1-30-2020

Quantitatively ranking incorrect responses to multiple-choice questions using item response theory

Trevor Smith
Rowan University, smithtr@rowan.edu

Kyle J. Louis

Bartholemew J. Ricci IV

Nasrine Bendjilali
Rowan University, bendjilali@rowan.edu

Quantitatively ranking incorrect responses to multiple-choice questions using item response theory

Trevor I. Smith[1,2,*] Kyle J. Louis,[1,2] Bartholomew J. Ricci IV,[1,3] and Nasrine Bendjilali[3]

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

Research-based assessment instruments (RBAIs) are ubiquitous throughout both physics instruction and physics education research. The vast majority of analyses involving student responses to RBAI questions have focused on whether or not a student selects correct answers and using correctness to measure growth. This approach often undervalues the rich information that may be obtained by examining students' particular choices of incorrect answers. In the present study, we aim to reveal some of this valuable information by quantitatively determining the relative correctness of various incorrect responses. To accomplish this, we propose an assumption that allows us to define relative correctness: students who have a high understanding of Newtonian physics are likely to answer more questions correctly and also more likely to choose better incorrect responses than students who have a low understanding. Analyses using item response theory align with this assumption, and Bock's nominal response model allows us to uniquely rank each incorrect response. We present results from over 7000 students' responses to the Force and Motion Conceptual Evaluation.

I. INTRODUCTION

Many instructional and research questions over the past three decades have been answered by examining student responses to multiple-choice research-based assessment instruments (RBAIs) [1,2]. Tens of thousands of students have provided responses to questions on the Force Concept Inventory (FCI [3]), the Force and Motion Conceptual Evaluation (FMCE [4]), and many others [5], and dozens of analyses have been published that use these results to measure student learning (see Refs. [6,1]). A common factor throughout most of these analyses is that students' responses are typically scored as being correct or incorrect; very little attention has been paid to which incorrect answers students choose. This dichotomous scoring scheme is very beneficial for simplifying student performance on a RBAI or growth in learning to a single number that may be compared between students or across populations. The simplicity of this analysis and the ability for instructors and researchers to compare their results with other datasets has contributed to the proliferation of RBAIs,

to the benefit of the physics education research (PER) community; however, the dichotomous scoring scheme implicitly ignores any information about students' choices that are not correct. All incorrect answers are treated equally, regardless of how similar or different they may be to the correct answer.

RBAIs are so powerful because their questions help to elicit students' core beliefs about how the world works in ways that mathematical or problem-solving questions often do not. Many of the incorrect "distractor" response choices correspond with deeply held intuitive understandings that fit well with everyday experiences (and correspond with historically accurate models) but conflict with the principles of Newtonian physics [1]. The ability to deeply probe students' conceptual understanding of physics and represent this understanding with a single numerical value is very powerful. The authors of the FMCE, in fact, argue against using a single numerical score to represent student understanding [7], instead favoring the examination of student performance on individual or small groups of questions [4], but the common practice persists. Moreover, the common practice of reporting normalized gain as a measure of student learning has been shown to be biased against students with little prior exposure to formal physics instruction [8].

Other analyses of student responses to RBAI questions examine specific choices that students make and relate these choices to various mental models [9], misconceptions [3], views [10], or pieces of knowledge [11,12] that

students may have or use when answering particular questions. These analyses provide a lot of rich information about students' ideas, but the processes of conducting these analyses are often quite time intensive, and the presentation and visualization of the results can be conceptually dense and difficult to interpret [13,14]. As such, these analyses are not nearly as common as reporting a single numeric score.

Our ultimate goal is to develop a method for analyzing RBAI data that combines the benefits of having an easy-to-use tool that generates a single numeric score with the deep insights that can only be seen by considering both correct responses and the good ideas that may be expressed in some incorrect responses. The first part of defining a metric to create a numeric score of this type is to determine whether or not some incorrect answers may be considered better than others. In this usage, "better" could mean closer to correct, incorporating more productive ideas, or indicating a higher level of understanding.

In addition to developing a more complete metric for reporting student achievement, analyses that provide more detailed information about students' responses, such as consistency plots [13–15] or conceptual dynamics [10], often examine how these responses change over time. Interpretations of the results of these analyses would benefit greatly from a well-defined ranking of incorrect responses. In that way, if a student chooses different incorrect responses to a particular question before and after instruction, one could argue whether or not that student's understanding of the course material increased even though they never chose the correct answer.

Considering one incorrect answer to be better than another can be a tricky business. A simple method for defining which responses are better than others would be to ask expert instructors and researchers to rank the responses; however, reasonable people may disagree on this ranking. Any decision about the relative correctness of various responses based on expert opinions would inherently contain personal biases from the experts surveyed and the decision makers. We want to minimize the potential for bias in our rankings of incorrect responses. As such, we carefully articulate an assumption for defining what makes one response better than another, and we choose an analysis method that correspond to that assumption to quantitatively rank incorrect responses based on students' response patterns.

Assumption: Students who choose correct responses on most questions are more likely to choose better incorrect answers than students who choose few correct responses.

This assumption is based on the premise that students who understand more about Newtonian physics are more likely to choose better incorrect answers than students who understand less physics, and these students are also more likely to choose a greater number of correct responses. This assumption is consistent with previous work that has used item response curves (IRCs) to examine and rank incorrect responses on both the FCI and the FMCE [16–19]. We expand on this prior work by using a nested-logit item response theory (IRT) model to simultaneously estimate students' overall understanding of Newtonian mechanics (the IRT latent trait, or person parameter) and determine how closely each response choice correlates with a high level of understanding using the estimated parameters of the model [20–25]. Based on this assumption we would claim, for example, that a student who only incorrectly answers one question is more likely to choose a response that is almost correct than a student who answers 20 questions incorrectly. In Sec. VII we discuss alternate assumptions and the implications for future analyses based on those.

To illustrate the applicability of our assumption, we analyzed more than 7000 students' matched pretest and post-test responses to the FMCE to demonstrate how quantitative analyses can provide information about which response choices may be better than others. We present a ranking of incorrect responses for all FMCE questions as well as the parameter values used to make these determinations.

II. DATA SOURCES AND PREPARATION

Our data come from two primary sources:
- sets of student responses to the FMCE provided to one author (TIS) by colleagues from four different colleges or universities, schools 1–4 [26], as part of current and previous research projects ($N = 952$), and
- student responses uploaded to PhysPort's Data Explorer ($N = 6336$) [27].

Some information is known about the instructional settings at schools 1–4 (all of which used research-based instructional materials of some sort), but this information is not available for the PhysPort data. For the purposes of the current analysis, we combine all data into one set of $N = 7288$ students. We are not interested in how instructional factors impact student learning for this analysis, or whether or not student responses are different before or after instruction. As such, we have combined all pretest and post-test responses into a dataset of $N = 14\,576$ response sets.

To prepare the data for analysis, we omit any responses that are inappropriate for a given question (e.g., a given response of E on question 45, which only includes options A, B, C, and D). We also omit response J (None of these answers is correct) from interpretations of our analyses because it does not represent a well-defined indication of what each student would consider correct: two students who choose answer A agree on what they consider to be correct, but two students who choose answer J may have very different ideas of what would be a correct answer, so we cannot claim similarities between the responses of students who choose J. We also removed response sets with

TABLE I.   Question clusters on the FMCE [11].

