# Motivations for using the item response theory nominal response model to rank responses to multiple-choice items

Trevor I. Smith[1,*] and Nasrine Bendjilali[2]

[1]*Department of Physics & Astronomy and Department of STEAM Education, Rowan University, Glassboro, New Jersey 08028, USA*

[2]*Department of Mathematics, Rowan University, Glassboro, New Jersey 08028, USA*

(Received 8 October 2021; accepted 23 March 2022; published 25 April 2022)

Several recent studies have employed item response theory (IRT) to rank incorrect responses to commonly used research-based multiple-choice assessments. These studies use Bock's nominal response model (NRM) for applying IRT to categorical (nondichotomous) data, but the response rankings only utilize half of the parameters estimated by the model. We present a mathematical argument for why this practice of using half of the NRM parameters when ranking responses is appropriate based on the primary question of multiple-choice tests: How can we use students' responses to test items to estimate their overall knowledge levels? We provide additional motivation for this practice by recognizing the similarities between Bock's NRM and the probability function of the canonical ensemble with degenerate energy states. As physicists often do, we exploit these mathematical similarities to gain new insights into the meaning of the IRT parameters and a richer understanding of the relationship between these parameters and student knowledge.

DOI: 10.1103/PhysRevPhysEducRes.18.010133

## I. INTRODUCTION

Item response theory (IRT) are mathematical models used to link examinees' responses to items on a test to a latent trait (students' ability or overall knowledge), via item response functions. IRT models assume that the latent trait (student's ability) is unknown but organized on a continuum, with the goal to use students' responses on a test to measure students' latent trait and determine their position on the latent trait continuum. IRT has become a popular method for analyzing data from multiple-choice research-based assessment instruments, with the latent trait $\theta$ interpreted as a student's overall understanding of the topic being tested (e.g., force and motion) [1–12]. Traditionally, IRT analyses model the probability of a correct response to an item as a function of item properties (such as item difficulty) for students with different latent trait values.

One of the most common and basic IRT models is the two-parameter logistic (2PL) model, which defines the probability that a student will answer a particular item $j$ correctly as a function of two item parameters, based on the student's latent trait $\theta$

$$P_j(1|\theta) = \frac{1}{1 + e^{-a_j(\theta - b_j)}}, \tag{1}$$

where $b_j$ and $a_j$ are the IRT difficulty and discrimination parameters for item $j$, respectively. The difficulty (location) parameter is the point on the latent scale with median probability. The discrimination (slope) parameter determines the rate at which the probability of choosing the correct answer changes given ability levels. Thus, items with high discrimination can be used to better map students' ability along the latent continuum.

The 2PL model may be reparametrized as

$$P_j(1|\theta) = \frac{e^{a_j\theta + d_j}}{1 + e^{a_j\theta + d_j}} \tag{2}$$

by defining $d_j = -a_j b_j$ [13]. In the psychometric literature, $a_j$ in Eq. (2) is often referred to as the "slope" parameter, and $d_j$ is often referred to as the "intercept" parameter due to the exponent $a_j\theta + d_j$ being a linear function of $\theta$.

Item response theory models have become popular tools in test development and assessing students' knowledge [1,4–6,9,14–20]. Ding and Beichner elevated awareness of IRT analyses by including them among a suite of methods for analyzing multiple-choice data more deeply than had been traditionally done in physics education research [2]. Wang and Bao used a three-parameter logistic model (3PL, including a "guessing" parameter) to analyze data from the

*smithtr@rowan.edu

Force Concept Inventory (FCI) [21] and showed that the IRT latent trait is somewhat correlated with overall test score [3]. Stewart *et al.*, Yang *et al.*, and Eaton and Willoughby have used multidimensional 2PL models to examine the substructure of tests such as the FCI and Force and Motion Conceptual Evaluation (FMCE) [22] using both exploratory and confirmatory factor analyses [7,8,23].

In 1972, Bock introduced the polytomous IRT nominal response model (NRM) to analyze responses to items with two or more nominal unordered categories [24]. Bock's NRM expresses "the probability that a subject of ability $\theta$ will respond to item $j$ in category $k$" as

$$P_j(k|\theta) = \frac{e^{a_{k,j}\theta + d_{k,j}}}{\sum_{i=1}^{n} e^{a_{i,j}\theta + d_{i,j}}}, \tag{3}$$

where the summation in the denominator is performed over all $n$ response options for a given item $j$ [24]. Unlike IRT models for dichotomously scored items that treat all incorrect response options as being equally incorrect by grouping them into a single category, Bock's NRM treats each response separately. The calculation of $\theta$ in the NRM incorporates which incorrect responses students select, thereby acknowledging that different incorrect responses may indicate different levels of understanding. The $a_{k,j}$ and $d_{k,j}$ parameters estimated using the NRM are determined for each response option $k$ for a given item $j$, and do not have the same interpretation as in the dichotomously scored models [25,26]. Additionally, the set of $\{a_{k,j}, d_{k,j}\}$ parameters must include anchoring conditions to be able to uniquely determine the parameter values (e.g., $a_{0,j} = d_{0,j} = 0, \ \forall \ j$) [24]; therefore a single $a_{k,j}$ or $d_{k,j}$ value cannot be considered high or low without comparison to the entire set of values for that item [27].

Since Bock's NRM does not assume ordered response categories, it can be used to empirically rank response categories where ordering of responses is of interest, and parameter values can be used to rank responses based on their relationship to the latent trait $\theta$ [25]. Ranking all responses to multiple-choice items can have many benefits for representing students' understanding of the material being tested, especially students who choose different incorrect responses before and after instruction. These rankings could be used to assign partial credit for responses that are not completely correct [11]. They could also be used in other analyses, such as transition matrices and consistency plots, which show how students change their responses to test items from pre- to postinstruction without explicitly assigning scores [28,29]. Well-established rankings of responses could inform interpretations of test results by showing whether students selected better, equivalent, or worse responses after instruction than they did before.

The use of $a_{k,j}$ parameters estimated by the NRM for ordering and comparing item responses based on their relationship to the latent trait $\theta$ is well established in the psychometric literature. This is based on the idea that a higher value of $a_{k,j}$ indicates a response that is more closely correlated with higher values of the latent trait and, therefore, better than a response with a lower value [30]. Wainer, Sirecki, and Thissen presented an argument for ordering responses based on the $a_{k,j}$ parameters by using the odds ratio between two different response options [31]. Bock and Moustaki paraphrased this work by stating that, "increasing $\theta$ implies greater probability of response in a higher category rather than lower if and only if the $[a_{k,j}]$ parameter of the higher category is greater than that of the lower" [30]. This is consistent with the interpretation of the $a_{k,j}$ parameters in other works, but the rationales focus on comparisons between pairs of response options rather than ranking the set of all responses to an item [24,32,33].

Smith, Louis, Ricci, and Bendjilali used an IRT nested-logit model that combines the 2PL with the NRM to analyze and rank incorrect response options to FMCE items based on the $a_{k,j}$ parameters [10]. Suh and Bolt developed the 2PL-NRM nested-logit model to analyze an item using the 2PL to determine the probability of selecting the correct response, and using the NRM to determine the relative probability of selecting each of the incorrect responses [34]. Eaton, Johnson, and Willoughby also used the 2PL-NRM to rank incorrect responses to FCI items, and suggested a method for assigning partial credit based on the $a_{k,j}$ parameters [11].

