (9) for $k_B = 1$.

Consider a paramagnet consisting of $N = 4$ dipoles. Each dipole can be in one of two states: $\uparrow$ and $\downarrow$. All possible orientations of the dipoles are shown below.



(i) Does this system have more microstates or more macrostates? Explain your reasoning.

(ii) What is the probability of finding the system in the most probable macrostate?

FIG. 4. The small paramagnet problem, the first problem provided to student groups in phase 2, adapted from the two versions of item 4 of the U-STEP, see Rainey *et al.* [19]. The problem was given either with (a) or (b), but here we illustrate them side by side for convenience.

students' discussions of this problem, and so we decided to collect more data for this particular context. At the end of all the interview sessions, the students were asked to reflect on the tasks as well as the course content in general. Such as, what they found challenging or confusing.

Next, we provide suggestions for how the problems can be solved, along with further comments on the design choices and subtleties of the tasks.

1. The four atoms problem

This problem was meant to probe the students' ability to define and distinguish between macrostates and microstates. Here the context is a small system with discrete, unbounded energy levels, where the total energy of the system is kept fixed. How they relate these concepts to entropy was also considered in this question. The main point was, however, to explore possible challenges that may arise due to the ambiguous and conceptually subtle definitions of macrostates and microstates. The representation of the configurations in Fig. 1 was intentionally vague. While the question states that the particles are distinguishable, there are no labels on the dots in the illustrations. Thus, these depictions do not fully specify the system. A microstate would specify the energy states of all four particles, which can be distinguished due to their confinement in a lattice (this was discussed in Sec. II C).

However, the concept of macrostates is neither obvious nor unambiguous here. The total energy of the system is fixed, and other macroscopic quantities are implied to be constant. Along with these observables, the macrostates could be defined by some variable parameters, denoted as α in Sec. II A. Let n_i be the number of particles in energy level $i = 0, 1, 2, 3, 4$. Then, for this particular context, we could, in principle, choose the set $\{n_i\}$ to specify a macrostate, along with the total energy $4e$. The configuration on the left in Fig. 1 would then correspond to a macrostate with $\{n_0 = 3, n_1 = 0, n_2 = 0, n_3 = 0, n_4 = 1\}$. The one on the right would be $\{n_0 = 0, n_1 = 4, n_2 = 0, n_3 = 0, n_4 = 0\}$. Thus, the different macrostates would be given by the possible ways of partitioning $4e$, which includes two additional configurations to the three shown in Fig. 1. It would then make sense to calculate the statistical weight Ω of each macrostate, the corresponding number of microstates, and rank the probabilities accordingly. Since the total energy is fixed, the most probable macrostate is the one with the largest number of microstates (postulate of equal probability). With these assumptions and definitions in mind, the correct ranking from most probable to least probable is the middle configuration (with $\Omega = \binom{4}{1} \cdot \binom{3}{2} = 12$ number of microstates), the left configuration (with $\Omega = \binom{4}{1} = 4$), and finally, the right configuration (with $\Omega = \binom{4}{4} = 1$).

Whether this definition of the macrostates is reasonable from a realistic perspective is, however, debatable. One could argue that the set of numbers $\{n_i\}$ would not be observable. On the other hand, this really is a "toy problem" with a very small and idealized system. Therefore, it might be beside the point to discuss the experimental, real-life aspects. An alternative approach to the question is to argue that the depicted configurations do not show specific microstates, nor are they to be considered macrostates, but something in between [38].

2. The three-state system problem

It is important to note that there are several implied assumptions in the context of the three-state system problem, see Fig. 2. Once again, this is an idealized "toy model" which allows for simple calculations.

However, it is not necessarily trivial to identify its subtle aspects nor to reason about the system conceptually. The system is obviously very small, with only two components, and it is quantum mechanical—only three possible energy states. Nothing is stated regarding the nature of the components other than that they are confined to a lattice, and that interactions between the sites are negligible. It is therefore implied that the particles, or whatever is confined to the lattice, are considered distinguishable (see our discussion in Sec. II C). Moreover, another unstated aspect is that any vibrational motion of the particles is neglected. Only the interaction corresponding to those three particular energy states is considered, consequently, the energy states are not only discrete but also bounded. This is typically not what we are used to experiencing in real life, where for instance the kinetic energy is unbounded. Many physics contexts also deal with unbounded energy states. Even in quantum systems such as harmonic oscillators, which have discrete but unbounded energy levels.

With all these idealizations in mind, the answer to (a) can be calculated using basic probability principles since the sites are independent and all energy states are assumed to be equally probable. The average values of U and U^2 are then given by the following:

$$\langle U\rangle = \langle U_1\rangle + \langle U_2\rangle = \frac{2}{3}(-1+0+1) = 0, \tag{6}$$

$$\begin{aligned}\langle U^2\rangle &= \langle (U_1)^2\rangle + 2\langle U_1\rangle\langle U_2\rangle + \langle (U_2)^2\rangle \\ &= \frac{2}{3}(1+0+1) = \frac{4}{3}.\end{aligned} \tag{7}$$

In part (b), the system is in contact with a heat bath and is assumed to be in thermal equilibrium with it, at temperature T. It is implied that the number of particles in our system remains constant and that no other relevant macroscopic properties change. Only thermal energy can be exchanged between the system and the heat bath. The probability for a site to be in one of the three energy states is then given by the Boltzmann distribution since the context has the required conditions (see Sec. II B). The probabilities will now depend on the temperature T, and they are not the same for each energy state anymore (except in the high-temperature limit, as we will see). The lattice sites are noninteracting and distinguishable. Thus the total partition function of the system is $Z = (Z_1)^2$ where Z_1 is the partition function for one of the sites. This is by definition, see Eq. (4), in this case given by

$$Z_1 = \sum_i e^{-\beta\epsilon_i} = e^{\beta} + 1 + e^{-\beta}, \tag{8}$$

where $\beta = 1/k_B T$. The average total energy can then be calculated according to Eq. (5), given $Z = (Z_1)^2$, and one finds that

$$\begin{aligned}\langle U\rangle &= -2\cdot\frac{\partial}{\partial\beta}\ln Z_1 = 2\cdot\frac{1}{e^{\beta}+1+e^{-\beta}}(-e^{\beta}+e^{-\beta}) \\ &= -4\cdot\frac{\sinh\beta}{1+2\cosh\beta}.\end{aligned} \tag{9}$$

The average value of U^2 can be calculated using the following:

$$\begin{aligned}\langle U^2\rangle &= \frac{1}{Z}\sum_i \epsilon_i^2 e^{-\beta\epsilon_i} = \frac{1}{Z}\frac{\partial^2}{\partial\beta^2}Z = \cdots \\ &= \frac{4\cosh\beta + 8\cosh 2\beta}{(1+2\cosh\beta)^2}.\end{aligned} \tag{10}$$

The solution to part (b) is thus given by Eqs. (9) and (10). We determine the high- and low-temperature behavior by taking the limits as follows. In the low-temperature limit, where $\beta \to \infty$, the system is frozen in the ground state. That is, $\langle U\rangle \to -2$ and $\langle U^2\rangle \to 4$. In the high-temperature limit, where $\beta \to 0$, all energy states become equally probable and the results match the answers in (a). Namely, $\langle U\rangle \to 0$ and $\langle U^2\rangle \to \frac{4}{3}$. The temperature dependence of $\langle U\rangle$ is illustrated in Fig. 3.

Discussing the high-temperature limit conceptually is not as straightforward as the calculation. As mentioned previously, several assumptions are made about the system and there is a subtlety in treating the temperature limits. The particles are assumed to have no vibrational energy in this model, which illustrates the need for more sophisticated reasoning about temperature using the theoretical definition in Eq. (2) in this context. That is, relying on the more fundamental quantity of entropy since temperature is just a characterization of a system's "willingness" to give up energy [1]. From a physics perspective, this three-state system is isomorphic to the ideal two-state paramagnet (mentioned in Sec. II B). If we give the three-state system problem more context, it could correspond to a small, ideal paramagnet of two spin-1 particles, instead of spin-1/2, in an external magnetic field. The assumptions, calculations, and unusual behavior of these two systems are essentially the same. For a detailed treatment of the two-state paramagnet, we refer the reader to literature on the topic, such as Refs. [1,27,31].

3. The small paramagnet problem

The small paramagnet problem, see Fig. 4, is also a version of the two-state paramagnet. There is an important distinction and there is no external magnetic field in this situation. We also point out the following implied assumptions. That the dipoles are noninteracting, dipole-dipole interactions are not considered, and the dipoles are distinguishable since they are spatially confined. Considering all these conditions, there is no specific energy associated with the states ↑ and ↓. Therefore, all of the 16 microstates depicted in the figures are equally probable. We recognize from the context that the quantity of interest here is instead the total magnetic moment, or magnetization, which is what we define our macrostates according to. That is, we have five different macrostates characterized by the number of dipoles "up" (or equivalently "down") of the four dipoles in our system; $(4\uparrow - 0\downarrow)$, $(3\uparrow - 1\downarrow)$, $(2\uparrow - 2\downarrow)$, $(1\uparrow - 3\downarrow)$, and $(0\uparrow - 4\downarrow)$. There are 16 microstates and 5 macrostates, so the answer to part (i) is that the system has more microstates than macrostates.