Cluster	Questions	Context
Force Sled	1–7	sled on ice
Reversing Direction	8–13, 27–29	coin toss/car on ramp
Force Graphs	14–21	toy car
Accel. Graphs	22–26	toy car
Newton III	30–39	collisions/pushing
Velocity Graphs	40–43	toy car
Energy	44–47	sled on hill

three or more blank or unscorable responses. This gave us a usable dataset of $N = 12\,388$ response sets.

The structure of the FMCE makes it an interesting focus for this work. Unlike many other RBAIs, the FMCE contains several questions for each physical scenario presented (e.g., a toy car moving horizontally, see Table I), and all questions in each set have the same set of response choices. This is particularly interesting because a response choice that corresponds with the most common intuitive answer to one question, may not relate to any documented reasoning for another question.

III. ITEM RESPONSE THEORY MODELS

Item response theory (IRT) uses students' responses to multiple-choice questions to simultaneously estimate each student's overall understanding of the material (a.k.a. the latent trait or person parameter, θ) and determine the probability that a student will be correct on each question given his or her understanding [28,29]. The latent trait is normalized such that the average value is $\langle \theta \rangle = 0$ and the standard deviation is $\sigma_\theta = 1$. In the two-parameter logistic (2PL) IRT model, the probability of a student answering a specific question correctly is given by

$$P(\theta) = \frac{1}{1 + e^{-a(\theta-b)}}, \tag{1}$$

where a is the discrimination parameter and b is the difficulty parameter. Some previous work has used the three-parameter logistic model to analyze RBAI data [30], but we feel that the inclusion of the third "guessing" parameter is inappropriate for our analyses given that student responses to the FMCE are concentrated in a small subset of responses for each question: they are not, in fact, guessing [7,11].

The interpretation of the parameters in the 2PL model may be understood by examining plots of $P(\theta)$ vs θ: Fig. 1 shows examples from several questions. The difficulty b is the value of θ at which $P(b) = 0.5$, and the discrimination a is proportional to the slope of the curve at $\theta = b$: $dP/d\theta|_b = a/4$. Questions 1 and 14 [Figs. 1(a) and 1(b), respectively] have similar difficulty parameters (the b value differs by less than 0.1 standard deviations of the

(a) Question 1
$a = 2.5$, $b = 0.45$

(b) Question 14
$a = 3.7$, $b = 0.54$

(c) Question 22
$a = 2.8$, $b = -0.27$

(d) Question 47
$a = 1.1$, $b = 0.10$

FIG. 1.   Item response theory plots for questions 1, 14, 22, and 47: probability of answering correctly as a function of the latent ability parameter θ. The plots help to illustrate the meaning of the discrimination a and difficulty b parameters in the two-parameter logistic (2PL) IRT model shown in Eq. (1).

latent ability θ), but Q14's higher discrimination parameter a shows up as a sharper transition from most likely incorrect to most likely correct, and a steeper slope at the midpoint of the curve. Question 22 in Fig. 1(c) has a similar discrimination to Q1 [similar slope at $P(\theta) = 0.5$], but the difficulty is much lower [shown by a shift to the left compared to Fig. 1(a)], with many below-average students ($\theta < 0$) being fairly likely to answer correctly. Question 47 in Fig. 1(d) has a difficulty parameter that is about average (close to zero), but the discrimination is relatively small, as shown by a shallow slope, and a more gradual transition from probably incorrect to probably correct than any of the other three. Higher values of discrimination a mean a sharp transition and steeper slope; lower values mean a gradual transition and shallower slope. Higher values of difficulty b mean a graph that is shifted to the right; lower values mean a graph that is shifted to the left.

Bock's nominal response model (NRM) provides the probability of a person choosing each possible response based on θ [20],

$$P_k(\theta) = \frac{e^{a_k(\theta-b_k)}}{\sum_{i=1}^{N} e^{a_i(\theta-b_i)}}, \tag{2}$$

where k indicates the particular response choice, and the summation is performed over all N response choices. According to Bock and Moustaki, the value of the a_k parameter may be used to rank the incorrect responses,

with a higher value indicating a response that is more closely correlated with the latent trait and, therefore, better than a response with a lower value [31]; however, the meaning of the a_k and b_k parameters is not as easily interpreted as the a and b of the 2PL model [21]. (See Supplemental Material [44] for a physicist's interpretation of these parameters that provides additional insight.) One shortcoming of the NRM is that the parameters are not uniquely defined, and a normalization constraint is required. This is often accomplished by setting the value of both parameters associated with one particular response to be fixed (at 0 or 1) and determining all other parameters relative to those. The NRM is excellent for analyzing data for which no prior information is available regarding the relative correctness of any of the responses; however, we have found that for our FMCE data, the parameters occasionally become reversed with choosing the correct response being associated with having a low value of θ (i.e., a poor overall understanding of Newtonian mechanics).

IV. RANKING INCORRECT RESPONSES

A. Using a nested logit model to quantitatively rank responses

In order to rank incorrect responses while properly accounting for the correct response, we use the 2PL-NRM nested logit model developed by Suh and Bolt [22]. In this model, the probability of a student choosing a specific incorrect response k is given by

$$P_k(\theta) = \left(1 - \frac{1}{1 + e^{-a(\theta - b)}}\right) \frac{e^{a_k(\theta - b_k)}}{\sum_{i \neq \text{correct}} e^{a_i(\theta - b_i)}}, \quad (3)$$

where the parenthetical term is the probability of being not correct from the 2PL model, and the second term is Bock's NRM with the summation being over only the *incorrect* responses. In this model, the values of all θ, a, and b parameters are calculated using the 2PL model, and all a_k and b_k parameters are determined using the NRM, given the 2PL results. We used the multidimensional item response theory (mirt) package in the R programming language to perform all IRT analyses [23,24,32].

To determine the ranking of incorrect responses, we calculated the values of a and b for every question, and a_k and b_k for each incorrect response choice. According to de Ayala, a dataset must have at least 10 times as many response sets as the number of parameters to be calculated for an IRT model to have good convergence [29]; with our dataset of $N = 12\,388$ response sets we are more than able to determine the 722 necessary parameters.

As a result of using the mirt package to apply the 2PL-NRM nested logit model, every response choice within each question has a unique a_k value, implying that all answers are meaningfully different from each other. The question is then whether or not any of the a_k values may be considered approximately equal to others, indicating approximately equal correlations with the θ parameter (i.e., response choices that are equally correct). To determine whether or not response choices are different from each other, we calculated the sampling distribution of values for each a_k parameter by selecting random sample of 7300 respondents using the sample function in R, and we used the mirt package to calculate each parameter [33]. We repeated this process over 100 000 times to create a set of values for each parameter.

The mirt package uniquely determines each value of a_k by setting one parameter equal to 1 for each question [24]. In order to ensure that we obtained a distribution of values for all parameters of interest, we chose to include one set of responses that included a "dummy" response and set $a_0 = 0$ for this response. All other a_k parameters are determined relative to a_0; as such, the a_k values are only meaningful when compared within the same question, and there are no thresholds for determining whether a particular value of a_k is high or low in and of itself.