In this paper, we provide new insights into why the ordering of the $a_{k,j}$ values from the NRM may be used as a proxy for ordering the "correctness" of each response option. In Sec. II, we provide a new mathematical rationale for ordering responses using $a_{k,j}$ as a proxy for correctness under certain assumptions. In Sec. III, we propose an analogy between the NRM and a well-known physical model: the canonical ensemble with degenerate energy states. Thinking about IRT parameters and variables in terms of familiar quantities like energy and temperature provides conceptual insight that augments the purely mathematical arguments in Sec. II. We show that the $a_{k,j}$ parameter may be considered analogous to the "energy" of response $k$, with the correct response consistently having the highest energy. The physical interpretation of NRM parameters has been briefly introduced by Smith *et al.* in the supplemental material of Ref. [10]. In the current work, we develop this further by providing additional details about the analogy and exploring how we can extend it by considering quantities analogous to thermodynamic variables, such as average energy.

## II. MATHEMATICAL RATIONALE FOR USING IRT PARAMETERS TO RANK RESPONSES

The IRT probability expressions in Eqs. (1)–(3) center on modeling the probability of a student selecting a particular response, given the value of their latent trait. In practice, however, the goal of testing is to use the responses students

TABLE I. The set of $a_{k,j}$ and $d_{k,j}$ parameters for item 14 of the FMCE calculated from our dataset. These values are the average of 10 000 repeated analyses with randomly generated starting values. The final row is the fraction of our dataset that gave each response: $N_{k,j}$ is the number of students who chose response $k$ for item $j$, and $N$ is the total sample size (22 263).

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| $a_{k,14}$ | −0.35 | −2.07 | −2.10 | −3.28 | 1.73 | −2.58 | −2.34 | 0.77 |
| $d_{k,14}$ | 3.97 | 0.31 | 0.98 | −2.04 | 1.14 | −1.89 | −2.10 | 0.32 |
| $N_{k,14}/N$ | 0.585 | 0.035 | 0.075 | 0.013 | 0.243 | 0.007 | 0.005 | 0.021 |

select to determine their overall understanding of the test material. It is, therefore, useful to consider the probability of a student having a particular value of the latent trait $\theta$, given their selected responses to test items.

### A. Bock's nominal response model and Bayes' formula

Bayes' formula provides a method for determining our desired probability of a student's latent trait $\theta$ given a specific response $k$ to item $j$,

$$P_j(\theta|k) = \frac{P_j(k|\theta)P(\theta)}{P_j(k)}, \tag{4}$$

where $P_j(k|\theta)$ is the NRM probability function from Eq. (3), $P(\theta)$ is the probability distribution of the latent trait $\theta$, and $P_j(k)$ is the probability that a student will select response $k$ to item $j$ independent of $\theta$.

Equation (3) may be used to rewrite Eq. (4) as

$$P_j(\theta|k) = \frac{P(\theta)}{P_j(k)} \frac{e^{a_{k,j}\theta + d_{k,j}}}{\sum_{i=1}^{n} e^{a_{i,j}\theta + d_{i,j}}}. \tag{5}$$

A theoretical distribution for $P_j(k)$ is not known; however, because this does not depend on $\theta$, it does not affect the shape of $P_j(\theta|k)$ but only the scale. A common choice for $P(\theta)$ is the standard normal distribution, with a mean of zero and a standard deviation of 1 [33].

### B. An example from the FMCE

We can use a large dataset to determine the response-specific set of parameters $\{a_{k,j}, d_{k,j}\}$ using IRT NRM analyses. We may then use these parameters to explore how the selection of each response option for a particular item relates to the probability that a student has a particular level of understanding $\theta$. As an illustrative example, Table I shows the parameter values for each response option for item 14 on the FMCE. The data used to determine these parameters consisted of a combination of 22 263 pretest and post-test responses from 14 200 students. Of these, 6336 response sets were obtained from the PhysPort DataExplorer database [35], 6912 response sets were collected via the Learning About STEM Student Outcomes (LASSO) website [36], and 952 response sets were collected at four different colleges and universities from across the U.S. All IRT analyses were performed using the MIRT package in the R computing environment [37–39].

To get a sense of the variability in the estimated parameter values of the model, we used the MIRT function's option to generate random values for the initial parameter estimates (GenRandomPars = TRUE), and we repeated the analysis of each dataset 10 000 times [40]. The values in Table I are the average values of model parameters using these 10 000 runs.

Figure 1 shows the category characteristic curves (CCCs) for item 14 derived from Eq. (3), which shows $P_{14}(k|\theta)$, the probability of selecting a particular response, given a student's latent trait $\theta$. As expected, the probability of selecting the correct response E is a monotonically increasing function of $\theta$ that asymptotically approaches 1: a higher level of understanding is always associated with a higher probability of choosing the correct response. The most common response is A, shown by higher probabilities than any other responses over the middle range of $\theta$ values, and the least common responses are F and G. Response D is notable for having a distinctly negative slope for all values of $\theta$, indicating that the probability of choosing response D is inversely related to students' overall understanding. For a given value of $\theta$, the CCCs represent the probability of selecting a particular response category for an item. Thus, given $\theta$, the probabilities add to 1.

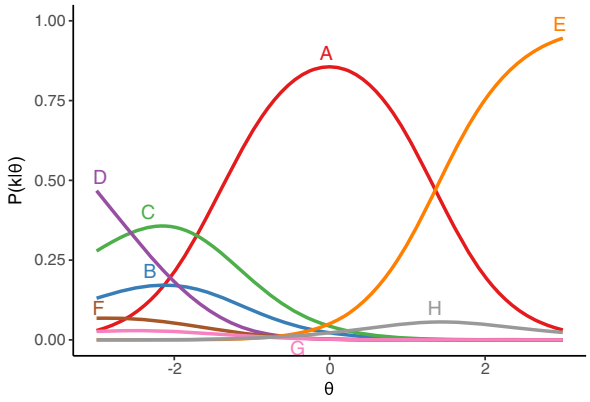


FIG. 1. IRT category characteristic curves showing the probability of selecting a particular response for item 14 of the FMCE, given a students' latent trait $\theta$. Curves were generated from Eq. (3) using the $\{a_{k,14}, d_{k,14}\}$ parameters from Table I.

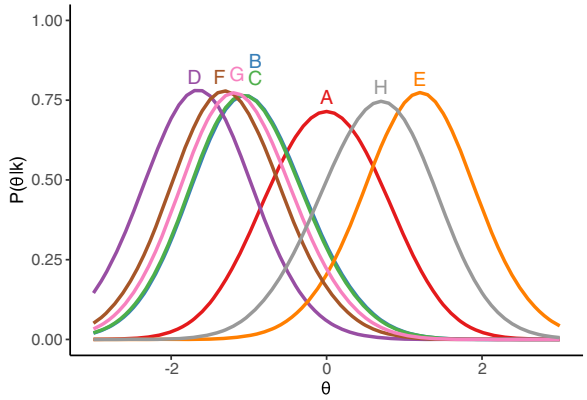


FIG. 2. Probability distributions for a student having a particular value of $\theta$, given that they chose one of the response options to item 14 on the FMCE. Curves were generated from Eq. (5) using the parameters from Table I, and assuming a standard normal distribution for $P(\theta)$ and values of $P_{14}(k)$ that normalize the probability distribution.

Using parameter estimates from Table I, and assuming a standard normal distribution for $P(\theta)$, Fig. 2 shows the probability density $P_{14}(\theta|k)$ from Eq. (5) for each response to item 14 on the FMCE, i.e., the probability of a student having a value of the latent trait $\theta$, given that they chose a particular response $k$. As expected, the probability distribution of $\theta$ given the correct response option E, $P_{14}(\theta|E)$, peaks at higher values of $\theta$ than any other distribution, with a mode at $\theta \approx 1.2$. This means that students who choose response E are more likely to have higher understanding of the topic being measured by the test item than students who choose any other response. Conversely, students who select response D are more likely to have lower values of $\theta$ than students who choose any other responses, with the peak of the $P_{14}(\theta|D)$ distribution occurring around $\theta \approx -1.7$.