Moreover, since we assume that all the microstates have the same energy and that the system is isolated, the microstates are equally probable. To calculate the probability of finding the system in a certain macrostate, part (ii), we can simply count the number of microstates corresponding to that macrostate and divide by the total number of microstates. The most probable macrostate is two dipoles in the "up" state and two in the "down" state, $(2\uparrow - 2\downarrow)$. There are six microstates of that macrostate so the probability is $P = 6/16 = 3/8$. We note that the representation of the microstates in version (a), see Fig. 4, almost organizes the macrostates in columns (except that the $(2\uparrow - 2\downarrow)$ has been split into two columns). As mentioned previously, the designers of the U-STEP [19] suspected that the length of the second and fifth columns were distractors, so they made the representation in (b) without an obvious grouping of microstates in rows or columns.

C. Data analysis

We conducted the qualitative data analysis through a grounded approach, in line with the study's overarching methodology. In essence, the starting point is an open coding of broad initial data, looking for initial themes or ideas. The coding is an interactive and comparative process, and it is iterative. It starts while gathering data and continuously emerges as the researcher engages with the data. This open coding phase corresponds to our first phase. We took notes during the interview sessions. Afterward, we looked through all the collected video data (Group A–D) and made notes of patterns and themes of student challenges on a macrolevel. Note that we were interested in overall challenges. We did not analyze specific questions (if solved correctly or not) separately or consider detailed identification of activated resources. Nor did we consider any quantitative analysis. We inductively generated initial codes from the data, and some relevant excerpts were transcribed to allow line-by-line coding. We discussed these initial codes among the authors, and we established and summarized the most frequent and/or significant ones. At this stage, we discussed the physics content of the emerging themes with a previous lecturer of the statistical mechanics course.

In the next stage, we used these initial codes to analyze the new data gathered in phase 2, both during and after the interview sessions. We constantly compared and scrutinized these codes to evaluate their significance, and we developed them to generate tentative categories. After establishing the tentative categories, we analyzed the video data from both phases again, making detailed notes that were coded based on the categories. Again, we transcribed relevant excerpts. In the end, we established ten categories of recurring challenges, which we later grouped into two broad themes. We discussed this among the authors and with colleagues in our research group. We summarized the categories, with a broad perspective of all the data, and we chose representative quotes to illustrate them (see Sec. IV). We note that while this exploratory study stopped here, future studies could continue to refine the abstraction and search for relationships between categories. In this phase, a relevant and appropriate external theory can act as a supplement to aid the analysis. In our thorough engagement with the data, we noticed that students' ideas and reasoning patterns seem to be context dependent. Individual students also shift between ideas frequently. We would, therefore, consider a resource, or knowledge-in-pieces, perspective [40] as a theoretical supplement in future work for a more focused analysis of significant episodes.

IV. EMPIRICAL FINDINGS

This study was initiated with the aim of exploring challenges that upper-division physics students face when solving statistical mechanics problems in groups. Here we present our empirical findings from two phases, based on video data of nine groups of two to three students. Once more we point out that the categories of challenges emerged from the data, without external theoretical frameworks or hypotheses in mind. We make no claim that the categories are extensive or generalizable. During the iterative analysis process, ten distinct challenges emerged. These were finally divided into two groupings: (A) *challenges with concepts* and (B) *challenges with problem-solving strategies*, including seven and three of the categories each. We present an overview of the final versions of the emerged categories in Table III. Note that these emerged through our prolonged engagement and iterative analysis of all the data. In the following, we summarize all the categories and provide transcribed excerpts as examples to illustrate what led to each category. How our findings relate to previous research, and suitable theoretical perspectives for further studies, is discussed in Sec. V.

A. Challenges with concepts

The first grouping of identified challenges concerns the concepts, and sometimes the terminology itself, of statistical mechanics. It is a broad and overarching grouping that includes seven categories, each of which involves common observed challenges with the concepts: (1) *macrostates and microstates*, (2) *entropy and temperature*, (3) *distinguishable and indistinguishable*, (4) *equilibrium*, (5) *energy*, (6) *temperature and heat bath*, and (7) *the Boltzmann distribution and partition function*.

TABLE III. Overview of the ten emerged categories of challenges that the students faced when solving statistical mechanics problems in groups. The categories are divided into two broad groupings, *(A) challenges with concepts* and *(B) challenges with problem-solving strategies*. Note that this table only summarizes the final version of the categories, and the main text demonstrates what led to each category through example excerpts from the data.

Emerged categories	Broad groupings
1. Struggling to define and disentangle macrostates and microstates and attributing statistical weight inappropriately	A. Challenges with concepts
2. Treating macroscopic quantities like entropy and temperature as microscopic properties of a system's components	
3. Confusing distinguishable and indistinguishable and not recognizing the cause nor the consequences of that distinction	
4. Associating an equilibrium state with increased order	
5. Struggling to recognize the consequences and causes of a system having discrete energy states when the states are bounded or the total energy is kept fixed	
6. Struggling to conceptualize temperature through the statistical definition and misinterpreting the meaning and effects of a heat bath	
7. Reasoning inconsistently about the Boltzmann distribution and the partition function, and their applicability	
1. Inappropriately applying typical mathematical relations from statistical mechanics and thermodynamics	B. Challenges with problem-solving strategies
2. Having difficulties managing conflicts between quantitative results and qualitative reasoning	
3. Defining new concepts that led to more confusion due to their inconsistency with established relations	

In general, the students struggled to navigate among the many different concepts, which were still rather new to them. The students seldom had a definite perception of key terms, and if they did, it was often incomplete or inappropriate in the considered context. These general challenges were also articulated by the students themselves, both spontaneously during the interview sessions and when asked to reflect on them in the end. They reflected on their confusion with the many different and abstract concepts. Sometimes unsure of the terms themselves, but more often the meaning of the concepts and how they are related. The students also emphasized that this was extra challenging due to the lack of focus on conceptual understanding in the course, which they felt was mainly focused on long derivations of mathematical expressions. These self-reported difficulties support our findings of observed challenges and show that the students were aware of them in many cases.

Next, we summarize each category of challenges involving concepts, which we refer to as A1–A7.

1. Struggling to define and disentangle macrostates and microstates and attributing statistical weight inappropriately

One of the most evident challenges that appeared across the student groups, and in several different contexts, involved the concepts of macrostates, microstates, and statistical weight or multiplicity. Our findings suggest that the students struggled with the ambiguous and context dependent definition of macrostates and microstates when discussing simple and discrete systems. This carried over to their treatment of the statistical weight. In general, the students were unsure how to define the macrostates appropriately in a particular context. Especially determining what scale is suitable to consider as macro or micro. Students repeatedly confused the macrostates or microstates of a system with the system, or parts of the system, itself. Similarly, they often attributed a statistical weight to the system in general, or to a microstate, rather than to a particular macrostate of the system. Similar patterns were identified in both phases, mainly in the students' discussions of the four atoms problem and the small paramagnet problem, respectively.

These challenges often led to disagreements within the groups, and the reasoning shifted back and forth throughout the tasks. Very few students had a precise idea about how to approach the definition of macrostates and microstates. The groups were notably torn between different conceptualizations. Group B, for instance, at one point stated that macrostates *is* the atom and the energy levels *are* the microstates, in the four atoms problem. Similarly, Ivan initially said that "*each particle is a microstate, and then we look at the whole unit [points to the square] as a macrostate*" for version (b) of the small paramagnet problem. Group I also discussed back and forth, similarly to group E, that the microstates are always attributed to single atoms or molecules. Such that there are 2 microstates (up or down) in this context and 16 macrostates (the depicted configurations).

Ivan: *But then I think like this: there are two microstates, we have up or down for each of these [points to an arrow in the figure] and there are 16 macrostates, because we can have 16 different configurations of up or down.*

These ideas led to confusion since it was inconsistent with for instance their ideas about entropy and that several microstates should be able to give rise to the same macrostate.

Another common pattern was many students' firm belief that macrostates must solely be defined by the total energy. For example, groups C and D wanted to reason that the depicted configurations in the four atoms problem represented microstates and that the macrostate was defined by the total energy $4e$. This led to confusion since they realized that the depictions did not specify the system fully since there were no labels on the atoms, so there might be more layers of states. We discuss this further in category B3. When turning to the discussions about statistical weight and entropy, they realized that they contradicted themselves by stating that statistical weight is associated with macrostates at the same time as attributing a statistical weight to the depicted configurations (that they believed to be microstates). For example, in group C:

Clara: *Can we just say that the [statistical] weight is the amount of degeneracies per microstate?*

Cole: *Yeah that's what I'm thinking. If we are talking about like statistical weight of a microstate.*

Which is problematic since it is inconsistent with the postulate of equal probability. This confusion was articulated by several of the groups. Group D was unsure how to talk about the issue:

Derek: *If that [middle figure] shows more than one microstate, can we describe it as a macrostate? No I don't think so.*

Dylan: *I still think that they show microstates. But then maybe in some way, like you said, that [right figure] is one microstate, here [middle figure] we also have microstates but we can, it is somehow several… in this figure.*

Derek: *[later] I don't want to call them microstates, because then it feels wrong to say that we have several microstates in a microstate. Because that doesn't make much sense.*

In the end, they held on to that macrostates should only be defined by quantities like energy, volume, etc. This was challenging for groups F, G, and H too. They were unsure how to deal with defining macrostates and microstates in the small paramagnet problem since there was no mention of energy. Some struggled to recognize that without an external magnetic field, there is no specific energy associated with spin up or down. The important macroscopic quantity that should define the macrostates is instead the magnetization.