B. Determining whether or not parameters are meaningfully different

Using the effsize package, we calculated a Hedges' g effect size to quantify the magnitude of the difference between the a_k values for each pair [34]. In our full dataset, every response to every question is selected in at least one response set. The a_k values reported in Table II are those determined from the full dataset, with the dummy response set included to uniquely determine each parameter. Given that the values of a_k parameters are only meaningful in relation to other values for the same question, the inclusion of the dummy response set has minimal impact on the overall results.

Figure 2 shows a graphical representation of the parameter distributions for question 19. There are several key features to notice about these distributions:
  (i) there is no distribution for the correct response because the 2PL-NRM does not calculate an a_k parameter for the correct response; it is automatically assumed to be the best,
 (ii) the distribution for a_A is higher than any other value of a_k, with only minimal overlap with a_E, indicating that A has the highest parameter, and is thus the best incorrect response, and
(iii) the distributions a_C and a_G are practically identical, indicating that these parameters have very similar values; thus, we would interpret them as being equally correct.

Other comparisons between various responses are a bit more ambiguous. The a_D and a_F distributions look quite

TABLE II. Ranking results from IRT analyses (questions 1–47). These results show the complete ranking for every response choice to every question, as well as the associated a_k values for each response, the value of Hedges' g used to determine the ranking between each pair of parameters, and the percent of the dataset that chose each response (rounded to the nearest percent).

Q1: B > D ≥ A ≥ C = E ≈ F ≥ G
- a_k: 0.77, −0.18, −0.48, −0.62, −0.77, −1.39
- g: 2.75, 1.25, 0.34, 0.28, 1.21
- %: 32, 66, 1, 0, 0, 0

Q2: D > C ≥ E ≥ G ≥ A = B > F
- a_k: 1.62, 0.54, 0.29, −0.14, −0.46, −0.51, −1.46
- g: 2.69, 1.04, 0.54, 0.16, 2.86
- %: 29, 4, 0, 0, 1, 65, 0

Q3: F > B ≥ E ≥ G > A = C ≈ D
- a_k: 0.59, 0.29, 0.13, −0.49, −0.61, −0.71
- g: 1.32, 0.73, 2.71, 0.45, 0.47
- %: 41, 1, 6, 7, 1, 41, 3

Q4: F > D > B = E ≥ C ≥ G > A
- a_k: 0.76, 0.27, 0.15, −0.02, −0.3, −0.7
- g: 2.05, 0.5, 0.78, 1.28, 1.66
- %: 31, 1, 3, 0, 63, 1

Q5: D > C > E > A ≥ B = G = F
- a_k: 1.31, 0.18, −0.45, −0.61, −0.62, −0.73
- g: 4.84, 2.56, 0.67, 0.06, 0.45
- %: 48, 4, 1, 3, 36, 1, 7

Q6: B > F > A = E ≥ G > C > D
- a_k: 0.91, 0.17, 0.16, 0.05, −0.36, −0.87
- g: 3.62, 0.06, 0.53, 1.98, 2.46
- %: 19, 12, 7, 16, 6, 34, 5

Q7: B > F = C = A > G > D = E
- a_k: 0.22, 0.14, 0.1, −0.39, −0.68, −0.78
- g: 0.34, 0.18, 2.23, 1.33, 0.43
- %: 42, 4, 6, 8, 2, 4, 32

Q8: A > D > B = E > C > G > F
- a_k: 1.1, 0.48, 0.37, 0.35, −0.17, −0.94
- g: 2.5, 0.43, 0.08, 2.08, 3.14
- %: 23, 1, 4, 9, 2, 44, 16

Q9: A > C = B > D ≥ G = E ≈ F
- a_k: 0.29, 0.27, −0.17, −0.5, −0.55, −0.63
- g: 1.85, 1.27, 0.16, 0.34
- %: 26, 1, 3, 67, 1, 3, 1

Q10: A > E ≥ D ≥ G ≥ B = C = F
- a_k: 0.76, 0.45, −0.2, −0.37, −0.4, −0.42
- g: 2.53, 1.2, 0.65, 0.11, 0.08
- %: 35, 1, 2, 54, 6, 1

Q11: A > D > C ≥ B ≥ E > G > F
- a_k: 0.53, 0.22, 0.11, −0.19, −0.46, −1.64
- g: 1.5, 0.58, 1.56, 1.38, 6.23
- %: 34, 1, 4, 3, 4, 48, 7

Q12: A > B = C > F > D > E > G
- a_k: 0.23, 0.16, −0.18, −0.48, −0.81, −1.14
- g: 0.32, 1.37, 1.41, 1.6, 1.15
- %: 36, 2, 1, 0, 60, 1, 0

Q13: A > D = E ≈ C ≥ B ≥ G ≥ F
- a_k: 0.45, 0.42, −0.63, −0.75, −0.87, −1.06
- g: 0.11, 4.8, 0.57, 0.57, 0.84
- %: 44, 1, 1, 3, 50, 1, 1

Q14: E > H > F > G > A ≥ D = C = B
- a_k: 2.07, 1.01, 0.3, −0.55, −0.84, −0.97, −0.99
- g: 2.7, 1.34, 1.89, 0.86, 0.32, 0.05
- %: 27, 0, 0, 62, 1, 5, 1

Q15: E > G ≥ D ≥ H = C > F > A ≥ B
- a_k: 0.94, 0.43, 0.19, 0.17, 0.04, −0.35, −0.62
- g: 1.27, 0.69, 0.03, 0.44, 1.37, 1.07
- %: 95, 0, 0, 0, 0, 1, 2

Q16: A > E > B = F > C > D = G ≥ H
- a_k: 0.94, 0.53, −0.26, −0.36, −0.95, −0.96, −1.3
- g: 1.5, 2.49, 0.34, 2.83, 0.05, 0.09
- %: 31, 1, 0, 0, 1, 66, 1, 0

(Table continued)

TABLE II. (Continued)

The table is rotated on the page. Each question lists its response options ranked by a_k, with comparison operators between adjacent options (top operator based on a_k; the g row gives the effect size with operators ≈, >, ≥). The % row gives the fraction of respondents choosing each option.

Q17: E > F = B > D ≥ G = C ≥ H ≥ A
a_k: 0.06, −0.02, −0.48, −0.64, −0.68, −0.81, −1.07
g: ≈ 0.25, > 2.36, ≥ 0.76, ≈ 0.22, ≥ 0.58, ≥ 1.23
%: 24, 0, 53, 9, 2, 49, 5, 4

Q18: B > A ≥ E > D = G = H > C > F
a_k: 1.07, 0.77, −0.26, −0.35, −0.35, −0.67, −0.97
g: ≥ 1.26, > 4.32, ≈ 0.39, ≈ 0, > 1.46, > 1.31
%: 27, 2, 1, 8, 9, 49, 1, 2

Q19: B > A > E > D ≈ F > C = G ≥ H
a_k: 2.31, 0.76, −0.16, −0.19, −1.05, −1.06, −1.41
g: > 4.98, > 2.83, ≈ 0.12, > 2.85, ≈ 0.01, ≥ 1.19
%: 27, 3, 0, 52, 1, 6, 5, 3

Q20: G > B ≥ A = D = C = E = H > F
a_k: 0.59, 0.27, 0.22, 0.21, 0.21, 0.19, −0.23
g: ≥ 1.57, ≈ 0.25, ≈ 0.04, ≈ 0.03, ≈ 0.1, > 2.31
%: 29, 0, 1, 2, 1, 0, 2, 64