Comparing Fig. 2 and Table I shows us that the order of the central peaks of each $P_{14}(\theta|k)$ follows the order of the set of $\{a_{k,14}\}$ values. The order of the set of $\{a_{k,14}\}$ as displayed in Table I in decreasing order is E, H, A, B, C, G, F, and D. The same analysis was performed for the remaining items on the FMCE and found that the order of the modes of the $\theta$ distributions for each response category follows that same order of the $a_{k,j}$ values estimated by the nominal response model (data not shown). Examining the order of modes of the $P_j(\theta|k)$ using the FMCE data for each response category supports the idea of using the $a_{k,j}$ parameters to rank incorrect responses.

### C. Relating parameter values to the probability density of the latent trait

To further explore the relationship between the value of $a_k$ and the location of the maximum of $P(\theta|k)$, we use Eq. (5) to identify the location of the peak for each curve $\theta_{pk}$. For simplicity of notation, we have removed the label "$j$" from the following expressions with the understanding that parameter values are specific for each item. At the peak, the slope of each curve must be zero:

$$\left.\frac{\partial P(\theta|k)}{\partial\theta}\right|_{\theta_{pk}} = 0. \tag{6}$$

The derivative of the expression in Eq. (5) is

$$\frac{\partial P(\theta|k)}{\partial\theta} = P(\theta|k)\left[\frac{P'(\theta)}{P(\theta)} + a_k - \sum_{r=1}^{n} a_r P(r|\theta)\right], \tag{7}$$

where the summation index in the last term has been changed to $r$ to avoid confusion with the undetermined index $k$. (Additional details of this mathematical derivation are included in the Supplemental Material [41].) The last expression of Eq. (7) is the (weighted) average value of $a_r$ as a function of $\theta$. Equation (7) can be rewritten as

$$\frac{\partial P(\theta|k)}{\partial\theta} = P(\theta|k)\left[\frac{P'(\theta)}{P(\theta)} + a_k - \langle a_r\rangle_{(\theta)}\right]. \tag{8}$$

Putting the results from Eq. (8) into Eq. (6) yields the relationship

$$a_k + \frac{P'(\theta_{pk})}{P(\theta_{pk})} - \langle a_r\rangle_{(\theta_{pk})} = 0, \tag{9}$$

$\therefore$

$$a_k = \langle a_r\rangle_{(\theta_{pk})} - \frac{P'(\theta_{pk})}{P(\theta_{pk})}. \tag{10}$$

To explore the behavior of $\langle a_r\rangle_{(\theta)}$, consider the derivative of $\langle a_r\rangle_{(\theta)}$ with respect to $\theta$.

$$\frac{\partial\langle a_r\rangle}{\partial\theta} = \frac{\partial}{\partial\theta}\left(\sum_{r=1}^{n}\frac{a_r e^{a_r\theta+d_r}}{\sum_{i=1}^{n} e^{a_i\theta+d_i}}\right) \tag{11}$$

$$= \langle a_r^2\rangle_{(\theta)} - \langle a_r\rangle_{(\theta)}^2 \tag{12}$$

$$= \langle (a_r - \langle a_r\rangle_{(\theta)})^2\rangle. \tag{13}$$

This is the variance of $a_r$ as a function of $\theta$, which must be positive. (Additional details of this mathematical derivation are included in the Supplemental Material [41]). This implies that $\langle a_r\rangle_{(\theta)}$ is a monotonically increasing function of $\theta$.

Most IRT models developed for educational testing purposes assume that the latent trait has a standard normal distribution for the purposes of estimating the model parameters [42–44]. Under this assumption of normality,

$$\frac{P'(\theta)}{P(\theta)} = -\theta, \tag{14}$$

and Eq. (10) becomes

$$a_k = \langle a_r \rangle_{(\theta_{pk})} + \theta_{pk}. \tag{15}$$

This implies that higher locations of the peak value $\theta_{pk}$ are directly associated with higher values of $a_k$. Given that higher peak $\theta_{pk}$ values are associated with higher levels of understanding (and vice versa), this provides a mathematical argument for using the $a_k$ parameters as a proxy measure of students' understanding. This, in turn, justifies using the $a_k$ parameters to rank responses.

Figure 2 shows us that the distribution of each $P_{14}(\theta|k)$ is fairly broad. Consequently, a student's response to a single item provides an imprecise measurement of $\theta$. In order to increase this precision, a student's responses to many items should be considered together: generally, the maximum likelihood method is used to estimate a student's ability $\theta$ using the IRT model and student responses to test items. This is consistent with common practice, in that multiple-choice assessments typically contain many items for measuring student knowledge and understanding.

## III. A PHYSICIST'S INTERPRETATION OF THE PARAMETERS IN THE NOMINAL RESPONSE MODEL

In the previous section, we presented a mathematical argument showing that the $a_k$ parameters from Bock's NRM are associated with a level of understanding represented by the latent trait $\theta$, further supporting the use of the $a_k$ parameters to rank item responses for polytomous data. In this section, we present an interpretation of these parameters by analogy to common thermodynamic systems, allowing us to better understand the relationship between $a_k$ and $\theta$. We also provide an argument for why the $d_k$ parameters are not useful for ranking response options.

### A. The canonical ensemble

We begin our physical analogy by recognizing that the denominator of the NRM probability expression in Eq. (3) can be considered a partition function. This denominator acts as a normalizing factor that is independent of the selected response $k$ but ensures that the sum of the probabilities over all response options is unity for every value of $\theta$. One of the simplest physical systems that involves partition functions is the canonical ensemble, which describes a generic thermodynamic system that has a fixed temperature and variable energy. Additional details for why we have chosen the canonical ensemble for our analogy rather than another model can be found in the Supplemental Material [41]. The probability of the system being in a particular *microstate* $m$ may be written as

$$P(m|T) = \frac{e^{-E_m/k_BT}}{\sum_{\ell=1}^{N} e^{-E_\ell/k_BT}} \tag{16}$$

TABLE II. Macrostates for a paramagnet with three distinguishable spin-1/2 particles in a magnetic field of strength $H$. Each spin has a magnetic moment with magnitude $\mu$. For the microstates, ↑ indicates that the spin is aligned with the field $H$ (spin up), and ↓ indicates that the spin is antialigned (spin down).

| Net magnetization | Energy | Degeneracy | Microstates |
|---|---|---|---|
| $+3\mu$ | $-3\mu H$ | 1 | ↑↑↑ |
| $+\mu$ | $-\mu H$ | 3 | ↓↑↑<br>↑↓↑<br>↑↑↓ |
| $-\mu$ | $+\mu H$ | 3 | ↓↓↑<br>↑↓↓<br>↓↑↓ |
| $-3\mu$ | $+3\mu H$ | 1 | ↓↓↓ |

or

$$P(m|\beta) = \frac{e^{-E_m\beta}}{\sum_{\ell=1}^{N} e^{-E_\ell\beta}}, \tag{17}$$

where the summation in the denominator is performed over all $N$ accessible microstates of the system, and $k_B = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant. Equation (17) is expressed in terms of the commonly used temperature parameter $\beta \equiv 1/k_BT$.

The model of the probability of a system being in a particular microstate in equations Eqs. (16) and (17) looks very similar (but not identical) to the IRT NRM presented in Eq. (3). In order to define a useful analogy between the two models, we can investigate similarities in the behavior of systems that are described by either the NRM or the canonical ensemble. To illustrate these similarities, we will use a paramagnetic system consisting of several noninteracting spin-1/2 particles in an external magnetic field $H$. Each particle can be either aligned with the field (spin up) or antialigned with the field (spin down). Any particle that is spin up has a positive magnetization $+\mu$ and a negative energy $-\mu H$, where $\mu$ is the magnitude of the magnetic moment of the particle. Conversely, particles that are spin down each have negative magnetization $-\mu$ and positive energy $+\mu H$. Table II shows the properties of the four macrostates of a paramagnet consisting of three distinguishable spin-1/2 particles, along with the associated microstates (defined by spin configuration). The probability of a paramagnetic system being in a particular microstate can be modeled using Eqs. (16) and (17), where the partition function in the denominators has eight terms for the 3-spin paramagnet: one for each of the microstates shown in Table II. We can also rewrite Eq. (17) to, instead, express the probability of the system being a in a particular *macrostate* $k$ as

$$P(k|\beta) = \frac{g_k e^{-E_k\beta}}{\sum_{i=1}^{n} g_i e^{-E_i\beta}}, \tag{18}$$

where $g_k$ is the degeneracy of (a.k.a., the number of microstates associated with) macrostate $k$, and $n$ is the total number of macrostates.