2. Treating macroscopic quantities like entropy and temperature as microscopic properties of a system's components

In category A1, we described the common challenges that students faced with respect to macrostates and microstates. This is related to the relative scale that one considers, and the choices you make in defining them. Similarly, this category also deals with the distinction and connection between macroscopic and microscopic perspectives, as was concluded to be challenging for students in Ref. [15]. In this case, we primarily consider this challenge for the quantities of entropy and temperature. Entropy and temperature are both observed on a macroscopic level, but the concepts are constructed from a microscopic point of view in statistical mechanics. Our findings suggest that the students struggle with this distinction, and they sometimes refer to these concepts as microscopic properties of a component of a system or of the component's states.

Despite recognizing and articulating that these issues depend on the considered scale, group B still struggled with it. When talking about heat baths in the context of the three-state system problem, Beth reasoned about temperature as a microscopic property, attributing it to the molecules individually:

Beth: *Like if we put that bottle in a heat bath, eventually all of the molecules are gonna have the same temperature. So then they will all be in the same configuration?*

Shortly thereafter, Ben reasoned about entropy, and consequently, probability, as a microscopic property when attributing entropy to the different energy states ($\epsilon = +1, 0, -1$) of a single component:

Ben: *We would get more entropy in certain states, right? [Later] Like our three different states, say that* $+1$ *has higher entropy than the others, then it is more likely that the two non-interacting lattices would choose the* $+1$.

As mentioned in the previous category, Group C was confused about their own formulation of the "*statistical weight of a microstate*," which implies a similar microscopic perception of entropy via its statistical definition. Group H also briefly struggled with the appropriate scale for temperature. Stating that "*now we have T for each state [referring to the available energy states for one lattice* *site]*" after recognizing that all the states are equally probable in part (a) of the three-state system problem.

3. Confusing distinguishable and indistinguishable and not recognizing the cause nor the consequences of that distinction

The concept of distinguishable or indistinguishable particles was an apparent challenge for many of the groups. The terms were emphasized as very confusing by groups A, B, and D. All groups except G, H, and I struggled with the concept during their problem solving. Groups H and I never explicitly discussed the term distinguishable and appropriately made assumptions about it without articulating it. Among the rest of the groups, the students were often unsure whether the components of a system should be considered distinguishable or not. If it was stated in the problem they did not reliably recognize why. This challenge was manifested both in the students' direct interpretation of the questions, and when they counted microstates. Sometimes students did not agree within the group whether their calculations had assumed distinguishable or indistinguishable particles.

Groups A–D expressed confusion about the concept distinguishable throughout the interview sessions and particularly struggled with it during the four atoms problem. Groups A and B kept coming back to it during that problem. Asking themselves why the problem states that the atoms are distinguishable, not recognizing that the cause is the confinement to the lattice (see Sec. II C). They were also unsure why this distinction matters, along with group D, choosing to ignore the word when they could not resolve the issue. There was a common tendency to count extra microstates by swapping atoms "within an energy level" and justifying it by pointing out that the atoms are distinguishable. However, as we discussed previously, exchanging two atoms with the same energy would not yield a new microstate. The data suggest that the students might be interpreting the graph of the energy space as a literal representation of the system in space.

Group C articulated similar conclusions when discussing the three-state system problem:

Cole: *It doesn't matter if we think they are distinguishable or not. The question doesn't say if they are distinguishable or not. Shall we just say that they are distinguishable?*

Cole seems to make that assumption without particular reason, which in this case happened to be appropriate. Group E also had challenges in the context of the three-state system problem, going beyond the appropriate distinction of, e.g., the microstate $(\epsilon_1 = 1, \epsilon_2 = 0)$ from $(\epsilon_1 = 0, \epsilon_2 = 1)$ to something similar to group B's own term "super-distinguishable." After drawing a picture of the system as a box with two smaller boxes inside, with labels "1" and "2" corresponding to the two lattice sites, Ethan suggested an inappropriate distinction:

Ethan: *You can have (1 and 1) and (1 and 1). Do you understand how I'm thinking?*
Eric: *Yeah …*
Ethan: *I mean if, if we have an* ϵ_1 *and an* ϵ_2*. Then we have a combination that* ϵ_1 *is in the first box and has [the value] 1, and then a combination that* ϵ_1 *is in the second box and has [the value] 1 at the same time as …*
Eric: *Yeah sure, but we can count with duplicates.*
Ethan: *But then, but either we don't do it, and then we never do it.*
Eric: *But then we never do it, it should lead to the same …*

Other groups from the second phase struggled less with the term distinguishable, possibly due to receiving the small paramagnet problem instead of the four atoms problem. Group F had some challenges initially, during their discussion of the small paramagnet problem, but resolved it shortly thereafter.

4. Associating an equilibrium state with increased order

This trend was less common than the other challenges but still appeared across a few different groups in the discussions in different contexts. Other groups discussed the term *equilibrium* very little, and if they did, it was without notable difficulty. It was mainly groups A, C, and E (and partly group B) that associated an equilibrium state with increased order, seemingly confusing it with a ground state with low entropy. Equilibrium seemed to be associated with a state in which the energy is most spread out, evenly among particles, such that each particle has the same energy. This is problematic for nonzero energies since in equilibrium, the entropy should be maximized and the system tends toward the macrostate with the most number of microstates, not the least.

When group A discussed the four atoms problem, Adam was conflicted with their discussion about the entropy and probability of the configurations (that they said should be macrostates for the questions to make sense). Counting the possible microstates of each macrostate suggested another answer than his "gut-feeling" told him. He argued that the energy should be as spread out as possible in the macrostate when it is in equilibrium, suggesting that the configuration in the right figure (with all atoms in the $1e$ energy level) is more probable. "*That's my intuitive understanding of what equilibrium means*," Adam said. It is similar to the discussion of the same problem in group C. They reasoned that there should be some kind of probability weight on each energy state, even though the total energy is fixed, and therefore all atoms in the $1e$ energy level is the most probable configuration. Beth argued, mostly based on recollection, that the molecules of a system in equilibrium with a heat bath will eventually all have the same temperature (see category A2) and therefore maybe all will be in the same configuration.

Group E displayed the same challenge with reasoning about equilibrium in the three-state system problem. They used the zeroth law of thermodynamics to, inappropriately, conclude that the two lattices must be in the same energy state in equilibrium.

Eric: *If we say that* $dE = dQ$, *then both* e_1 *and* e_2 *have to be in equilibrium with each other if they are in equilibrium with* T_0 *[temperature of heat bath]. Isn't that the zeroth law of thermodynamics? Something like that. That if two systems are in equilibrium with a third, then the first and second are in equilibrium with each other.*
Ethan: *Yeah exactly, that's right.*
Eric: *So ehm, if we say that the system is in equilibrium, then we suddenly have a constraint that they should be equal. Yeah. So equilibrium [writes* $\epsilon_1 = \epsilon_2$ *next to "thermo 0" on the whiteboard]. And that means that, they can still be* 1, 0, *or* -1.
Elias: *But both have to be the same?*
Ethan: *Exactly, so we just have the combinations 11, 00, and* $-1 - 1$.

The zeroth law states that in equilibrium, the macrostate of a system remains the same over time. It does not put a constraint that properties must be the same on a microscopic level. This is another indication of students' struggle with scale, with respect to the concept of macro and micro.

5. Struggling to recognize the consequences and causes of a system having discrete energy states when the states are bounded or the total energy is kept fixed

This category of challenges is related to the behavior of the considered systems' energy states and total energy. The energy states are discrete in all the problems. For the four atoms problem, the levels themselves are unbounded but the total energy of the system is kept fixed. For the three-state system problem, the energy states are instead bounded. As we discussed in Sec. III B, energy is not relevant in the context of the small paramagnet problem since there is no external magnetic field, and the dipole-dipole interaction is assumed to be neglected. These properties of the systems are not made explicit in the problem statements, except that the total energy is kept fixed in the four atoms problem. The students struggled to recognize the reasons why the energy exhibited a particular behavior and the consequences of that.

Several groups from the first phase were conflicted about the consequences of the fixed total energy for the probabilities in the four atoms problem. As we mentioned in category A4, group C reasoned that the probability of the energy levels should be different, not recognizing how that is inappropriate in the context of keeping the total energy fixed. We note that one of the trends we presented in category A1, about students' belief that the energy *must* solely define the macrostates, is related to this category. For example, several groups in the second phase were confused about the energy in the small paramagnet problem. They seemed convinced that there must be an energy involved to solve the problem, failing to recognize that the context implies that the energy is irrelevant.