Q21: E > B > A = D ≈ H = F > C > G
a_k: 0.56, −0.01, −0.38, −0.47, −0.5, −0.69, −1
g: > 2.86, > 1.77, ≈ 0.43, ≈ 0.19, > 0.87, > 1.52
%: 36, 2, 8, 2, 25, 15, 2, 7

Q22: A ≥ C ≥ B > D > G > F = E
a_k: 0.76, 0.57, 0.17, −0.4, −1.1, −1.16
g: ≥ 0.91, ≥ 1.6, > 1.93, > 2.12, ≈ 0.09
%: 55, 1, 1, 0, 0, 1, 42

Q23: B > A > C ≥ F > G = E ≥ D
a_k: 0.59, 0.58, −0.61, −0.69, −0.74, −0.99
g: > 5.67, ≈ 0.38, ≈ 0.27, ≈ 0.95, ≥ 1.03
%: 43, 2, 1, 8, 41, 5, 4

Q24: C > D > E ≈ F > B > G > A
a_k: −0.2, −0.67, −0.68, −0.91, −1.17, −1.61
g: > 2.22, > 0, ≈ 0, > 1.41, > 1.51, > 2.34
%: 53, 1, 1, 4, 33, 4, 3, 2

Q25: B > A > C > D = F > E ≥ G
a_k: 1.08, 0.5, −0.65, −0.69, −1.14, −1.45
g: > 2.47, > 4.37, ≈ 0.13, > 1.97, ≥ 1.24
%: 40, 6, 1, 1, 41, 5, 4

Q26: C > G ≥ D ≥ B > E ≥ A
a_k: 0.54, 0.42, −0.12, −0.35, −0.78, −1.32
g: ≥ 0.55, ≥ 2.31, ≥ 0.9, > 1.79, ≥ 2.37
%: 55, 1, 0, 2, 3, 37

Q27: A > D ≥ B ≥ E ≥ C ≥ G > F
a_k: 0.3, 0.06, −0.11, −0.31, −0.92, −2.1
g: ≥ 1.17, ≥ 0.92, ≥ 1.06, > 3.3, > 6.13
%: 42, 1, 5, 4, 7, 36, 5

Q28: A ≥ E ≥ B = C > F > D > G
a_k: 0.31, 0.16, 0.15, −0.23, −0.61, −0.98
g: ≥ 0.64, ≈ 0.07, > 1.55, > 1.52, > 1.4
%: 38, 2, 2, 1, 0, 0, 57

Q29: A > E > D > C > F > G ≈ B
a_k: 1.07, 0.17, −0.46, −0.81, −0.85, −0.9
g: > 3.62, > 2.34, > 1.37, > 0.81, ≈ 0.16
%: 45, 5, 1, 3, 6, 1, 40

Q30: E > B ≥ F ≈ C ≥ A ≥ D
a_k: 0.25, 0.1, −0.1, −0.22, −0.38
g: ≥ 0.58, ≈ 0.74, ≥ 0.44, ≥ 0.62
%: 42, 1, 54, 1, 5, 14

Q31: E > D = F ≈ C ≥ A ≥ B
a_k: −0.01, −0.04, −0.14, −0.39, −0.61
g: ≈ 0.09, ≈ 0.41, ≥ 0.95, ≥ 0.85
%: 47, 1, 54, 5, 1, 5

Q32: E > F = A ≈ D = B ≥ C
a_k: −0.06, −0.16, −0.27, −0.37, −0.5
g: ≈ 0.41, ≈ 0.44, ≈ 0.39, ≥ 0.55
%: 44, 7, 5, 2, 39, 2

TABLE II. (Continued)

Q33: E > D ≥ F ≥ B ≥ C ≥ A
a_k: 0.2, 0.08, -0.14, -0.29, -0.45
g: 0.56, 1.03, 0.73, 0.7
%: 93, 1, 1, 2, 2

Q34: E > F = A ≈ D ≥ C ≈ B
a_k: -0.13, -0.14, -0.23, -0.24, -0.35
g: 0.01, 0.37, 0.06, 0.42
%: 41, 1, 3, 2, 1, 50

Q35: A > E > B = C ≥ D
a_k: 0.77, -0.13, -0.25, -0.57
g: 3.6, 0.47, 1.27
%: 56, 1, 38, 2, 2

Q36: A > E ≥ C ≥ B > D
a_k: 0.51, 0.16, -0.01, -0.54
g: 1.51, 0.72, 2.32
%: 23, 1, 68, 3, 3

Q37: A > E > B > C ≥ D
a_k: 0.29, -0.09, -0.56, -0.78
g: 1.84, 2.27, 1.07
%: 65, 4, 3, 22, 4

Q38: A > E = B ≥ C ≥ D
a_k: 0.27, 0.16, 0.05, -0.2
g: 0.47, 0.53, 1.09
%: 23, 2, 66, 5, 3

Q39: E > A ≥ C = D > B
a_k: -0.12, -0.4, -0.42, -1.28
g: 1.25, 0.07, 3.96
%: 51, 3, 4, 35, 6

Q40: A > E ≥ B ≥ F ≥ C ≥ G ≥ H ≥ D
a_k: 0.29, 0.12, 0.08, -0.18, -0.4, -0.6, -0.84
g: 0.77, 0.22, 1.09, 0.84, 0.55, 0.72
%: 88, 2, 2, 0, 1, 0, 0, 7

Q41: F > E > B > D = A ≥ C > G ≥ H
a_k: 0.46, 0.06, -0.28, -0.36, -0.48, -0.87, -0.99
g: 2.05, 1.69, 0.39, 0.55, 1.98, 0.56
%: 72, 1, 7, 1, 1, 5, 7, 5

Q42: B > E > G > F = D ≥ H ≥ A ≈ C
a_k: 0.06, -0.37, -0.38, -0.43, -0.44, -0.61, -0.69
g: 1.9, 0.02, 0.26, 0.06, 0.79, 0.42
%: 78, 1, 0, 1, 2, 7, 2, 7

Q43: D > B ≥ G ≈ C = F = A ≥ H > E
a_k: 0.03, -0.09, -0.12, -0.21, -0.26, -0.27, -0.77
g: 0.54, 0.14, 0.41, 0.19, 0.06, 2.1
%: 89, 1, 0, 1, 0, 4, 4, 1

Q44: B > D = C ≥ A
a_k: -0.03, -0.04, -0.25
g: 0.06, 0.81
%: 44, 2, 4, 49

Q45: B > C = A ≈ D
a_k: -0.17, -0.21, -0.24
g: 0.14, 0.11
%: 54, 6, 37, 3

Q46: A > B = D ≥ C
a_k: 0.04, -0.02, -0.3
g: 0.24, 1.11
%: 44, 16, 17, 22

Q47: A > D = B ≥ C
a_k: 0.01, -0.01, -0.26
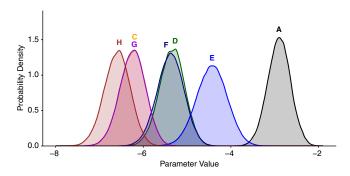g: 0.08, 1.03
%: 46, 13, 22, 19

FIG. 2. Distributions of a_k parameters for question 19. Plot created using the geom_density function in the ggplot2 package, and the directlabels package for R [35,36].

similar, but not as similar as a_C and a_G. The a_H distribution is noticeably shifted to the left of a_C and a_G, but there is still quite a bit of overlap. We use Hedges' g to quantify the magnitude of the difference between each pair of distributions of the a_k values: if g is small ($g < 0.5$) we conclude that the parameter values are effectively equal and the responses are equally correct, if g is very large ($g \geq 1.3$) we conclude that the parameters are significantly different and that the responses represent different levels of understanding, and if $0.5 \leq g < 1.3$ then we cannot make a conclusive determination.