In the following sections, we relate the physical quantities in Eqs. (16)–(18) with the NRM variables in Eq. (3) by considering the behavior of the 3-spin paramagnet at low temperatures, as well as the implications of the negative sign in the canonical ensemble, which is conspicuously absent in the NRM.

### B. The low-temperature limit

Consider the behavior of the canonical ensemble in the extremes of the continuous variable $T$. Equation (16) exhibits the well-known phenomenon that the state with the lowest energy (a.k.a. the ground state) has the highest probability in the low-temperature limit because the summation in the denominator is dominated by the term associated with the lowest-energy state as $T \to 0$.

Similarly, one response dominates the probability distribution of the NRM in Eq. (3) in an extreme limit of $\theta$, but in this case the limit is reversed: the probability of a student selecting the response with the highest $a_k$ value (i.e., the correct response) approaches one as $\theta \to +\infty$. This can be seen in Fig. 1 with the probability of the correct response E nearly reaching one at $\theta = 3$ for item 14 on the FMCE. The similarity between the canonical ensemble and the NRM is that, at an extreme value of the continuous quantity ($T$ or $\theta$) the probability of one particular value of the discrete quantity ($E_k$ or $a_k$) approaches one, with the probabilities of all other values of the discrete quantity (necessarily) tending to zero. The difference is that in the canonical ensemble the lowest value of $E_k$ dominates the probability as $T \to 0$, and in the NRM the highest value of $a_k$ dominates the probability as $\theta \to +\infty$.

To address one of these differences, we may use Eq. (17) or (18) with the temperature parameter $\beta$. In this case, the low-temperature limit occurs at $\beta \to +\infty$; therefore, the lowest value of $E_k$ dominates the probability distribution as $\beta \to +\infty$, which is more similar to the behavior of the NRM. We can now say that $\theta$ is analogous to the temperature parameter $\beta$, and that $a_k$ is analogous to the system energy $E_m$ or $E_k$ [45]. Rewriting the NRM probability from Eq. (3) as

$$P(k|\theta) = \frac{e^{d_k} e^{a_k\theta}}{\sum_{i=1}^{n} e^{d_i} e^{a_i\theta}} \tag{19}$$

highlights the similarities with Eq. (18) by separating an exponential term that includes $\theta$ from one that does not. Equation (19) also shows that the exponential term $e^{d_k}$ has a mathematical similarity to $g_k$ in Eq. (18); however, as we discuss more fully below (see Sec. III G), we do not make a direct connection between $d_k$ and $g_k$.

### C. Negative temperature

As $\theta \to +\infty$ in the NRM, the probability of the highest-$a_k$ response tends toward one. Conversely, as $\theta \to -\infty$, the probability of the lowest-$a_k$ response tends toward one (e.g., the curve for response D is a monotonically decreasing function in Fig. 1). These properties of the NRM suggest that, to complete our analogy, we need to consider both positive and negative values of the temperature parameter, and we need to be able to define values of $\beta$ for which the highest-energy state is most probable. Mathematically, the highest-energy state must dominate the probability distribution as $\beta \to -\infty$, which is the same as $T \to 0^-$.

To better understand the notion of negative temperature, consider the definition of temperature,

$$T \equiv \left(\frac{\partial U}{\partial S}\right)_{V,N,\ldots} \tag{20}$$

where $U$ is the average thermodynamic energy of the system, $S$ is the entropy of the system, and $\{V, N, \ldots\}$ are any other extensive quantities of the system that are held fixed. If $T < 0$, then a decrease in entropy $S$ must be accompanied by an increase in average energy $U$ (and vice versa). Purcell and Pound experimentally created a system with negative temperature by preparing a lithium flouride crystal system with a very low (positive) temperature in the presence of a magnetic field, and then rapidly reversing the direction of the field [46]. Before the field reversal, the system had most particles aligned with the field and was in a state with an average energy near the minimum value and a very low entropy. Immediately after the field reversal, the spins that had previously been aligned with the field were then antialigned, creating a state with a very large (and positive) average energy, but still a very low entropy. From this point, allowing some particles to switch orientation to be aligned with the field simultaneously increased the entropy of the system while decreasing its average energy, resulting in a negative temperature. Hakonen and Lounasmaa also discussed this phenomenon, stating, "A negative temperature describes a state with population inversion where the higher energy levels have more particles than the lower levels" [47]. This population inversion means that negative (absolute) temperatures may be thought of as actually being "higher" than positive infinity with "the average energy per system [being] larger… than the mid-energy of the available levels" [46].

According to deBoer, the "hottest" systems are at $\beta \to -\infty$, and the "coldest" systems are at $\beta \to +\infty$ [48]. If we consider the largest positive values of $\theta$—associated with the highest levels of understanding of the tested topic—to be analogous to the hottest systems, then

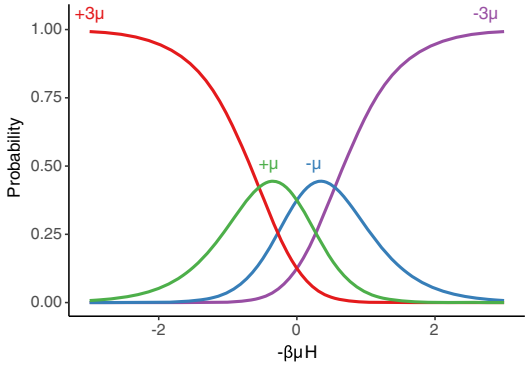


FIG. 3. Probability distributions for the 3-particle spin-1/2 paramagnet as a function of the temperature parameter $-\beta$. The horizontal axis has been scaled by $\mu H$ to obtain a dimensionless quantity. Because of the negative sign, colder temperatures are on the left and hotter temperatures are on the right. Macrostates are labeled by the net magnetization.

we must associate the latent trait $\theta$ with the negative of the temperature parameter $-\beta$. Likewise, the largest negative values of $\theta$—associated with the lowest levels of understanding—are analogous to the coldest systems.

Figure 3 shows the probability distributions for the four macrostates of the 3-particle spin-1/2 paramagnet as a function of $-\beta$ (scaled by $\mu H$ to create a dimensionless quantity). There are many similarities between the probability curves for the canonical ensemble in Fig. 3 and the NRM curves in Fig. 1: the curve associated with the highest $E_k$ or $a_k$ value approaches one on the right side of the plot, the curve associated with the lowest $E_k$ or $a_k$ value approaches one on the left side of the plot, and the curves with neither the highest nor the lowest $E_k$ or $a_k$ values each have a maximum at a single value of $-\beta$ or $\theta$, respectively.

### D. Summarizing the analogy

We are now in a position to propose an analogy between the IRT NRM and the canonical ensemble:

$$\theta \rightarrow -\beta, \tag{21}$$

$$a_k \rightarrow E_k. \tag{22}$$

In this interpretation, each student may be thought of as a knowledge reservoir with a specific "temperature" (related to $\theta$). Each item on an assessment is a system that has various macrostates identified with the specific response options, and each of these macrostates has an associated "energy" $a_k$. Each "item system" interacts with the student reservoir, and the probability of finding the system in each of its macrostates depends on the energy of each macrostate, as well as the temperature of the student.