The main confusion about the discrete and bounded energy states of the three-state system became apparent in the discussions about the temperature limits. It seemed to arise from the difficulty of reasoning about what interaction the energy could be related to. This is something that is not explicit in the problem statement, and the problem is very much stripped from context, but the possible energy states are precisely defined (mathematically). When prompted to discuss what interaction the energy states could correspond to, many groups answered that it could be some kind of molecular vibrations or harmonic oscillators (which are generally unbounded). There was a disagreement in group I which highlighted that one reason for this confusion could be how the students interpret the problem statement. Ivan proposed that the system could be a harmonic oscillator, but Ike countered that it could not be because those have infinite energy levels. Ivan then said, "*But nothing says that this* $[\epsilon = -1, 0, 1]$ *is all the energy levels.*" In their reasoning about the three-state system problem, all groups either spontaneously discussed the problem in contrast to a two-state paramagnet or were prompted to when completely stuck. In some ways, this example was a very productive analogy for the students, a topic we will explore in future work, but very few groups recognized that the two problems are essentially isomorphic. Group E struggled to see the similarities between the energy states of the two-state paramagnet and the three-state system. Not realizing that some kind of interaction is behind the three discrete, bounded energy states (which could be the interaction between spin-1 particles and an external magnetic field):

Ethan: … *But are you sure that this [the three-state system] is subject to the same rules as the dipole?*
Eric: *No but it, I mean. So to speak, it is like they are just two lattices, there are no E-fields, no B-fields, nothing like that. Then it is that if we have a higher temperature, yes then we should get a higher energy, it it kind of like, that's just how it is.*
Ethan: *That is completely true!*

One of the few groups who struggled less with this was group I. Ike even recognized the isomorphism, but there were disagreements within the group. Ivan articulated his confusion about the energy states when comparing the two contexts:

Ivan: *I think that my confusion lies in that in this case I can't see a direct parallel to why they [the lattice sites] would have any kind of potential energy. Because in the first case [referring to the two-state paramagnet] we have something that has a potential energy because of something externally applied, we give them higher temperature and then look at what happens. Here we have two things with some energy, and then we give them more temperature, which in my mind is more energy, and then we should, in fact, look at the energy. But like I can't see any potential energy that is affected by the temperature.*

This quote also illustrates the next category of challenges, related to the concepts of temperature and heat bath.

6. Struggling to conceptualize temperature through the statistical definition and misinterpreting the meaning and effects of a heat bath

Almost none of the groups utilized the precise statistical definition of temperature, see Eq. (2), productively to reason about the interview problems. Note that the students had been introduced to it during the course. It was hardly considered overall, except by group A who initially struggled to use it both quantitatively and qualitatively in the three-state system problem. They treated temperature as discrete and tried to quantify their results, saying "*Should we do the different T?*", when discussing the temperature limits. One aspect of their confusion was the idea of negative temperature, which emerged from their calculations but they struggled to conceptualize.

Moreover, none of the groups productively conceptualized temperature in a more general way by considering the more fundamental concept of entropy. Moore and Schroeder [31] suggested that problems such as the two-state paramagnet could push students toward this. Our findings, particularly from the three-state system problem, suggest that this is challenging. We instead see trends to conceptualize temperature as a measure of energy, in particular, kinetic energy, which caused difficulties with reasoning about the temperature limits of the three-state system. As indicated in the quote of Ivan in the previous category, and later followed by Ike who said "*Because the kinetic energy… The temperature is related to the kinetic energy, but the kinetic energy is like, it is never the kinetic energy which has discrete states like that [referring to the three-state system], right?*". One exception was Alex from group A who demonstrated a more sophisticated understanding of temperature toward the end of the interview. He recognized that the statistical definition is more general and can be utilized in contexts that do not consider kinetic energy, such as the three-state system problem.

We also identified a strong and fast association (referred to as "intuition" by several of the students themselves) that the total energy should be maximized in the high-temperature limit. This trend was apparent in nearly all the groups, see an example from group D:

Dylan: *… the probability for each of those small energies will not be the same anymore, right?*
Derek: *When you change the temperature?*
Dylan: *I'm thinking that if we put it in a heat bath with very high energy, then it will like excite the small system, right?*
Derek: *Mm. Which would do what to the small system?*
Dylan: *To raise, to get this small system to a higher energy. I mean, to make [pauses, reads the questions] like that it will "pop up" [laughs], that it will increase the energy in our small system.*
Dylan: *[later] … maybe they can only be in these three states, but then maybe all of them will end up in the higher energy state.*

Discussions about the analogy of the two-state paramagnet, as mentioned in the previous category, further highlighted the students' struggle to conceptualize temperature in this context. In almost all groups, the students were able to reason more productively about the temperature limits of the two-state paramagnet. Several groups noticed the similarities with the three-state system but few recognized the isomorphism. One idea that emerged from three of the groups was that the heat bath, in the three-state system problem, corresponds to the magnetic field for the two-state paramagnet. Thus neither recognizing that there must be some kind of heat bath involved in the paramagnet problem to discuss the temperature limits. Nor realizing that there must be a similar interaction to the field dipole also in the three-state system to obtain energy states of that kind (related to the previous category).

Another idea, which caused confusion among about half of the groups, was that the heat bath implies interaction between the lattice sites. Consequently, it affects the probabilities, whether the partition function is separable, and whether they can assume that the particles are distinguishable or not. Again this suggests that the students struggle to recognize the assumptions of the context. Such as that the lattice sites are assumed to be noninteracting, and the heat bath is assumed to interact weakly with the system, only affecting it by exchanging energy. Group B also discussed that the heat bath makes everything "*vibrate the same way*," and Beth said,

Beth: *One thing that I just thought of, when you have a heat bath, it means that all of the… the lattices, basically everything is going to start looking like each other, so then the probability… so it's like we have just one configuration.*

This was similar to group E who reasoned that the two lattices strive toward being in the same energy state. Also stating that for a certain temperature, this would correspond to the "*same combo as the heat bath*." This was related to their discussions about equilibrium, see category A4.

7. Reasoning inconsistently about the Boltzmann distribution and the partition function, and their applicability

Another common challenge that emerged from our data was students' inconsistent reasoning about the Boltzmann distribution and the partition function. In general, the students struggled with the meaning of the concepts, where they come from, how to apply them, and in which contexts it is justified. Some groups seemed to have almost no established conceptualization of the Boltzmann distribution. Others displayed fragmented ideas and shallow associations with them. For example, groups A and B did not use it at all, but when prompted to talk about it toward the end of the session, they had some associations with probability and the term *heat bath*. They also remembered pieces of the mathematical expression of the Boltzmann factor, although without the crucial quantity of energy. Group C used fragments of the expression during their discussion about the three-state system problem, without referring to it as the Boltzmann distribution, but they struggled to connect the pieces. Only after some prompting by the interviewer could they connect it to what they knew about probability in general.

Group D and the groups from the second phase seemed more familiar with the terms Boltzmann distribution and the partition function. They used the names spontaneously, but the concepts remained challenging. The data suggest that many students did not recognize the connection to probability distributions, and they applied it to inappropriate contexts. For example, four groups attributed a Boltzmann distribution to the energy states already in part (a) of the three-state system problem, when the problem explicitly states what the energy states are assumed to be equally probable. The students articulated that the states are equally probable at the same time as they utilized the Boltzmann distribution. They wrote down an expression for the corresponding partition function and followed the procedure that would be appropriate in part (b), not appearing to recognize that in doing so, they do *not* assume that the states are equally probable. Group F even stated that probabilities are not related to the partition function, as they approached part (a):

Fiona: *Then we can, ehm, just compute the partition function and take the derivation of it.*
Fran: *Yeah.*
Filip: *Yeah, because now we want, will have to compute the value of U?*
Fiona: *[pause] Okay so, they all have equal probabilities? Ehm.*
Filip: *The probabilities never show up in the partition function anyways.*
Fiona: *That's true. We just sum over all states. Yeah.*

After struggling with the mathematical expressions and feeling stuck, they were prompted to discuss the probability more and compare the conditions in (a) and (b). The group eventually realized that they did something "*too complicated*" but still failed to recognize why their approach was inappropriate. As Fiona said, "*Okay so (a) seems quite forward, straightforward to compute. But the question is why doesn't the more difficult one work? I think we should think about the Boltzmann distribution more [laughs], but I can't find it at the moment*." Additionally, groups F, G, and H wanted to apply the approach with the Boltzmann distribution to the small paramagnet problem, where it is neither appropriate nor relevant.