C. The final rankings

Table II show the IRT ranking results for each question on the FMCE, including the a_k value for each response, the value of Hedges' g for each nearest-neighbor comparison, and the percentage of the dataset that chose each response. Consider the ranking shown for question 19 as it relates to Fig. 2. The a_C and a_G parameters are nearly identical (when rounded to two decimal places), and the effect size between their distributions is negligibly small ($g = 0.01$). The effect size between D and F is also quite small ($g = 0.12$), suggesting that choosing either of these two responses indicates a similar level of understanding, and the effect size between G and H is large, but not above our threshold for different responses ($g = 1.19$) [37]. All other effect sizes for question 19 are very large, indicating considerably different values of a_k that correspond to different levels of understanding.

Question 19 on the FMCE presents students with a situation in which a toy car "moves toward the left and is speeding up at a steady rate (constant acceleration)" and asks them to choose an appropriate graph of force vs time. The correct response to question 19 is B: a graph with a constant negative value (zero slope). According to these results, the best incorrect answer is A: a graph with a constant positive value (zero slope). The second-best incorrect response is E: a graph with a constant zero value (zero slope). All other responses are graphs with

nonzero slope. This suggests that realizing that a constant acceleration indicates a constant force is indicative of an above-average understanding of basic Newtonian mechanics. This result alone may not be revolutionary to anyone who has taught introductory mechanics, but the implication that claiming that zero force is required to make an object speed up is a better answer than selecting a graph showing a changing force may be more surprising. Response E is chosen by fewer than 1% of the dataset, but these students seem to otherwise have a fairly strong understanding of Newton's laws as measured by the FMCE.

Table III shows IRT rankings that have been filtered to only include responses given by at least 1.00% of the population. For some questions (such as 1, 16, and 22) the difference is quite stark, with only the correct and one incorrect answer choice remaining. For many of the questions (such as 2, 14, and 26) the rankings remain the same, but many of the responses that are seen as equivalent to others (or ambiguously ranked) have been eliminated. Moreover, a smaller fraction of the rankings in Table III are "≥" as compared to Table II (29% vs 35%), and a greater fraction of rankings are ">" (56% vs 36%). This suggests that many of the ambiguities in rankings may be attributed to the relatively low probability of choosing those responses at all levels of physics understanding, which could result in relatively broad distributions of parameter values generated by randomly selecting subsets of data.

V. RELATING RESPONSE RANKINGS TO IRT PLOTS

We can use plots of the IRT curves to better understand these rankings. Figure 3 shows the 2PL-NRM curves for every response to several question. (The 2PL-NRM IRT plots for all other questions are included in the Appendix.) In the filtered ranking, question 1 only includes two dominant responses. This is consistent with Fig. 3(a) in which responses A and B dominate at all values of θ (overall understanding), and all other responses have near zero probability of being chosen. Figure 3(c) shows that Q14 is a bit more interesting: there are still two dominant responses (the correct E and a single incorrect A), but less-common incorrect responses are chosen differently by students with different θ values. Students who choose response H are likely to have an above-average understanding ($\theta > 0$) with the H curve having a roughly symmetric probability distribution centered around $\theta = b \approx 0.75$, while response C is mostly chosen by students with below-average understanding and is more and more likely with lower values of θ. This is consistent with the ranking in Table III with H being a better response than the most common A, and C being a worse response. We can also see in Fig. 3(c) that responses B and C on Q14 have a

TABLE III.   IRT ranking results: filtered so that only responses given by at least 1% of the population are included.

Q1:	B	>	A														
Q2:	D	>	C	>	A	=	B										
Q3:	F	>	E	≥	G	>	C	=	D								
Q4:	F	>	E	>	G												
Q5:	D	>	C	>	E	>	A	≥	B	≥	F						
Q6:	B	>	F	>	A	=	E	≥	G	>	C	>	D				
Q7:	B	>	F	=	C	=	A	>	G	>	D	=	E				
Q8:	A	>	D	>	B	=	E	=	C	>	G	>	F				
Q9:	A	>	B	>	D	>	E										
Q10:	A	>	D	>	B	=	C										
Q11:	A	>	B	≥	C	>	E	>	G	>	F						
Q12:	A	>	B	>	D												
Q13:	A	>	C	≥	B	>	F										
Q14:	E	>	H	>	A	>	C	=	B								
Q15:	E	>	A	≥	B												
Q16:	A	>	C														
Q17:	E	>	B	>	D	≥	G	=	C	≥	H	≥	A				
Q18:	B	>	A	>	D	=	G	=	H	>	C	>	F				
Q19:	B	>	A	>	D	>	C	=	G	≥	H						
Q20:	G	>	H	>	F												
Q21:	E	>	B	>	A	>	D	=	H	=	F	≥	C	>	G		
Q22:	A	>	E														
Q23:	B	>	A	=	C	>	F	=	G	>	D						
Q24:	C	>	F	>	B	>	G	>	A								
Q25:	B	>	A	>	D	=	F	>	E	≥	G						
Q26:	C	>	G	>	B	>	E	>	A								
Q27:	A	>	B	≥	E	≥	C	>	G	>	F						
Q28:	A	>	E	≥	B	>	D										
Q29:	A	>	E	>	C	>	F	=	B								
Q30:	E	>	B	≥	F	≥	C	=	A	≥	D						
Q31:	E	>	D	=	F	=	C	≥	A	≥	B						
Q32:	E	>	F	=	A	=	D	=	B	≥	C						
Q33:	E	>	D	≥	F	>	C	≥	A								
Q34:	E	>	F	=	A	=	D	=	C	=	B						
Q35:	A	>	E	>	B	=	C	≥	D								
Q36:	A	>	E	>	C	≥	B	>	D								
Q37:	A	>	E	>	B	>	C	≥	D								
Q38:	A	>	E	=	B	≥	C	≥	D								
Q39:	E	>	A	≥	C	=	D	>	B								
Q40:	A	>	E	≥	B	>	C	>	D								
Q41:	F	>	E	>	B	>	D	≥	C	>	G	≥	H				
Q42:	B	>	E	>	D	=	H	≥	A	=	C						
Q43:	D	>	B	≥	C	≥	A	=	H								
Q44:	B	>	D	=	C	≥	A										
Q45:	B	>	C	=	A	=	D										
Q46:	A	>	B	=	D	≥	C										
Q47:	A	>	D	=	B	≥	C										

similar shape, with the highest probability of choosing each being at the low end of the θ axis; this is consistent with these responses being considered equivalent in Table III even though the value of the probability is quite different for each. It is also important to notice that the curves for the correct responses (B for Q1 and E for Q14) are identical to the 2PL curves for the questions shown in Fig. 1.