The process of administering a test can then be seen as indirectly measuring the temperature of the reservoir (the knowledge level of the student) by directly measuring the macrostates of multiple systems (responses to items) that are all in thermal contact with the same reservoir (student) [49]. The connection between the $k$th macrostate and student knowledge is determined by the $a_k$ and $d_k$ parameters for each item. Treating individual incorrect responses as being representative of different values of $\theta$ provides a much more precise estimate of each student's actual level of understanding than does treating all incorrect responses as being equivalent by scoring items dichotomously.

The biggest difference between the probability distributions for the paramagnet shown in Fig. 3 and the IRT probability distributions in Fig. 1 is that the IRT plots are not symmetric about any particular value of $\theta$. The symmetry of Fig. 3 is due to the fact that the energy values of the macrostates for the paramagnet are evenly spaced with $\Delta E = 2\mu H$ for adjacent energy levels. The $a_k$ energy values are not necessarily evenly spaced, and both the range of $a_k$ values and the differences between adjacent levels may be different for each item.

Our analogy includes a connection between parameters associated with the various macrostates of a system ($E_k$ and $a_k$), as well as a connection between quantities related to the system's surrounding environment ($-\beta$ and $\theta$). One piece that is missing from our analogy is a connection between the other macrostate-specific parameters: $d_k$ in the NRM, and $g_k$ in the canonical ensemble. Comparing Eqs. (18) and (19) shows that the quantity $e^{d_k}$ serves a similar mathematical function as $g_k$, acting as a statistical weight for the exponential term containing the energy parameter; however, it would be inappropriate to consider $e^{d_k}$ analogous to degeneracy. In fact, the concept of degeneracy cannot be defined within this analogy. We discuss this limitation of our analogy in more detail is Sec. III G.

### E. Ranking responses by energy

Considering the values of the energy of each macrostate in the canonical ensemble is necessary for distinguishing between different temperature regimes. If lower energy states are more probable than higher energy states, then the temperature is positive; if higher energy states are more probable than lower energy states, then the temperature is negative. Because negative temperatures are considered hotter than positive temperatures, the higher energy states are associated with hotter temperatures, and the lower energy states are associated with colder temperatures. In our IRT analogy, this means that higher values of $a_k$ are associated with higher understanding, and lower values of $a_k$ are associated with lower understanding, which is exactly consistent with the psychometric literature [30,31] and our mathematical arguments in Sec. II [50]. The probability of the highest-$a_k$ response (the correct response) approaches one as $\theta \rightarrow +\infty$, and the probability of the lowest-$a_k$ response approaches one as $\theta \rightarrow -\infty$. In the 3-spin paramagnet example, we order the macrostates according to their energies; likewise, in the NRM it makes

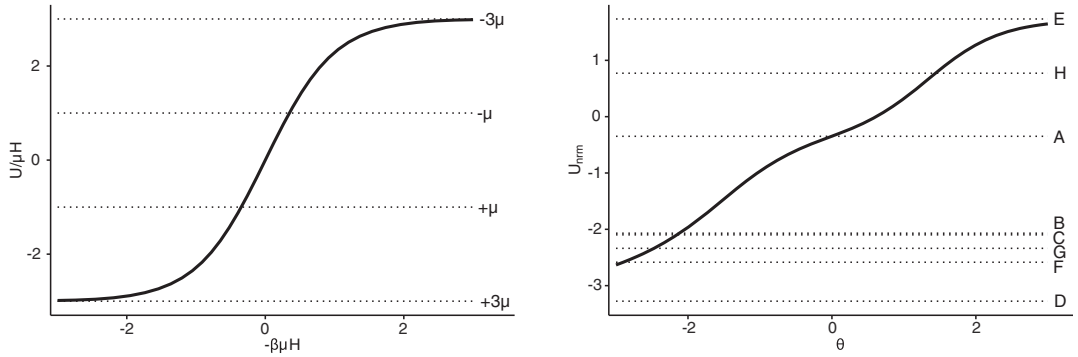


FIG. 4. (a) Average energy $U$ for the three-particle spin-1/2 paramagnet, plotted as a function of the temperature parameter $-\beta$. All axes have been scaled to obtain dimensionless quantities. The dotted lines show the energy value for each macrostate, labeled by magnetization. (b) The IRT equivalent of average energy $U_{\text{nrm}}$ plotted as a function of the latent trait $\theta$ for item 14 of the FMCE. The dotted lines show the $a_{k,14}$ energy value for each response option.

sense to order the response options according to their $a_k$ values.

We can also see the impact of the $a_k$ values in Fig. 1. At the lowest values of $\theta$ shown ($\theta = -3$, left side of Fig. 1) the response with the lowest $a_k$ value is the most probable: $a_D = -3.28$. As $\theta$ increases, the probability of selecting response D declines sharply. The curves for responses F and G appear to have maxima at or near the left edge of the plot, and those probabilities also decline with $\theta > -3$; this is consistent with the values of $a_F = -2.58$ and $a_G = -2.34$ being the next lowest. The curves for responses A, B, and C all have distinctly positive slopes at $\theta = -3$, but, while the probabilities for B and C start declining for $\theta \gtrsim -2$, the probability of selecting response A increases until $\theta \approx 0$. This is consistent with $a_A > a_B \approx a_C$ ($a_A = -0.35$, $a_B = -2.07$, $a_C = -2.10$). Finally, although the curves for responses E and H have very small slopes at $\theta = -3$, we can see that the probabilities of each of these increase for higher values of $\theta$. The probability of selecting response H does not peak until $\theta \approx +1.5$—consistent with H having the second-highest $a_k$ value, $a_H = +0.77$. The correct response E has the highest $a_k$ value ($a_E = +1.73$), and, of course, its probability curve increases over the entire range of $\theta$, asymptotically approaching one. Based on these $a_k$ values, we would rank the responses to item 14 as E > H > A > B ≈ C > G > F >D.

Given that students in our analyses are in the range $-3 < \theta < +3$ (with $\theta$ scaled to have a mean of 0 and a standard deviation of 1), and the fact that FMCE items have up to nine response options, we rarely see the probability of the lowest-$a_k$ option go above 0.8 in our IRT plots for any item, but in all cases the probability of the correct response approaches unity at $\theta \approx +3$.

In order to justify our analogy between the IRT NRM and the canonical ensemble, we have applied the somewhat unfamiliar context of negative absolute temperature; however, the end result of our analogy does not require a thorough understanding of negative temperatures. The main result is the association of the $a_k$ parameter with energy. As mentioned above, by interpreting $a_k$ as the energy of response $k$, the $a_k$ values may be used to rank the responses to each item; this is consistent with the results of our mathematical arguments in Sec. II.

### F. Extending the analogy

Having a well-defined analogy between the NRM and the canonical ensemble also allows us to define a quantity analogous to the thermodynamic average energy $U$ of a system. The average energy for the canonical ensemble is simply the weighted average of the energy of each macrostate of the system at a given temperature,

$$U(\beta) = \sum_{k=1}^{n} E_k P(k|\beta) = \sum_{k=1}^{n} E_k \left( \frac{g_k e^{-E_k \beta}}{\sum_{i=1}^{n} g_i e^{-E_i \beta}} \right). \tag{23}$$

Figure 4(a) shows a plot of the average energy of the 3-particle spin-1/2 paramagnet as a function of the negative of the temperature parameter $-\beta$. At the coldest temperatures, the average energy approaches the value of the lowest-energy macrostate: $\lim_{-\beta \to -\infty} U = -3\mu H$. As temperature increases, the average energy smoothly increases until, at the hottest temperatures, it approaches the value of the highest-energy macrostate: $\lim_{-\beta \to +\infty} U = +3\mu H$.