We turn our attention back to the three-state system problem. The student groups who approached part (a) inappropriately eventually realized that their approach was meant for part (b). Seemingly through their association between the Boltzmann distribution and the term *heat bath*, either when reading question (b) or being prompted to reread the questions. However, they continued to struggle with mathematics and qualitative reasoning, particularly with the temperature limits (as described in category A6). Many students did not recognize that since the energy states are Boltzmann distributed, they are equally probable in the high-temperature limit. In some cases, they failed to recognize this despite having found the correct mathematical expressions. Students' ideas about the partition function were also fragmented. Sometimes they productively reasoned that it is a normalizing factor, but in other instances not connecting it appropriately. Only group I recognized the generality of the partition function. That the concept is not specific to something Boltzmann distributed, for example, by correctly stating that it is just a constant in part (a).

B. Challenges with problem-solving strategies

The second grouping of categories concerns identified challenges with the students' problem-solving strategies. This is also a broad grouping, as there are many aspects to approach a problem. Here we present three of the recurring challenges observed from our data. These are divided into categories involving (1) *inappropriately applying typical mathematical relations from statistical mechanics and thermodynamics*, (2) *having difficulties managing conflicts between quantitative results and qualitative reasoning*, and (3) *defining new concepts that led to more confusion due to their inconsistency with established relations*. We note again that these categories are not extensive, and that we have limited our analysis to challenges. Several productive problem-solving strategies emerged from the data as well. Namely, certain ways in which the student groups sometimes overcame a barrier or resolved a conflict, which we intend to explore further in future work.

Our findings suggest that the students generally struggled to interpret the problem statements and representations. For instance, they struggled to recognize the implicit assumptions of the contexts. It was also challenging to find an appropriate strategy for solving the problems. Additionally, quantitative reasoning was often challenging for the students in itself, as they struggled with the calculations (many times concerning exponential functions and dealing with limits). These findings are based on observations and are described further in each category.

First, before summarizing the emerged categories based on observations, we also want to highlight how these general challenges align with the students' self-reported difficulties. This is interesting from a metacognitive perspective, and it provides some support for our interpretation of the data. Toward the end of the interviews, the students articulated their general struggles with problem-solving strategies when prompted to reflect on the session and the course content. For instance, group I expressed that statistical mechanics feels "*hand-wavy*" just like thermodynamics. Moreover, they felt the need for more support with the "*reasoning heavy*" problems; how to navigate which assumptions are made, where it comes from, and why we do it. Many groups emphasized that it is difficult to discern all the abstract and "*unintuitive*" problems. Cole described the problem-solving experience through analogies:

Cole: *It's kind of like we take a stroll with no destination [laughs]. We are just walking randomly, in random directions […] We're like rolling a dice to decide our next step [laughs].*

Similarly, Fiona articulated that the interview session illustrated the issue she has with statistical mechanics.

Fiona: *Because I always have a problem of knowing why to use which formula, because we derive them and I understand the derivation of it […] But it's never explained, or I never get why we apply which model for which situation, and what is the difference between the situations.*

Our data suggest that these difficulties led the students to employ problem-solving strategies that sometimes were unproductive or inappropriate. This induced more challenges. In the following, we summarize the three identified categories of such tendencies, referred to as B1–B3.

1. Inappropriately applying typical mathematical relations from statistical mechanics and thermodynamics

One of the most evident patterns that emerged from the data was a tendency to apply a certain mathematical relation, or set of relations, in an inappropriate context. Sometimes by default or by making unjustified assumptions. In many cases, the students referred to a "stat mech approach." By this, they meant using the Boltzmann distribution for the energy states, finding the partition function, and taking derivatives of it to find the desired quantities. This was in some cases the students' default approach, which they employed without discussing if it was appropriate. In other cases, they motivated the approach by stating some unjustified assumption, like "*we assume it's Boltzmann distributed, we always do that in stat mech*." This tendency was mainly related to the Boltzmann distribution, a concept that we already described as one of the main challenges for the students. One reason could be this inconsistent conceptualization, being unsure of what the concept means and in which contexts it is applicable. Another reason could be that the students expect a certain solution approach simply because they have seen it in many or most of the examples in the course. One explicit example of this came from Hans. He suggested a way to solve part (a) of the three-state system problem, seemingly based on the surface features of the problem statement:

Hans: *Should we take it in a 'stat mech' approach? Because they are talking about lattices, epsilons?*

Group H first realized that this was inappropriate after being prompted by the interviewer to discuss the differences between part (a) and part (b) when they had been stuck for a while. At the end of the interview, they were asked to reflect on what, if anything, had felt confusing when solving the problems. One of the things they mentioned was that they had done part (b) before part (a) for this problem. Upon being asked why they thought that they had approached it in this way, they answered that perhaps they were so used to following the same procedure that they were eager to do it like that. That part (b) felt like a more natural example in the course content, and they did not know when to assume that something is Boltzmann distributed. As we mentioned in category A7, many groups did the same thing for the three-state system problem. Group F was one of the groups that did so by default. Only after facing a conflict in their reasoning did they ask themselves if it was justified:

Fran: *So maybe the question is; are we allowed to use the Z function here?*

Filip: *I mean, somehow we always use the partition function.*

Fiona: *Mm, we assume that it is Boltzmann distributed.*

Several groups even did so for the small paramagnet problem. For example, group H was conflicted about how to find the energy states: "*if we think in stat mech, the partition function mentions the energy, how would the energy differ?*," which is irrelevant in this context.

The early groups from the first phase of data collection, who seemed even more unsure about what the Boltzmann distribution is, displayed a tendency to engage in a more haphazard "plug and chug" approach. Mainly using thermal physics relations that they might have been more familiar with from their thermodynamics course. When group H expressed confusion about the difference between part (a) and part (b) of the three-state system problem, they proposed to "*go back to thermodynamics*." Another example is group A which struggled to find a suitable way of solving the three-state system problem. They approached it by, for example, trying various manipulations of the statistical definition of temperature and some differential relations of thermodynamic quantities. They were at times vocal about recognizing that they were unsure about their approach, for instance:

Adam: *Do we know what we are doing?*
Alex: *No, or like yes, but we want more things!*

There was also some disagreement within the group. Alex suggested that they just needed more things to play around with while Adam was more sceptical. Alice repeatedly voiced her concern about whether they actually needed all of the relations. Group B articulated a similar strategy when discussing the temperature limits of the same problem:

Ben: *Like logically for me, if these like, well we had more heat, more energy, right? Which would correspond to more excitation, more excitation would mean higher energy level which would mean more* $+1$*:s instead of* -1*:s and zeros, right?*
Beth: *Yeah.*
Ben: *That's what I logically think but I have no recollection of this being explicitly stated. So I don't know. [long pause, laughs].*
Bill: *I mean, we could just, we could maybe like experiment around with the fundamental thermodynamical relation, but I don't know what we would actually do with it… [laughs].*

Group C had a similar strategy, wanting to apply certain thermodynamics relations, such as $S = \frac{dQ}{T}$ and $S = PdV + dQT$, already in the four atoms problem.

2. Having difficulties managing conflicts between quantitative results and qualitative reasoning

This category describes the challenge that the student groups faced when attempting to deal with conflicts that arose when their quantitative results differed from their expected result based on qualitative reasoning. In some cases, the groups attempted to resolve the conflict by, for instance, revising or even changing the calculations to fit the result based on their initial ideas. They sometimes referred to these initial ideas as their "*gut-feeling*" or "*intuition*" [41]. This was problematic when the calculations were correct and they continued to rely on their unproductive qualitative reasoning, despite facing inconsistencies. In many cases, the conflicts were left unresolved due to, for example, the students ignoring it, being frustrated after trying many approaches, or running out of time.

We structure the following examples from our data by first identifying the conflict and then describing the students' struggle to manage the conflict. One example was mentioned in category A4, about the concept of equilibrium. In the context of the four atoms problem, there was tension in both groups A and C between the students' qualitative reasoning about the probabilities and their quantitative result from calculations of the number of microstates per macrostate. In group C, for example, Cole expressed that it "*feels weird that this [the left configuration] is more probable than that one [the right configuration], it feels counter-intuitive*." In their attempts to resolve this conflict, the group struggled in many ways, such as turning to a plug and chug approach. The confusion grew more the more they tried to manage the conflict, which also gave rise to new conflicts. As more time passed, they dropped the discussion without resolving it.

The most common example was, however, the conflict arising from the discussions about the high-temperature limit of the three-state system. Most of the groups agreed that they expected the average total energy to reach a maximum value, $+2$, in the high-temperature limit. For the groups who managed to carry out the calculations correctly, the students were conflicted when the quantitative result (average total energy goes to zero) did not match their expectations. When it came to managing this conflict, several of these groups relied on their initial qualitative reasoning more than their quantitative, which was unproductive for solving the problem correctly in this case. Some groups even changed their calculations in an attempt to match their persistent "intuitive answer", as group F phrased it:

Fiona: *Yeah. But is it, is this incorrect, the '*$E = 0$*'?*
Fran: *Well it doesn't map with our intuitive answer.*
Fiona: *Yeah, but I feel like our intuition is still right.*
Fiona: *[A bit later] Okay, so do we say this is, our value is correct or our intuition is correct? Maybe we could look in our calculation again?*
Filip: *Yeah, maybe made a mistake somewhere.*
Fiona: *I guess high temperature means we have, for both states they should be in the highest energy state, and we assume the highest energy state to be 1?*

Another aspect of this challenge was that the students also struggled with mathematics, introducing even more uncertainty in their problem-solving strategies. This could be one possible contribution to why these conflicts were difficult to manage. Another could be conflicts between students in the group. This added more tension and disagreement to the problem-solving process, such as in group I where the students' qualitative reasoning differed and they also struggled with the calculations. Ike said "*if we think physically, and not mathematically, it feels reasonable that it is zero, right?*". Ivan, on the other hand, reasoned that the heat bath "adds heat," which should mean that the average energy goes to 2. However, even after they corrected their quantitative calculations and it matched Ike's qualitative reasoning, the group was still unsure about the high-temperature limit. Likely due to their restricted conceptualization of temperature (see category A6).