Question 19 in Fig. 3(e) is very similar to Q14 in that there is a single most-common-incorrect response (D), a better response (A) that has a relatively narrow symmetric probability distribution centered around $\theta \approx b > 0$, and some worse responses (C, G, H) that have their highest probabilities at the low end of the θ axis. The major differences between the plots for Q19 and Q14 are that the most-common-incorrect response

(a) Question 1   (b) Question 8   (c) Question 14

(d) Question 18   (e) Question 19   (f) Question 47
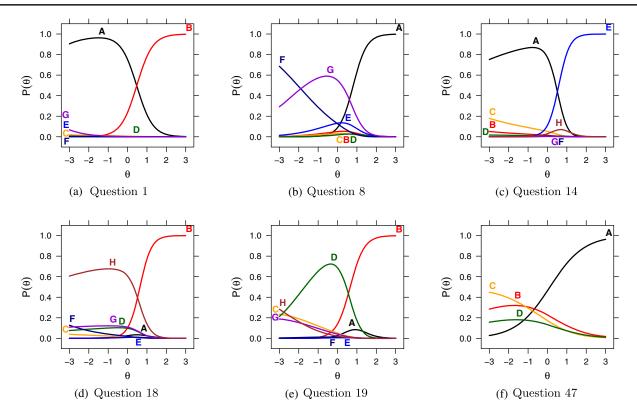
FIG. 3.   IRT plots from the 2PL-NRM nested logit model for questions 1, 8, 14, 18, 19, and 47: probability of choosing each response as a function of the person parameter θ (representing overall understanding of Newtonian mechanics).

has a much broader range of values in the $\theta < 0$ regime for Q19. In fact, response D is approximately equally probable to responses C, G, and H around $\theta \approx -3$. Moreover, the probability curves for the equivalent responses C and G are basically on top of each other for the entire range of θ. This is strong evidence that students choose these responses in roughly the same proportions, and it may indicate that students are choosing them for the same reasons. Response H also has similar probabilities to C and G, but has a distinctly more negative slope than either, indicating a greater likelihood of being chosen by students with lower θ values.

The plot for question 18 in Fig. 3(d) is even more complex than that of Q14, but there are several similarities between them. Once again there is a single dominant incorrect response (H), but it is less dominant than the most common response to Q1 or Q14. This is largely due to G and D being equivalent to H (according to Tables II and III), and all three sharing a similar shape in which the probability is relatively uniform for $\theta < 0$ (but has a slight peak around $-1 < \theta < -0.5$) and decreases to zero for $0 < \theta < 2$. Responses D and G each have probabilities between about 0.1 and 0.15 for $\theta < 0$, which accounts for the probability of the most common response H never rising above 0.75. Also in Fig. 3(d) we can see a small bump in probability for the

better-than-common response A around $\theta \approx 0.75$, and the worst responses (C and F) are most common at the lowest values of θ. Questions 14, 18, and 19 all indicate that students with above-average understanding ($\theta > 0$) are more likely to choose both the correct response and the highest-ranked incorrect response than students with $\theta < 0$.

The 2PL-NRM plot for question 47 in Fig. 3(f) shows an example of a question on the FMCE for which there is not a single dominant incorrect response. Once again we can see the ranking from Table III in the shape of the curves: A is correct, D and B have similar shapes with maximum probabilities around $\theta \approx -1.5$, and C with the highest probability at the low end of the θ axis with a distinctly negative slope. What makes Q47 really interesting is that none of the responses is uniformly zero over a broad range of θ values. All incorrect responses have probabilities above 0.15 for $-3 < \theta < -1$, and the probabilities for two of the incorrect responses (B and C) are roughly equal to the correct response around $\theta \approx -0.5$ (with response D only about 0.10 lower). The relatively high probabilities of all incorrect answer choices on Q47 may contribute to the low value of the discrimination parameter a (see Fig. 1).

Question 8 [Fig. 3(b)] shows an interesting example of a case where there are multiple responses (B, C, D,

and E) ranked higher than the most common incorrect response G. Only response F is ranked lower than G (with $g \approx 3.1$). On the FMCE, question 8 is the first in a set of three that asks students about the forces on a toy car as it rolls up and down a ramp. In question 8, the car is moving up and slowing down: response A is correct that the net force on the car is constant and down the ramp, response G is that the force is up the ramp and decreasing, while response F is that the force is up the ramp and increasing. Responses B and C both indicate a force down the ramp (increasing and decreasing, respectively), response D indicates zero net force, and response E is that the force is up the ramp and constant. All of these better-than-most-common responses agree with the correct answer in one way (either the direction of the force or the fact that it is constant), while G and F have nothing in common with the correct response.

From a visual perspective, the approximate value of θ for which the probability curve for a particular response is maximized indicates the ranking for that response [38]. The curve for the correct answer is always a monotonically increasing function of θ; answers that are better than a most-common incorrect answer [like H on Q14 in Fig. 3(c) and A on Q18 in Fig. 3(f)] have relatively narrow probability distributions with peaks around $\theta = b$; the worst response in each rank has its highest value at the low end of the displayed θ axis with a distinctly negative slope, indicating that lower values of θ would yield even higher probabilities of choosing those responses. Plots of IRT probability curves for all FMCE questions are included in Figs. 6–8 in the Appendix.

VI. ANOMALOUS RESULTS

We used Rowan University's high-performance computing cluster to generate the distribution of values for each parameter. In examining the results, we noticed that each parameter had two distinct distributions: a lower distribution containing approximately 99% of the analyses from the sampled data, and a higher distribution containing 1% of the analyses. All claims from the previous section were based on the main (99%) distribution, which is typified in Fig. 2. Figure 4 shows the full parameter distributions for questions 19, including the anomalous values for each parameter as shown by the small bumps on the right side. Figure 5 shows an enlarged version of the smaller (1%) distribution [39].

Comparing Figs. 2 and 5, we see that the overall shapes and locations of the parameter distributions are similar. The distributions in Fig. 5 are not as smooth or as broad as those in Fig. 2, but this may be attributed to the sampling method. Possibly the most interesting feature of these distributions is that there is not overlap between the left (99%) and right (1%) distributions of
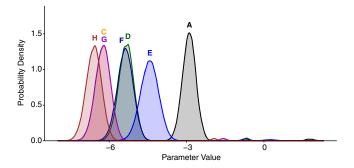

FIG. 4. The complete parameter distribution for question 19. The large labeled curves on the left side of the plot contain 99% of the results (also shown in Fig. 2), and the small bumps on the right side of the plot contain about 1% of the results (also shown in Fig. 5). A gap in the horizontal axis of about 1.5 exists between the left and right distributions for all parameters.

any parameter (which is not clear in any of the figures). For example, consider the distribution of the parameter for response A in Fig. 4: the left distribution for the parameter a_A (also shown as the black curve in Fig. 2) contains values between -4.2 and -1.9, while the right distribution (also shown as the black curve in Fig. 5) contains values between $+1.3$ and $+2.3$. This leaves a gap of 3.2 that contains no results.