With our thermodynamic interpretation, the $\langle a_r \rangle_{(\theta)}$ term in Eq. (10) from Sec. II C can be interpreted as the average energy of the item "system" at a given value of $\theta$,

$$U_{\text{nrm}}(\theta) = \langle a_k \rangle_{(\theta)} = \sum_{k=1}^{n} a_k \left( \frac{e^{a_k \theta + d_k}}{\sum_{i=1}^{n} e^{a_i \theta + d_i}} \right). \tag{24}$$

Figure 4(b) shows an example plot of $U_{\text{nrm}}(\theta)$ for item 14 of the FMCE. As expected, The average NRM energy $U_{\text{nrm}}(\theta)$ is a monotonically increasing function of the

temperature parameter $\theta$, which is exactly what we found in Eq. (13). At the extremes of the $\theta$ distribution (either positive or negative), the probability is dominated by a single response (see Fig. 1). As with the paramagnet example, the average energy in each of these limits then becomes the $a_k$ value for that specific response. For large positive values, the correct response E dominates the probability, and for large negative values, this is the response D, which has the lowest $a_k$ value.

$$\lim_{\theta\to+\infty} U_{\mathrm{nrm}}(\theta) = a_{E,14} = +1.73, \tag{25}$$

$$\lim_{\theta\to-\infty} U_{\mathrm{nrm}}(\theta) = a_{D,14} = -3.28. \tag{26}$$

Figure 4(b) also contains dotted lines showing the $a_k$ energy value for each of the response options to item 14. The value of $\theta$ where the average energy $U_{\mathrm{nrm}}$ crosses each dotted line is the same as the value of $\theta$ for which each IRT probability curve is maximized in Fig. 1. This suggests that the location of the maximum of each IRT curve may also be used to rank responses to an item.

### G. Limitations of the analogy

As mentioned in Sec. III D, our analogy does not provide a complete connection between the $d_k$ parameter in the NRM and the degeneracy $g_k$ in the canonical ensemble. These parameters can be compared mathematically—the quantity $e^{d_k}$ and $g_k$ both act as a statistical weight—but the meaning of these parameters is quite different. To explore this more, consider the zero point for the continuous quantities $\theta$ and $\beta$. A situation in which the temperature parameter $\beta = 0$ represents infinite temperature ($T \to \pm\infty$). In this high-$T$ limit, all microstates are equally probable, and the probability of a given macrostate depends only on the degeneracy

$$\lim_{\beta\to0} P(k|\beta) = \frac{g_k}{\sum_{i=1}^{n} g_i}. \tag{27}$$

A similar limit can be taken for Eq. (19) to obtain the probability of choosing a response $k$ given an average ability $\theta = 0$,

$$P(k|0) = \frac{e^{d_k}}{\sum_{i=1}^{n} e^{d_i}}. \tag{28}$$

However, the meaning of the $\theta = 0$ case is very different from the $\beta = 0$ case. In IRT analyses, the zero point for $\theta$ is an arbitrary choice, and choosing an anchoring value and a scale for $\theta$ is an important aspect of both performing and interpreting IRT analyses [51]. As mentioned in Sec. II, a common choice in IRT analyses is to set the mean of the $\theta$ distribution for a given population to zero, and the standard deviation to one.

Consider the effect of choosing a different anchor and scale such that the distribution of the rescaled quantity $\theta^*$ is assumed to have a mean of $\varepsilon$, and a standard deviation of $\gamma$. This can be accomplished by defining

$$\theta^* \equiv \gamma\theta + \varepsilon. \tag{29}$$

This rescaled latent trait would result in different sets of parameters $\{a_k^*, d_k^*\}$ as well. We can express these new parameters in terms of the original parameters by expressing the probability function from Eq. (3) in terms of $\theta^*$:

$$P(k|\theta) = \frac{e^{a_k\theta+d_k}}{\sum_{i=1}^{n} e^{a_i\theta+d_i}} \tag{30}$$

$$= \frac{e^{a_k(\theta^*-\varepsilon)/\gamma+d_k}}{\sum_{i=1}^{n} e^{a_i(\theta^*-\varepsilon)/\gamma+d_i}}. \tag{31}$$

Each exponent may be rewritten by distributing the $1/\gamma$ factor to explicitly show which terms include $\theta^*$ and which do not.

$$\frac{a_k}{\gamma}(\theta^* - \varepsilon) + d_k = \frac{1}{\gamma}a_k\theta^* - \frac{\varepsilon}{\gamma}a_k + d_k. \tag{32}$$

We can now define new parameters

$$a_k^* \equiv \frac{1}{\gamma}a_k, \tag{33}$$

$$d_k^* \equiv d_k - \frac{\varepsilon}{\gamma}a_k, \tag{34}$$

such that

$$P(k|\theta^*) = \frac{e^{a_k^*\theta^*+d_k^*}}{\sum_{i=1}^{n} e^{a_i^*\theta^*+d_i^*}}. \tag{35}$$

Under this rescaling, the $a_k^*$ energy parameters are simply stretched by a factor of $1/\gamma$. This is equivalent to expressing energy in different units (e.g., ergs vs joules): the relative order and spacing of the set of $\{a_k^*\}$ is the same as it was for $\{a_k\}$. Therefore, rankings of responses defined by $\{a_k^*\}$ would be the same as those defined by $\{a_k\}$, and the shape of a plot of $\langle a_k^*\rangle$ would be exactly the same as Fig. 4(b), with only the values on the axes changing. Each $d_k^*$ parameter, however, is shifted from $d_k$ by an amount that depends on the value of the associated $a_k$ parameter for that response. As such, responses with high $a_k$ values will be shifted differently from responses with low $a_k$ values. Depending on the choice of $\varepsilon$, this could result in the order of the set of $\{d_k^*\}$ values being completely different than the order of the set of $\{d_k\}$ values.

For example, in Fig. 1, the $\theta$ distribution is scaled to have a mean of zero. The values of the probability curves at

$\theta = 0$ give us information about the relative distribution of $d_k$ values: response A is the most probable (by far), so $d_A$ is the highest; responses C and E are about equally probable; responses B and H are also about equally probable, but lower than C and E; responses D, F, and G are so low that they are difficult to distinguish. This is consistent with the $d_{k,14}$ values in Table I: $d_A > d_C \approx d_E > d_B \approx d_H > d_F > d_D \approx d_G$. If, however, we rescale $\theta^*$ by setting $\varepsilon = 2$, then all of the values on the horizontal axis in Fig. 1 will increase by two; therefore, the location of $\theta = -2$ in Fig. 1 will be the new zero point for $\theta^*$. At this location, response C is the most probable, and responses A and D are approximately equal, followed closely by response B. This would suggest an order for the $\{d_k^*\}$ of $d_C^* > d_A^* \approx d_D^* > d_B^* > d_F^* > d_G^* > d_E^* \approx d_H^*$. This order is clearly very different from what is shown in Table I, highlighting the fact that the order of the set of $\{d_k\}$ parameter values is not invariant under translation of $\theta$. This provides a strong rationale for ignoring the $d_k$ parameters when ranking responses.

As with all analogies, there are some features of the reference model that are not applicable to the target model. The shifting of the order of the $\{d_k^*\}$ values highlights the fact that the concepts of microstates and degeneracy are not applicable to the NRM: the degeneracy of a macrostate in the canonical ensemble could never be affected by an arbitrarily defined reference value. This suggests that thermodynamic quantities that rely on those concepts (e.g., entropy) do not have analogs in the NRM. This disconnect hinges on the meaning of negative values in the NRM and the canonical ensemble. White Brahmia *et al.* describe the many different physical interpretations of negative quantities [52]. The model of temperature defined in Eq. (20) establishes a distinct difference between positive and negative values of $\beta$: negative energy microstates are more probable in systems with $\beta > 0$, and vice versa for systems with $\beta < 0$. In contrast, there is nothing inherently different about positive and negative values of $\theta$ because $\theta$ is defined relative to an arbitrary reference. Though mathematically similar, this difference in the meaning of negative quantities is crucial to recognize when applying our analogy between the canonical ensemble and the NRM. The useful similarity between the models is that higher values of $\theta$ represent greater knowledge or understanding, just like higher values of $-\beta$ represent hotter systems, and lower values of $\theta$ represent less knowledge or understanding, just like lower values of $-\beta$ represent colder systems.