3. Defining new concepts that led to more confusion due to their inconsistency with established relations

Another trend, although not as common as the other two in this grouping, was students defining a new concept in an attempt to resolve a confusion. This led to new challenges when the new concept turned out to be inconsistent with other concepts and relations between them. We identified this strategy in the context of challenges with the concepts of macrostates and microstates and distinguishable, as described in categories A1 and A3. For groups A–D, this was mainly during the discussions about the four atoms problem. Adam expressed it explicitly, "*I'm just totally making stuff up!*" when the group discussed if there could be even more layers of states. Macrostates, "regular" microstates, and what they referred to as "sub-microstates." However, their concept of sub-microstates caused more confusion within the group, due to its inconsistency with the postulate of equal probability. They knew that the postulate is associated with "regular" microstates as well as entropy, which they correctly associated with the number of microstates per macrostate. Group C faced the same issue when discussing a "*degeneracy of microstates.*" Similarly, group D wanted to say that there are "*different ways to build a microstate*." Group B also defined a layering of states, in a *hierarchy*:

Ben: *So there is kind of a hierarchy. At the bottom we have the configuration of the atoms in their different energy levels. The next point would be the ways we can configure the atoms in the energy levels. Like have one in* 4*e the rest in zero [left configuration], or one in* 3*e*, *one in* 1*e two in* 0*e [middle configuration]. And then the next after that is total energy level, we can just choose. Like* 5*e is the total energy,* 6*e total energy or whatever. Like there is a hierarchy, and the macro always has to be on top*.

Similarly to groups A–D, group E discussed an additional "*layer*" of states during the small paramagnet problem. After some back-and-forth discussions about the concepts of macrostates and microstates, the group felt more confident in answering that there were five macrostates and two microstates (up and down). They were then confused about what to call the configurations in the figure and jokingly suggested that these 16 configurations are the "*middle states.*"

Group B defined another new concept, which emerged from the common confusion about distinguishable vs indistinguishable, in the context of counting states. They were, however, the only group that defined a new concept in this situation. They related their concept of "*super-distinguishable*" to counting "swaps within an energy level" as yielding new microstates. As we have discussed previously, this does not have any physical meaning and rather suggests that the students might be interpreting the energy representations (drawing dots in the energy levels) as something spatial. There was some disagreement within the group, with Bill noticing that it does not make much sense to say "*where in the state something is*." Interestingly, we did not identify this particular confusion during the second phase of data collection. Then the small paramagnet problem, with actual spatial representations, had replaced the four atoms problem. Further exploration of how the students discerned and used representations to reason about these problems is a promising avenue for future research.

V. DISCUSSION

First of all, we provide a general discussion of our findings and alternative perspectives about the grouping of our categories. Overall, we noticed that the topic of statistical mechanics seemed to present various challenges for the students. These challenges involved a broad range of relevant terms and concepts, as well as their approaches to solving problems. The students often appeared to have vague or haphazardly constructed ideas about key concepts. Moreover, they struggled to interpret the problem statements and recognize the relevant assumptions of the contexts. As the students expressed themselves, they lacked an understanding of the connections between all the concepts and they were also unsure of in which situations they could apply the corresponding mathematical expressions. These difficulties were often magnified by the intrinsic mathematical demands of the exercises, as well as their abstractness. We identified the challenges through our observations of the problem-solving sessions, as well as through students' self-reporting at the end of the activities.

As we noted already in the findings section, there are possible connections between the identified challenges. They were also not observed independently of one another. Considering the purpose of our study, we decided to keep a broad perspective and present our findings as separate challenges without looking deeper into the relationships between them. We note that the possible connections are an interesting aspect to explore in future work. We discuss this briefly in Sec. V C, along with potentially suitable theoretical perspectives. Additionally, we recognize that the categories could be grouped differently. If some connections are considered, one alternative to grouping A could be to highlight that categories A2, A4, A5, and A6, are all related to the central concepts of entropy, temperature, and energy. These are in turn related to each other, for example, through the theoretical definition of temperature, see Eq. (2). Categories A1 and A3 mainly deal with counting states and probability distributions, while A7 draws from all of these aspects. Moreover, the challenges in grouping B are connected to the challenges in grouping A. Successful solutions mostly require both proficiency in problem-solving skills as well as understanding the concepts.

In the following, we first summarize how our findings compare to other research. Next, we discuss how our findings can inform physics instruction. Then, we suggest ideas for future work and comment on some possible theoretical frameworks that could be fruitful to apply in such studies. Finally, we consider the limitations of our study.

A. Findings in relation to previous research

Parts of our findings, from both main groupings of challenges, share similarities with previously published results. Several of the conceptual challenges presented in grouping A have been identified in various contexts by other researchers. For example, students' difficulty in distinguishing between macrostates and microstates has been observed for systems with varying simplicity [12,13,17]. This corresponds to one of our most common challenges described in category A1. We note that while some of the graduate students in Ref. [17] could appropriately classify macrostates for a question similar to the four atoms problem, many of them still struggled with generating suitable definition of the macrostates. We also found that the students struggled with the definitions as well as the concepts behind them. This bears similarities with the conclusion by Lo [22], who reported that "*most students did not recall a rigorous definition of microstates and macrostates, though most have the idea that microstates somehow define the macrostates*" (p. 86). However, the main finding related to macrostates and microstates in the current PER literature seems to be focused on the challenge of distinguishing between the two concepts. Loverude [12] also reported that students struggled to utilize mathematical expressions for the statistical weight, similar to one of the traits in A1.

The related findings in category A2 also support the conclusions by, e.g., Leinonen *et al.* [15]. Namely, the large gap between macroscopic and microscopic perspectives of thermal physics concepts, such as entropy, makes the connection and transfer between those levels difficult to grasp. Our findings in category A2 are also similar to an observation reported in Crossette *et al.* [17], where they suggested that some students may be projecting macrostate properties on to constituent microstates. Additionally, our previous discussion is supported by the closing remarks in Leinonen *et al.* [15]. There they claimed that this transfer to statistical perspectives could confuse students even more since it demands abstract models, approximations, and assumptions.

As we mentioned in the introduction, the concept of entropy has been investigated extensively from many different perspectives, also from a statistical mechanics point of view. Likewise, entropy is prominently featured in our data. Our findings suggest that the students had many ideas about the concept and struggled to utilize it as the more fundamental concept than, for example, temperature. It would be possible to identify various metaphors or resources that our students used to reason about entropy. Similarly to Haglund *et al.* [20] and Haglund [21] in the context of thermodynamics, and Loverude [14] and Crossette *et al.* [17] in the context of upper-division thermal physics. However, we did not consider this perspective as a part of the analysis. The two latter studies presented a challenge that corresponds to our category A4, involving students' conceptualization of equilibrium. Loverude [14] found that "*For example, many students used a previously unreported association between the equilibrium state of a system and an increase in order, rather than disorder*" (p. 1).

Another evident similarity in reported student challenges is related to the Boltzmann distribution. Students have been found to struggle to understand what it is and fail to recognize when it is applicable, as reported by Smith *et al.* [6]. Our findings in category A7 resemble this. Furthermore, it relates to our identified challenges regarding students' tendency to apply typical thermal physics relations inappropriately (category B1) since several of the inappropriately applied mathematical relations were connected to the Boltzmann distribution. Moreover, category B1 shares attributes with the recurring finding of students' excessive use of the second law of thermodynamics, as reported by, for example, Christensen *et al.* [3]. We also note that category B2, involving conflicts in the students' problem-solving process, resembles some of the findings in Geller *et al.* [42] about tension in students' reasoning about entropy and spontaneity. Additionally, like our remark in the general description of grouping B, Geller *et al.* [42] reported that part of these challenges were also expressed by the students: "*At times students recognized these tensions in their own understanding, and at other times we identified tensions that students did not fully articulate themselves*" (p. 397).