We found the same anomalous results in the distributions for all parameters in all questions. In 1003 out of the 105 600 analyses all of the a_k parameters are about 2–5 units larger than in the other 104 597 analyses. The same 1003 analyses provide the higher results for all parameters, and for all questions the distributions of the higher values are similar in shape and relative location to the distributions for the lower values. Given the similarities in the distributions, and the consistency of the same 1% of analyses providing the differences for all parameters, we feel comfortable using the results from the 104 597 main analyses to make claims about the differences and similarities
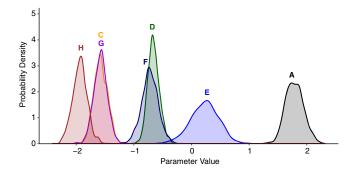

FIG. 5. Enlarged version of the right side of Fig. 4. These distributions represent about 1% of the total results. The overall shapes and locations of the parameter distributions are similar to Fig. 2.

between the a_k parameters for each response choice as we did in Secs. V and IV. We include the anomalous results in this section in order to maintain the transparency of our analysis methods.

VII. SUMMARY AND FUTURE DIRECTIONS

The 2PL-NRM nested logit IRT model may be used to rank incorrect responses to FMCE questions by considering the calculated a_k value as a measure of the correlation between choosing a particular response and the value of the θ parameter representing overall understanding of Newtonian mechanics (as measured by the FMCE). We have shown that using random samples of a large dataset can generate distributions of values for each a_k that allow us to determine whether or not these parameters are meaningfully different, and we used Hedges' g as a statistical measure to quantify these differences. We made particular choices regarding the values of g that we consider to represent parameters that are approximately equal, those that are definitely different, and those that could go either way, and we have reported the value of g for all comparisons to allow the reader to evaluate the validity of our claims or determine a ranking based on other choices of thresholds.

In many cases the responses that could not be determined to be definitely different or approximately the same as others are those that are rarely chosen by students. Future research will focus on clarifying these comparisons and trying to determine a robust ranking for all responses to every question. One way to accomplish this will be to perform similar analyses on other large datasets. Online data collection and analysis tools such as PhysPort's Data Explorer [27] and the Learning About STEM Student Outcomes (LASSO) [40] make this task much more achievable than it would have been even a decade ago.

As stated in Sec. I, our ultimate goal is to develop a method for defining student scores on the FMCE (and other RBAIs) that incorporates both the number of questions answered correctly and the value of the good ideas contained in other responses. That task is beyond the scope of the current study, as it would require more robust rankings of incorrect answers than are available from one type of analysis based on one assumption of what makes one response *better* than another.

The results presented here are based on the assumption that students who understand more about physics will answer more questions correctly on the FMCE, and will also select better incorrect responses than students who understand less overall physics. We also ignored whether data were collected before or after instruction: we did not care how students obtained their understanding, just that they had some level of understanding when they chose their responses. In future work, we will explore the implications

of different assumptions for what makes one response better than another.

One such assumption is that students will choose better responses after instruction than before instruction. This assumption is supported by the fact that (on average) students are more likely to choose correct responses after instruction than before: even low class-averaged gains from traditional instruction tend to be positive [2,7,41]. A method consistent with this assumption would be to look for asymmetric transitions between response choices in matched pretest and post-test data using a McNemar-Bowker chi-square test [42,43].

Another assumption that could be made is that students are more likely to choose correct responses after instruction if they chose better incorrect responses before instruction. Using conditional probabilities would allow us to identify a progression of responses to each question, with students moving up the progression being considered getting closer to correct. This is consistent with Thornton's conceptual dynamics in which students move between various views as they progress toward the correct response [10].

Each of these methods could be used to test the rankings presented above and help clarify the ambiguously ranked responses. These methods and assumptions may also be applied to other research-based assessments to determine a robust ranking of incorrect responses for any multiple-choice question.

ACKNOWLEDGMENTS

APPENDIX: ADDITIONAL 2PL-NRM PLOTS

Figures 6–8 are included to allow the reader to see how the rankings in Table II relate to the probability of choosing each response, given the value of a student's θ parameter. Each curve is labeled near the maximum value (with slight adjustments to avoid overlapping labels and curves), so the horizontal location of the label provides an approximate ranking of the responses (right is better, left is worse). Ambiguously ranked responses often show up as lines near zero probability for all values of θ.
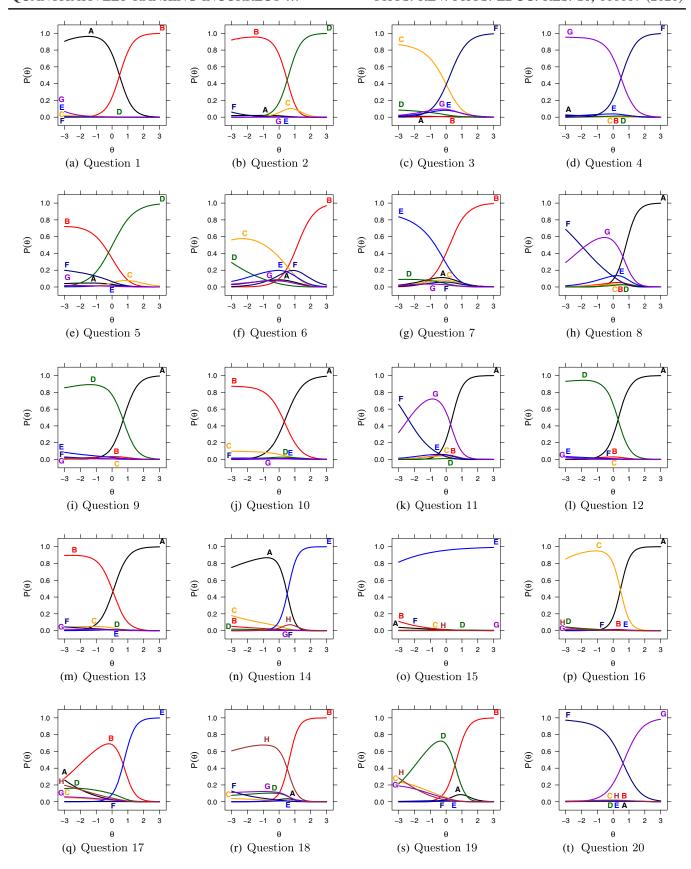
FIG. 6.   IRT plots from the 2PL-NRM nested logit model for questions 1–20: Probability of choosing each response as a function of the person parameter θ (representing overall understanding of Newtonian mechanics). Plots created using the mirt and directlabels packages in R [24,32,36].
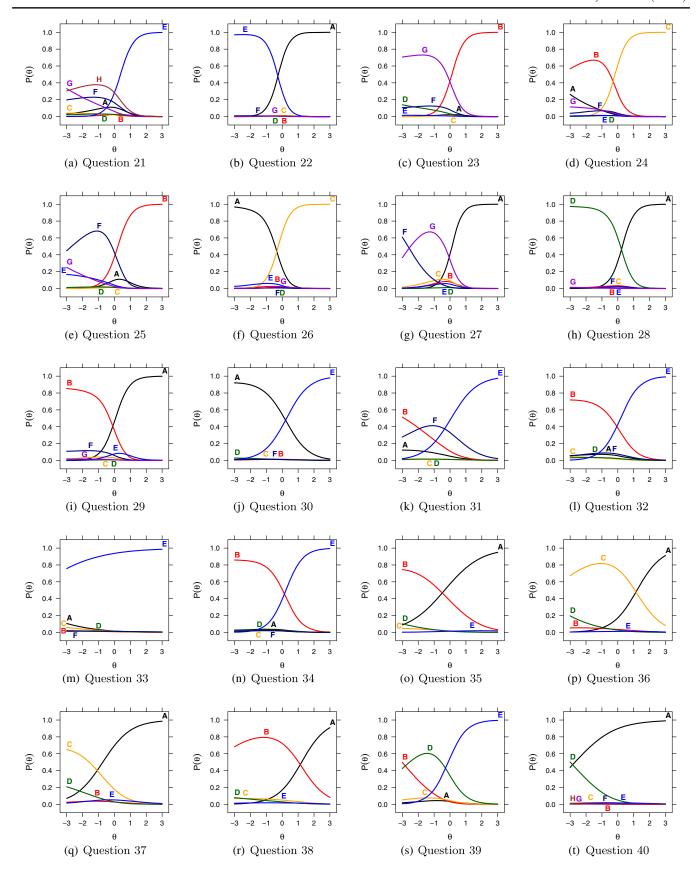
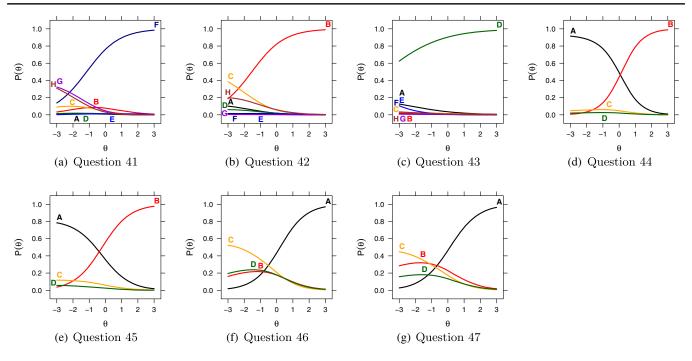FIG. 7.   IRT plots from the 2PL-NRM nested logit model for questions 21–40.