Equation (20) provides a theoretical definition of $T$ (and, by extension, $\beta$), but no such definition exists for $\theta$. In practice, this means that an underlying distribution for $\theta$ must be assumed within a population to be able to estimate item parameters, and then calculate the latent trait for each student. Conversely, macrostate parameters in the canonical ensemble may often be defined by theoretical models of microstates (like those shown in Table II). Even when a specific physical model is not known for a given system, the temperature of a system (and its surroundings) may often be measured directly; therefore, there is no need to assume an underlying distribution of temperatures within a collection of thermodynamic systems.

We believe strongly that the analogy that we have developed throughout Sec. III is very useful for relating responses to test items to macrostates of a physical system, but we think it is important to acknowledge the limitations of this analogy. Not all properties of systems described by the canonical ensemble are applicable to the NRM. We believe that the strongest feature of the analogy is the interpretation of the $a_k$ parameter as the energy associated with a particular response option, and higher-energy states (responses with higher $a_k$) being associated with hotter systems (students with higher values of $\theta$). We do not attempt to define a theoretical definition of $\theta$ similar to Eq. (20), nor do we attempt to define what might be equivalent to a "microstate" in the NRM. Care must be taken when extending this analogy to ensure that all claims are well justified.

## IV. CONSIDERATIONS FOR APPLYING IRT ANALYSES

There are several aspects of IRT analyses that must be considered before applying any of the methods described above. These considerations do not directly affect our results, but they are important for anyone who may want to apply our results to their own work. One consideration for using the NRM to rank responses to multiple-choice items is the need for fairly large datasets to be able to obtain reliable estimates of the parameter values. A general guideline is to have at least 10 response sets for every parameter being calculated [53]: this would be a minimum data set of 7220 respondents for the FMCE and 3000 for the FCI, which is significantly higher than typical classes. In order to apply the results of our work in typical classroom settings, the IRT parameter values would first need to be established using a large representative dataset. The parameters could then be used to identify higher and lower incorrect responses for use in analyzing smaller datasets for instructional or research purposes.

Another consideration is that all items on the test must function properly: that is, increasing the latent trait $\theta$ must result in an increased probability of choosing the correct response for all items. In practice, items on well-designed tests tend to function properly, but it is conceivable that a single misplaced item could have a correct response that is anticorrelated with the correct responses to the other items on the test. Any analyses of such items should be subject to scrutiny. Additionally, assessments like the FMCE and FCI have been shown to be multidimensional, suggesting that they are simultaneously measuring understanding of several different topics [7,8]. A unidimensional test (or subset of items) may be necessary to uniquely define rankings.

Ideally, a ranking of responses for each item could be determined that is consistent across multiple datasets and be shown to be independent of student population. These rankings could then be applied to analyze data from any student population to measure and represent student understanding and learning in more nuanced ways than are available with traditional dichotomous scoring methods.

## V. CONCLUSIONS

We have presented a mathematical argument for using the $a_k$ parameters from Bock's nominal response model for ranking responses to multiple-choice items based on the relationship between the $a_k$ parameter and the value of $\theta$ that maximizes the conditional probability of $\theta$ given a response option $k$. This conditional probability distribution expresses the overall goal of using multiple-choice assessments: using students' chosen responses to determine their overall knowledge and understanding of the tested material. We also presented an analogy between Bock's NRM and the probability of a system being in a particular macrostate at a given temperature for the canonical ensemble that shows a direct correspondence between the $a_k$ parameter and the energy of a system with fixed temperature. Both the mathematical argument and the physical interpretation support the claims of Wainer *et al.*, and Bock and Moustaki that the value of the $a_k$ parameter may be used to rank the incorrect responses, with a higher $a_k$ value indicating a response that is more closely related with higher values of the latent trait $\theta$ and, therefore, better than a response with a lower $a_k$ value [30,31]. These arguments also support the results of Smith, Louis, Ricci, and Bendjilali [10], and Eaton, Johnson, and Willoughby [11] who both used the $a_k$ parameters in the nested-logit 2PL-NRM to rank incorrect responses.

Our physical analogy also highlights the need to consider differences in parameters rather than their absolute values. In systems with a finite number of macrostates, no value of energy can be defined as universally high or low. The zero point of the energy scale can be arbitrarily set without changing any of the mathematical or physical relationships. As such, the $a_k$ parameters must be compared within each item to determine which are higher and lower. A single parameter value is meaningless without the context of the other parameter values for the same item. Additionally, parameter values cannot be directly compared between items, just as the energies of the macrostates of a paramagnetic system in an external field $H$ cannot be directly related to the energies of the macrostates of a similar system in an external field of strength $3H$. Both of these systems would have macrostates with energies $\pm 3\mu H$, but they would be very different: in the original system with magnetic field $H$, the $\pm 3\mu H$ macrostates each have one associated microstate with all spins in the same direction, but in the $3H$-field system the $\pm 3\mu H$ macrostates would each have three associated microstates that each have 1 or 2 spins in each direction.

The physical analogy presented above may also be used to interpret the process of administering a multiple-choice test. By administering a test like the FMCE, instructors and researchers typically want to measure a student's overall understanding of a particular topic, which is represented by the IRT person parameter $\theta$. Using the physical analogy from Sec. III we can think of each student as a thermal reservoir, and the process of testing as a way of (indirectly) measuring temperature. Each item on the test can be considered a system with a set of states (response options), with the correct response option corresponding to the macrostate with the highest energy value. A student responding to an item is analogous to placing the item system in thermal contact with that student's thermal knowledge reservoir: there will be an associated probability of the system being found in each macrostate, related to energy of the states (mitigated by a statistical weight), and the temperature of the student reservoir. Measuring the macrostates of multiple different item systems in contact with the same student reservoir (e.g., using 47 items on the FMCE) allows us to indirectly measure its temperature (knowledge) based on the macrostate of each system. Dichotomous scoring is akin to only measuring whether or not each system is in its highest-energy macrostate. This gives an imprecise measure of temperature. Measuring energy for each macrostate provides a more complete picture of the temperature of the student reservoir; therefore, treating incorrect responses as having different values based on their energy could provide more complete information about each student's level of understanding.

IRT analyses of large datasets serve the purpose of identifying the energy associated with each response to each item. Once these parameters are well established, they can be used to inform more detailed measurements of students' knowledge and understanding by considering all responses rather than just those that are correct. Additional research will be required to establish the full set of energy values for both correct and incorrect responses, and to relate these energy values to a useful metric for representing students' levels of knowledge and understanding.

## ACKNOWLEDGMENTS

We thank Paul Bergeron and Samuel Lofland for helpful conversations about the interpretations of the IRT parameters. We also thank our very thoughtful referees for challenging us to think more deeply about our analogy. This project was supported by the National Science Foundation through Grant No. DUE-1836470.

[1] Y.-J. Lee, D. J. Palazzo, R. Warnakulasooriya, and D. E. Pritchard, Measuring student learning with item response theory, Phys. Rev. ST Phys. Educ. Res. **4**, 010102 (2008).

[2] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).

[3] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).

[4] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).

[5] C. S. Wallace, T. G. Chambers, and E. E. Prather, Item response theory evaluation of the Light and Spectroscopy Concept Inventory national data set, Phys. Rev. Phys. Educ. Res. **14**, 010149 (2018).