Finally, based on our data, we suspected that the students possess several loosely connected ideas about the considered concepts, which they sometimes struggled to navigate around. Our findings suggest a need for increased conceptual understanding, which was also self-reported by the students. These findings echo the conclusions of Lo's Ph.D. thesis [22], which considered a range of key concepts, as well as Loverude [14] who reported that students frequently shifted between ideas in his study. In the context of the velocity distribution function, Erceg *et al.* [16] suggested that traditional physics instruction can sometimes help students learn mathematical operational knowledge, without developing the corresponding conceptual understanding. We touched upon this in grouping B. We mentioned that our students sometimes were able to succeed in the mathematical calculations but nevertheless struggled to solve the problem due to their qualitative reasoning. The remaining findings (categories A3, A5, A6, and B3) are, to our knowledge, unreported challenges in the context of upper-division thermal physics content.

B. Implications for instruction

The categories presented in this study can provide helpful insight to upper-division thermal physics instructors, for example, into potential challenges that their students might face. Being aware of possible difficult areas can guide teachers as to where they might place more emphasis or adjust their teaching. Keep in mind that this study has not investigated any instructional interventions. Therefore, the following serves as recommendations based on our current findings and should be considered in relation to the knowledge and context of each statistical mechanics instructor.

In general, our findings suggest that there is a need for an increased focus on conceptual understanding in the teaching activities. The introduction of new concepts would benefit from a clear connection to both the students' prior knowledge, other concepts from the course, and to the appropriate mathematical relations. We believe that students would benefit from discussing concepts and problem-solving strategies in groups, as well as from explicit and articulate guidance from the instructor. This is neither limited to this specific physics topic, nor is it a novel recommendation in PER. Nevertheless, it is an important advice and it is worth repeating. We want to stress that physics terms and concepts can evolve and change meaning from one area of physics to another, and these transitions of concepts could be expressed explicitly to students. These key concepts are, after all, necessary to understand the topic, frame questions appropriately, and solve problems.

Due to the complexity of statistical mechanics, instructors often utilize idealized "toy models" to illustrate concepts and allow simpler calculations, even explicit counting of states. We note that this has beneficial features, but we also encourage teachers to consider that using simple systems as examples does not imply that the intended message is easily received by the students. Moreover, based on our findings we suspect that stripping such idealized problems of context, such as in the three-state system problem, could introduce even more challenges. Schroeder [1] proposed that the two-state paramagnet is beneficial since "*it forces us to think primarily in terms of entropy rather than temperature. Entropy is the more fundamental quantity, governed by the second law of thermodynamics. Temperature is less fundamental; it is merely a characterization of a system's "willingness" to give up energy, that is, of the relationship between its energy and entropy*" (p. 102). However, the subtle and "unintuitive" aspects of models like this can be very challenging for students, as we observed for the analogous three-state system. Therefore, we suggest that instructors draw attention to the purpose of using such examples (such as the suggestion by Moore and Schroeder), make the assumptions of the situation explicit, and consider what real-life applications can be used as credible context. For example, instructors may consider carefully walking the students through the two-state paramagnet while explicitly highlighting the assumptions and important choices in each step of the problem-solving process. In line with the recommendation of Crossette *et al.* [17], we suggest that instructors directly address the subtleties of the relationship between entropy, temperature, and energy with their students.

C. Potential theoretical perspectives and future work

Considering the purpose and methodology of this study, we intentionally refrained from employing a certain theoretical perspective in our analysis and presentation of our findings. However, during the course of this project, we noticed interesting aspects to explore further in the future, where we suspect certain theoretical frameworks could be beneficial as lenses. One observation is that our two main groupings, involving challenges with concepts (A) and problem-solving strategies (B), could be considered through the resources perspective [40] by investigating students' conceptual and epistemological resources, respectively. The resources framework can be used to analyze interesting parts of our data at a finer grain size by identifying activated resources. Which could be productive or unproductive from the perspective of the physics discipline. Potential findings could then inform instructors to identify students' productive resources and help them use these resources in appropriate ways. This could be done, for example, by implementing certain exercises, shifting their focus, and leveraging productive analogies.

As we mentioned earlier in the discussion, our categories A2, A4, A5, and A6, are all related to the central concepts of entropy, temperature, and energy. In statistical mechanics, these concepts are, in turn, connected through the theoretical definition of temperature. Students' intuitions regarding key concepts seemed to play an important role in their choices during problem solving, as captured in the category B2. Future work could investigate the evolution of students' conceptualization of temperature from everyday life, to classical thermodynamics, to statistical mechanics. Insights from such research could, in turn, inform how we can help students develop a cross-cutting understanding of temperature and other concepts. This is especially important since we have observed that the process of transitioning between the domains of everyday life, classical thermodynamics, and statistical mechanics appears to be closely related to many of the identified challenges.

A related theoretical angle on problem-solving strategies is the concept of *framing*. That is, the activation of a set of resources depending on the situation or activity (e.g., Hammer *et al.* [43]). Students have resources for framing activities epistemologically. In other words, the *epistemological framing* refers to students' ideas about what knowledge is appropriate to consider in a specific context, as described by Bing and Redish [44]. Paying attention to *epistemic games* [7] that students engage in may offer additional insights. Moreover, the possible connections between the challenges identified in our study could be investigated from the viewpoint of combinations or connections between conceptual and epistemological resources. This has been considered by, for instance, Richards *et al.* [45] who studied conceptual breakthroughs in student reasoning. Another example is Samuelsson [46] who developed a new framework, a synthesis of the resources framework and social semiotics, which considers combinations of conceptual, epistemological, and semiotic resources. We suspect that representations play a significant role in how students interpret and approach tasks (categories B1–B3), as well as in what conceptual challenges they experience (categories A1–A7). For example, our identified challenges in A1 and A3, seem to be related to how the students interpret the representations provided in the tasks, as well as the representations they generate themselves. Specifically, students tended to interpret the energy-space representation in the four atoms problem as a spatial representation, likely leading them into the pitfall of "exchanging atoms within an energy level." Future work could investigate how an actual spatial representation of the same situation might change the students' reasoning and approach. A framework that focuses on the function and use of representations, such as social semiotics [47], would then be a suitable theoretical lens.

We note that there are even more theoretical viewpoints regarding problem-solving strategies. Such as the *cognitive heuristics* described by Talanquer [48] in the context of chemistry education. Another is the dual-process theories (DPTs), which were considered by Kryjevskaia *et al.* [49] to formulate a schematic model of reasoning pathways. We believe this model could be fruitful in analyzing our data in terms of intuitive and analytic processes. Additionally, their characterization of *possible hazards* in reasoning pathways could be used as a lens to look deeper into our identified challenges and conflicts in the students' problem-solving strategies (categories B1–B3). We mentioned in categories A5 and B2 that in their reasoning about the three-state system problem, students employed analogies that drew on their intuitions for related physics phenomena. These episodes could be analyzed deeper in terms of intuition and analogical reasoning. For example, future work could investigate how students' existing cognitive resources play into the process of problem solving in statistical mechanics. Our findings suggest that such resources can play a significant role, especially when the students rely heavily on them. They may even determine whether the students successfully solve the problem or not.

This exploratory study considered identifying recurring challenges, while productive ideas or resources to deal with those challenges were not categorized systematically at this stage. Future work, as we discussed previously, could take this into account and provide more specific and valuable advice for teachers. In addition to being aware of potential difficulties, those findings could inform instructors how they can approach these student challenges. We envision that such research could lead to practical suggestions on how to, for example, vary the context of exercises to help students develop skills to frame physics tasks productively. Additionally, future work is necessary to develop instructional material to aid instructors in addressing these challenges explicitly.

D. Limitations

Two apparent limitations of our study are the small sample size and the selection bias due to volunteering student participants. Consequently, our findings cannot be generalized over all possible populations of students, and they are not meant to be. In other words, they should not be considered a complete categorization of student challenges in statistical mechanics. Our intention was instead to gain insight into challenges in student reasoning by collecting and analyzing rich, detailed data. Thus, the findings must be considered in relation to the specific contexts. However, they can still provide suggestions and guidance applicable to similar contexts. While all student groups had the chance to attend the lectures covering the necessary content for our designed problems, the interview sessions were spread out during the course, more so in the first phase. This is a potential bias, due to the unequal amount of time the participants had to process the course content before the interviews.

Moreover, we acknowledge that our analysis is an interpretation of the students' expressed reasoning. This does not necessarily reflect the whole picture or their individual perspectives. While there are many advantages to group interviews, as we mentioned in Sec. III A, it is important to consider the undesired influences of the group dynamic, the *in vitro* setting, and the role of the moderator. Despite our efforts to minimize these effects, it is impossible to eliminate them. We also point out the consequences of performing a large-grained analysis, such as that nuances and details can be overlooked. That being said, the findings still provide inspiration for further studies where the interesting trends can be investigated deeper.

VI. CONCLUSIONS

We set out to explore potential challenges that upper-division physics students face when solving statistical mechanics problems in groups. In this paper, we present findings from nine groups of two to three students, who were volunteers from a statistical mechanics course at a large Swedish university. We present findings consisting of ten categories of recurring challenges that emerged from the data, divided into two broad groupings. Namely, seven categories of challenges with concepts and three categories of challenges with problem-solving strategies. The first seven categories involve different problematic trends among the student groups related to key concepts such as macrostates and microstates, distinguishable and indistinguishable particles, temperature, entropy, energy, equilibrium, heat bath, the Boltzmann distribution, and the partition function. The last three categories feature the following unproductive patterns; the inappropriate application of common relations, the difficulty to manage tensions between calculated results and qualitative reasoning, and the definition of new and inconsistent concepts.