FIG. 8.   IRT plots from the 2PL-NRM nested logit model for questions 41–47.

[1] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter RBAI-1: Research-based assessment instruments in physics and astronomy, Am. J. Phys. **85,** 245 (2017).

[2] J. Von Korff, B. Archibeque, K. A. Gomez, S. B. Mckagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell, Secondary analysis of teaching methods in introductory physics: A 50k-student study, Am. J. Phys. **84,** 969 (2016).

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

[4] 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).

[5] PhysPort, (2011–2019), https://www.physport.org, Last accessed on Mar. 20, 2019.

[6] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10,** 020119 (2014).

[7] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the force and motion conceptual evaluation and the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **5,** 010105 (2009).

[8] J. M. Nissen, R. M. Talbot, A. N. Thompson, and B. Van Dusen, Comparison of normalized gain and Cohen's *d* for analyzing gains on concept inventories, Phys. Rev. Phys. Educ. Res. **14,** 010115 (2018).

[9] L. Bao and E. F. Redish, Model analysis: Representing and assessing the dynamics of student learning, Phys. Rev. ST Phys. Educ. Res. **2,** 010103 (2006).

[10] R. K. Thornton, Conceptual dynamics: Following changing student views of force and motion, AIP Conf. Proc. **399,** 241 (1997).

[11] T. I. Smith and M. C. Wittmann, Applying a resources framework to analysis of the Force and Motion Conceptual Evaluation, Phys. Rev. ST Phys. Educ. Res. **4,** 020101 (2008).

[12] T. I. Smith, M. C. Wittmann, and T. Carter, Applying model analysis to a resource-based analysis of the Force and Motion Conceptual Evaluation, Phys. Rev. ST Phys. Educ. Res. **10,** 020102 (2014).

[13] T. I. Smith, Telling new stories using old data: FMCE edition, in *Proceedings of the 2015 Physics Education Research Conference, College Park, MD*, edited by A. D. Churukian, D. L. Jones, and L. Ding (AIP, New York, 2015), pp. 315–318.

[14] I. T. Griffin, K. J. Louis, R. Moyer, N. J. Wright, and T. I. Smith, A multi-faceted approach to measuring student understanding, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA*, edited by D. L. Jones, L. Ding, and A. Traxler (AIP, New York, 2016), pp. 132–135.

[15] M. C. Wittmann and K. E. Black, Visualizing changes in pretest and post-test student responses with consistency plots, Phys. Rev. ST Phys. Educ. Res. **10,** 010114 (2014).

[16] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, An item response curves analysis of the Force Concept Inventory, Am. J. Phys. **80,** 825 (2012).

[17] P. J. Walter and G. Morris, Assessing student learning and improving instruction with transition matrices, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA*, edited by D. L. Jones, L. Ding, and A. Traxler (AIP, New York, 2016), pp. 376–379.

[18] M. Ishimoto, G. Davenport, and M. C. Wittmann, Use of item response curves of the Force and Motion Conceptual Evaluation to compare Japanese and American students' views on force and motion, Phys. Rev. Phys. Educ. Res. **13,** 020135 (2017).

[19] 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, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH*, edited by L. Ding, A. Traxler, and Y. Cao (AIP, New York, 2017), pp. 380–383.

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

[21] 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.

[22] Y. Suh and D. M. Bolt, Nested logit models for multiple-choice item response data, Psychometrika **75,** 454 (2010).

[23] R. P. Chalmers, mirt: A multidimensional item response theory package for the R environment, J. Stat. Softw. **48** (2012).

[24] R. P. Chalmers, Multidimensional Item Response Theory (mirt) (2017), R package version 1.29.

[25] K. J. Louis, B. J. Ricci, and T. I. Smith, Determining a hierarchy of correctness through student transitions on the FMCE, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC*, edited by A. Traxler, Y. Cao, and S. Wolf (AIP, New York, 2018).

[26] Analyses of data from School 1–2 was published in Ref. [12].

[27] PhysPort, Data Explorer (2011–2017), https://www.physport.org/DataExplorer/Preview.cfm, Last accessed on Mar. 20, 2019.

[28] F. B. Baker, *The Basics of Item Response Theory* (ERIC Clearinghouse on Assessment and Evaluation, College Park, MD, 2001).

[29] R. J. de Ayala, *The Theory and Practice of Item Response Theory* (Guilford Press, New York, 2008).

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

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

[32] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria 2019).

[33] We chose 7300 response sets to ensure we had at least 10 times as many response sets as parameters, as recommended in Ref. [29]. Samples were created without replacement.

[34] M. Torchiano, effsize: Efficient Effect Size Computation (2018), R package version 0.7.4.

[35] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York, 2016).

[36] T. D. Hocking, directlabels: Direct Labels for Multicolor Plots (2018), R package version 2018.05.22.

[37] The effect size between C and H is also in the ambiguous range: $g = 1.20$.

[38] The label for each curve is placed as close to the peak as possible without overlapping with either another label or another curve.

[39] The vertical axis in Figs. 2, 4, and 5 is probability density, which is normalized based on the visible distribution; therefore, the values on the vertical axis should not be compared between plots.

[40] Learning Assistant Alliance, Learning About STEM Student Outcomes (LASSO) (2012–2017), https://learningassistantalliance.org/modules/public/lasso.php, Last accessed on Feb. 27, 2018.

[41] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66,** 64 (1998).

[42] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika **12,** 153 (1947).

[43] A. H. Bowker, Test for symmetry in contingency tables, J. Am. Stat. Assoc. **43,** 572 (1948).

[44] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.16.010107, for a physicist's interpretation of the parameters of the nominal response model in the context of the canonical ensemble.