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

[7] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart, Multidimensional item response theory and the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010137 (2018).

[8] J. Yang, C. Zabriskie, and J. Stewart, Multidimensional item response theory and the Force and Motion Conceptual Evaluation, Phys. Rev. Phys. Educ. Res. **15**, 020141 (2019).

[9] C. Zabriskie and J. Stewart, Multidimensional Item Response Theory and the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **15**, 020107 (2019).

[10] T. I. Smith, K. J. Louis, B. J. Ricci, and N. Bendjilali, Quantitatively ranking incorrect responses to multiple-choice questions using item response theory, Phys. Rev. Phys. Educ. Res. **16**, 010107 (2020).

[11] P. Eaton, K. Johnson, and S. Willoughby, Generating a growth-oriented partial credit grading model for the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **15**, 020151 (2019).

[12] J. Stewart, B. Drury, J. Wells, A. Adair, R. Henderson, Y. Ma, Á. Pérez-Lemonche, and D. Pritchard, Examining the relation of correct knowledge and misconceptions using the nominal response model, Phys. Rev. Phys. Educ. Res. **17**, 010122 (2021).

[13] Please note that the parameter $d_j$ does not have a particular interpretation in the psychometric literature analogous to discrimination or difficulty.

[14] J. A. Marshall, E. A. Hagedorn, and J. O'Connor, Anatomy of a physics test: Validation of the physics items on the Texas Assessment of Knowledge and Skills, Phys. Rev. ST Phys. Educ. Res. **5**, 010104 (2009).

[15] S. Rakkapao, S. Prasitpong, and K. Arayathanitkul, Analysis test of understanding of vectors with the three-parameter logistic model of item response theory and item response curves technique, Phys. Rev. Phys. Educ. Res. **12**, 020135 (2016).

[16] R. Henderson, P. Miller, J. Stewart, A. Traxler, and R. Lindell, Item-level gender fairness in the Force and Motion Conceptual Evaluation and the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **14**, 020103 (2018).

[17] Y. Xiao, J. C. Fritchman, J. Y. Bao, Y. Nie, J. Han, J. Xiong, H. Xiao, and L. Bao, Linking and comparing short and full-length concept inventories of electricity and magnetism using item response theory, Phys. Rev. Phys. Educ. Res. **15**, 020149 (2019).

[18] P. Eaton, K. Johnson, B. Frank, and S. Willoughby, Classical test theory and item response theory comparison of the brief electricity and magnetism assessment and the conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **15**, 010102 (2019).

[19] J. Hansen and J. Stewart, Multidimensional item response theory and the Brief Electricity and Magnetism Assessment, Phys. Rev. Phys. Educ. Res. **17**, 020139 (2021).

[20] J.-i. Yasuda, N. Mae, M. M. Hull, and M.-a. Taniguchi, Optimizing the length of computerized adaptive testing for the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **17**, 010115 (2021).

[21] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[22] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula, Am. J. Phys. **66**, 338 (1998).

[23] P. Eaton and S. Willoughby, Identifying a preinstruction to postinstruction factor model for the Force Concept Inventory within a multitrait item response theory framework, Phys. Rev. Phys. Educ. Res. **16**, 010106 (2020).

[24] R. D. Bock, Estimating item parameters and latent ability when responses are scored in two or more nominal categories, Psychometrika **37**, 29 (1972).

[25] R. Ostini and M. L. Nering, *Polytomous Item Response Theory Models*, Quantitative Applications in the Social Sciences, Vol. 07–144 (Sage Publications, Newbury Park, 2006).

[26] F. B. Baker, *Item Response Theory: Parameter Estimation Techniques* (Marcel Dekker, New York, 1992).

[27] The need for anchoring conditions for the set of $\{a_{k,j}, d_{k,j}\}$ also means that parameter values cannot be directly compared across items.

[28] P. J. Walter and G. Morris, Assessing Student Learning and Improving Instruction with Transition Matrices, presented at PER Conf. 2016, Sacramento, CA, 10.1119/perc.2016.pr.089.

[29] T. I. Smith, K. A. Gray, K. J. Louis, B. J. Ricci, and N. J. Wright, Showing the dynamics of student thinking as measured by the FMCE, presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.pr.090.

[30] R. D. Bock and I. Moustaki, Item response theory in a general framework, in *Handbook of Statistics*, Vol. 26, edited by C. R. Rao and S. Sinharay (Elsevier, New York, 2007), Chap. 15, pp. 469–514.

[31] H. Wainer, S. G. Sireci, and D. Thissen, Differential testlet functioning: Definitions and detection, J. Educ. Measure. **28**, 197 (1991).

[32] D. Thissen, L. Cai, and R. D. Bock, The nominal categories item response model, in *Handbook of Polytomous Item Response Theory Models*, edited by M. L. Nering and R. Ostini (Routledge, Taylor & Francis Group, New York, 2010), Chap. 3, pp. 43–75.

[33] R. D. Bock and R. D. Gibbons, *Item Response Theory* (Wiley, Hoboken, 2021).
[34] Y. Suh and D. M. Bolt, Nested logit models for multiple-choice item response data, Psychometrika **75**, 454 (2010).
[35] PhysPort, Data Explorer (2017), https://www.physport.org/.
[36] Learning Assistant Alliance, Learning About STEM Student Outcomes (LASSO) (2021), https://learningassistantalliance.org/modules/public/lasso.php.
[37] R. P. Chalmers, mirt: A multidimensional item response theory package for the R environment, J. Stat. Softw. **48**, 1 (2012).
[38] R. P. Chalmers, Multidimensional item response theory (mirt) (2017), https://cran.r-project.org/web/packages/mirt/index.html.
[39] R Core Team, *R: A Language and Environment for Statistical Computing* (2020), https://cran.r-project.org.
[40] Parameter values were quite stable, with standard error after 10 000 analyses being around 0.1 for each parameter.
[41] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.18.010133 for the details of our mathematical derivations, and additional justification for the physical analogy.
[42] R. D. Bock and M. Lieberman, Fitting a response model for n dichotomously scored items, Psychometrika **35**, 179 (1970).
[43] R. D. Bock and M. Aitkin, Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm, Psychometrika **46**, 443 (1981).
[44] R. J. Wirth and M. C. Edwards, Item factor analysis: Current approaches and future directions, Psychological Methods **12**, 58 (2007).
[45] The meaning of the negative sign in the exponent of Eq. (18) is discussed in greater detail in Sec. III C.
[46] E. M. Purcell and R. V. Pound, A nuclear spin system at negative temperature, Phys. Rev. **81**, 279 (1951).
[47] P. Hakonen and O. V. Lounasmaa, Negative absolute temperatures: "Hot" spins in spontaneous magnetic order, Science **265**, 1821 (1994).
[48] J. de Boer, Temperature as a basic physical quantity, Metrologia **1**, 158 (1965).
[49] In this analogy "temperature" cannot be measured directly.
[50] Please note: Because the zero point of $\theta$ is arbitrarily defined, we cannot necessarily relate positive values of $\theta$ with negative temperatures (or vice versa), but we can associate higher values of $\theta$ with hotter temperatures, and lower values of $\theta$ with colder temperatures. See Sec. III G for more details.
[51] S. E. Embretson and S. P. Reise, *Item Response Theory for Psychologists*, Multivariate Applications Books Series (Lawrence Erlbaum Associates Publishers, Mahwah, 2000), Chap. 6, pp. 129–133.
[52] S. White Brahmia, A. Olsho, T. I. Smith, and A. Boudreaux, Framework for the natures of negativity in introductory physics, Phys. Rev. Phys. Educ. Res. **16**, 010120 (2020).
[53] R. J. de Ayala, *The Theory and Practice of Item Response Theory* (Guilford Press, New York, 2008).