Based on our findings, we suggest ideas for future work and discuss potential implications for instruction. In general, we recommend instructors of upper-division thermal physics courses to consider keeping our identified challenges in mind and adjusting the teaching to account for them. Whether as a proactive measure or as a response to recognizing a challenge. Due to the many complex and subtle concepts in statistical mechanics, we suspect that it is worthwhile making some aspects more explicit when introducing concepts and problem-solving approaches, particularly in the case of highly abstract and idealized systems. Future work is, however, necessary to propose more concrete advice for teachers. Finally, we express how this exploratory study inspires promising ideas for future research. Such as considering different theoretical perspectives to investigate the role of students' epistemological framing, intuitions and analogies, and representations in these contexts. Such studies can hopefully provide further insight into students' reasoning and provide specific implications for instructors.

ACKNOWLEDGMENTS

We thank Christopher Robin Samuelsson and the rest of our colleagues in the Uppsala University Physics Education Research Group, as well as visitor to the group, Saalih Allie, for valuable discussions that helped shape this project. Guido Festuccia is thankfully acknowledged for discussing the findings and reviewing the physics content. A special thanks to all the student participants as well.

E. K. designed the project, collected the data, and carried out several iterations of data analysis, under the supervision of B. G. The authors discussed the findings from the data analysis continuously as the process evolved. E. K. wrote the original draft of the manuscript. Both authors contributed to writing the final version.

[1] D. V. Schroeder, *An Introduction to Thermal Physics* (OUP Oxford, Oxford, 2021).
[2] 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).
[3] W. M. Christensen, D. E. Meltzer, and C. A. Ogilvie, Student ideas regarding entropy and the second law of thermodynamics in an introductory physics course, Am. J. Phys. **77**, 907 (2009).
[4] 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).
[5] 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).
[6] 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).
[7] T. J. Bing and E. F. Redish, Epistemic complexity and the journeyman-expert transition, Phys. Rev. ST Phys. Educ. Res. **8**, 010105 (2012).
[8] D. E. Meltzer, Observations of general learning patterns in an upper-level thermal physics course, AIP Conf. Proc. **1179**, 31 (2009).
[9] D. B. Mountcastle, B. R. Bucy, and J. R. Thompson, Student estimates of probability and uncertainty in Advanced Laboratory and Statistical Physics Courses, AIP Conf. Proc. **951**, 152 (2007).
[10] T. I. Smith, D. B. Mountcastle, and J. R. Thompson, Identifying student difficulties with conflicting ideas in statistical mechanics, AIP Conf. Proc. **1513**, 386 (2013).
[11] 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).
[12] M. E. Loverude, Student understanding of basic probability concepts in an Upper-Division Thermal Physics Course, AIP Conf. Proc. **1179**, 189 (2009).
[13] M. E. Loverude, Investigating student understanding for a statistical analysis of two thermally interacting solids, AIP Conf. Proc. **1289**, 213 (2010).
[14] M. E. Loverude, Identifying student resources in reasoning about entropy and the approach to thermal equilibrium, Phys. Rev. ST Phys. Educ. Res. **11**, 020118 (2015).
[15] 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).

[16] N. Erceg, I. Aviani, V. Mešić, M. Glunčić, and G. Žauhar, Development of the kinetic molecular theory of gases concept inventory: Preliminary results on university students' misconceptions, Phys. Rev. Phys. Educ. Res. **12**, 020139 (2016).

[17] 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).

[18] 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).

[19] K. D. Rainey, M. Vignal, and B. R. Wilcox, Validation of a coupled, multiple response assessment for upper-division thermal physics, Phys. Rev. Phys. Educ. Res. **18**, 020116 (2022).

[20] J. Haglund, S. Andersson, and M. Elmgren, Language aspects of engineering students' view of entropy, Chem. Educ. Res. Pract. **17**, 489 (2016).

[21] J. Haglund, Good use of a 'Bad' Metaphor: Entropy as disorder, Sci. Educ. **26**, 205 (2017).

[22] W. C. H. Lo, Student understanding of statistical mechanics, Ph.D. thesis, North Carolina State University, NC, 2022.

[23] M. E. Loverude and B. S. Ambrose, Editorial: Focused collection: PER in Upper-Division Physics Courses, Phys. Rev. ST Phys. Educ. Res. **11**, 020002 (2015).

[24] We identified a few cases where student groups were considered in the context of statistical mechanics content. Students were interviewed individually or in groups in Ref. [10] as well as in Ref. [11], the latter also considered videotapes of small student groups working on a tutorial together.

[25] E. Koerfer, Student intuitions in statistical mechanics [Student project, Uppsala University (2022)].

[26] E. Koerfer and B. Gregorcic, Student Intuitions in Statistical Mechanics [Conference presentation abstract], GIREP 2022 Effective Learning in Physics from Contemporary Physics to Remote Settings, Ljubljana (2022).

[27] F. Mandl, *Statistical Physics*, The Manchester Physics Series (Wiley, Chichester, 1987), 2nd ed.

[28] While some examples appear to be exceptions to this, like the seemingly continuous energy range of a knocked out electron from an ionized atom, the system can always be forced to have discrete states by enclosing it in a finite volume. This is comparable to limiting the length of a waving string to a finite value. The general waveform is then described by a Fourier series with discrete harmonic modes rather than a Fourier integral [27].

[29] F. Reif, *Fundamentals of Statistical and Thermal Physics* (McGraw-Hill, New York, 1965).

[30] C. Kittel and H. Kroemer, *Thermal Physics* (W.H. Freeman, San Francisco, CA, 1980), 2nd ed.

[31] T. A. Moore and D. V. Schroeder, A different approach to introducing statistical mechanics, Am. J. Phys. **65**, 26 (1997).

[32] J. J. Sakurai, *Modern quantum mechanics* (Addison-Wesley, Reading, MA, 2011), 2nd ed.

[33] B. Glaser and A. Strauss, *Discovery of Grounded Theory: Strategies for Qualitative Research* (Routledge, New York, 2017).

[34] S. Booth and F. Marton, *Learning and Awareness* (Routledge, New York, 1997).

[35] V. Braun and V. Clarke, Thematic analysis, in *APA Handbook of Research Methods in Psychology, Vol. 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*, APA Handbooks in Psychology® (American Psychological Association, Washington, DC, 2012), pp. 57–71.

[36] K. A. Ericsson and H. A. Simon, How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking, Mind Cult. Act. **5**, 178 (1998).

[37] D. W. Johnson and R. T. Johnson, *Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning* (Prentice-Hall, Inc, Englewood Cliffs, NJ, 1987), 2nd ed.

[38] One example in the literature defines a new term (the distribution), something between a macrostate and microstate, for a particular system such as this one. It says how many particles are in each energy state but not which ones they are; this latter information is given by the microstate [39].

[39] D. Trevena, 2—some basic ideas, in *Statistical Mechanics* (Elsevier Ltd., New York, 2002), pp. 5–17.

[40] D. Hammer, Student resources for learning introductory physics, Am. J. Phys. **68**, S52 (2000).

[41] These are the students' own words, we will not define *intuition* here but we discuss how it has been defined in the literature in Sec. V, and we note that it is potentially interesting for future studies.

[42] B. D. Geller, B. W. Dreyfus, J. Gouvea, V. Sawtelle, C. Turpen, and E. F. Redish, Entropy and spontaneity in an introductory physics course for life science students, Am. J. Phys. **82**, 394 (2014).

[43] D. Hammer, A. Elby, R. E. Scherr, and E. F. Redish, Resources, framing, and transfer, in *Transfer of Learning from a Modern Multidisciplinary Perspective* (Information Age Publishing, Greenwich, CT, 2005), pp. 89–119.

[44] T. J. Bing and E. F. Redish, Analyzing problem solving using math in physics: Epistemological framing via warrants, Phys. Rev. ST Phys. Educ. Res. **5**, 020108 (2009).

[45] A. J. Richards, D. C. Jones, and E. Etkina, How students combine resources to make conceptual breakthroughs, Res. Sci. Educ. **50**, 1119 (2020).

[46] C. R. Samuelsson, Resourceful students: Engaging students in active and systematic investigations in laboratories involving thermal phenomena, Ph.D. thesis, Uppsala University, 2023.

[47] J. Airey and C. Linder, Social semiotics in university physics education, in *Multiple Representations in Physics Education*, edited by D. F. Treagust, R. Duit, and H. E. Fischer, Models and Modeling in Science Education (Springer International Publishing, Cham, 2017), pp. 95–122.

[48] V. Talanquer, Chemistry education: Ten heuristics to tame, J. Chem. Educ. **91**, 1091 (2014).

[49] M. Kryjevskaia, P. R. L. Heron, and A. F. Heckler, Intuitive or rational? Students and experts need to be both, Phys. Today **74**, No. 8, 28 (2021).