Generating a growth-oriented partial credit grading model for the Force Concept Inventory

Philip Eaton,* Keith Johnson, and Shannon Willoughby

*Department of Physics, Montana State University,
1325-1399 S. 6th Avenue, Bozeman, Montana 59715, USA*

(Received 30 July 2019; published 23 December 2019)

Traditionally, multiple choice assessments are graded in a dichotomous manner, where selecting the correct option for a question awards 1 point and the selection of an incorrect option awards 0 points. As a result of this grading scheme, all incorrect response options are treated as being equally incorrect regardless of potential differences in their relative correctness, intentional or otherwise. We propose a partial credit grading model for the Force Concept Inventory (FCI) that allots points to the incorrect responses. This was done using slope parameters from the two-parameter logistics nominal response model (2PLNRM), a model from polytomous item response theory (PIRT). The resulting scores from the partial credit model represent student growth towards a proper Newtonian mindset, as measured by the FCI. Observations indicate that this model accounts for student progression through prominent misconceptions (i.e., impetus) as their world views become more Newtonian. As expected, we find that student total scores increase as a result of the model, but the average overall gains on the assessment are essentially unchanged. The data used in this analysis were maintained and organized by PhysPort and included about 20 000 responses from first semester introductory physics courses at multiple universities. Ultimately, this partial credit model allows instructors to more accurately gauge the growth of their students over the course of instruction. Additionally, as a result of these partial credit scores, we are able to identify potentially malfunctioning questions on the FCI that may be sources of error in measuring student abilities.

DOI: [10.1103/PhysRevPhysEducRes.15.020151](https://doi.org/10.1103/PhysRevPhysEducRes.15.020151)

I. INTRODUCTION

Traditionally, grading a multiple-choice assessment is done in a dichotomous manner, i.e., the student responses are either correct or incorrect. This method of grading places all of the incorrect options (otherwise called distractors) within an assessment on the same level of “incorrectness” by assigning them all scores of zero. However, dichotomous scoring neglects the possibility that there could exist varying degrees of incorrectness within the distractors themselves. In this paper, we propose a partial credit model that will assign nonzero values to the distractors of any given question for the well-vetted 1995, 30-question version of the Force Concept Inventory (FCI) [1], which can be found at PhysPort [2] or in Ref. [3].

Some questions on the FCI that are said to probe force identification contain incorrect options that can readily be identified as having “levels of incorrectness” [1,4–6]. Consider question 11 where a hockey puck slides with a constant velocity on a frictionless floor. One option states

that there is only a force of gravity acting on the puck and another suggests that the force of gravity and a force in the direction of motion both act on the puck. However, the velocity of the puck is constant, so there is no force in the direction of motion. This implies that the second option could be interpreted as being less correct compared to the first. As a result, a partial credit grading model for the FCI may be appropriate to fairly assess students’ understanding.

Within physics education research (PER), there have been many investigations into statistical properties of the FCI [4–29]. From these studies a wealth of information has been gathered about how the FCI functions and/or about how students are thinking about the FCI. Some of these studies have called into question the validity of the FCI, while others offer support to the statistical rigor of the assessment. It is appropriate to state that the FCI is not perfect. However, with more studies analyzing the statistical nature of the FCI, comes more information on how to properly create a conceptual instrument. This information is vital to the PER community when new assessments are generated in the future, which use the results of these analyses.

Partial credit scoring methods for multiple choice assessments are yet another means by which one may further understanding of student response behavior. Polytomous item response theory (PIRT) is a framework which may allow for partial credit modeling. Specifically, PIRT is a

*philip.eaton@montana.edu

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

model-driven analysis which generates parameters for correct and incorrect response options, and then assigns students a latent trait ability score based on their selection of correct *and* incorrect response options [30]. Using the process proposed by Louis, Ricci, and Smith [17], this study generates a partial credit grading model for the FCI using a two-parameter logistic nominal response model (2PLNRM) from PIRT [31]. The partial credit scoring model present in this study is developed from the 2PLNRM, employed for use on the FCI with a dataset containing 19 745 student responses.

Louis, Ricci, and Smith, in Ref. [17], used the 2PLNRM to propose an ordering for the relative “correctness” of the distractors within individual questions on the Force and Motion Conceptual Assessment [32]. Indicated in their study is an assumption that must be made when using the 2PLNRM in order to generate an ordering for incorrect response options:

Students who choose correct responses on most questions are more likely to choose better (or more sophisticated) incorrect answers (for questions they got incorrect) than students who choose few correct responses.

This assumption makes conceptual sense; however, it may not give a scale of correctness in the strictest sense of the word. As will be seen in the discussion section, some orderings of distractors given by the 2PLNRM do not appear to be ordered in terms of correctness from an expert perspective. Instead, for some questions the ordering of the distractors appears to use alternative world views (i.e., misconceptions) in conjunction with a Newtonian world view. That is, for particular questions the model assumes that the average student will go through a prominent alternative world view, such as impetus [1,11], before they can arrive at the desired Newtonian world view. Thus, the resulting partial credit scores generated using the 2PLNRM inherently carry the interpretation of “progression towards proper Newtonian thinking.”

A partial credit model that gives scores based on how close a student is to the proper world view can be thought of as a growth-oriented grading model. Compared to the stricter dichotomous scoring, partial credit scoring gives students the chance to see how much they have moved towards a proper conceptual framework. For example, a student who receives the same dichotomous total score pretest and post-test on the FCI may not actually be responding to each assessment in an identical manner. This student could have shifted from one incorrect world view into another one while in the process of advancing towards a proper Newtonian world view. This kind of motion between distractors is not captured in dichotomous grading, but can be reflected in a partial credit model. Thus, partial credit scores should not be thought of as an amount of correctness contained in a distractor. Instead, they should be interpreted as how close a student is to proper thinking

while potentially still struggling with more complicated and harder-to-break misconceptions.

To generate a partial credit model of this nature, this study seeks to answer the following research questions:

RQ 1: What ordering for the response options on the FCI can be generated when using a two-parameter logistic nominal response model, and as a result what partial credit grading criteria can be proposed and what information might a partial credit model reveal about the questions on the assessment?

RQ 2: How do student gains on the FCI compare between using dichotomous scoring vs the proposed partial credit scoring?

II. DATA

The data used in this analysis were maintained and organized by PhysPort [2]. PhysPort is an online resource for physics instructors, a part of which is collecting physics education assessments and maintaining a database of student responses for many of the assessments. For this study the data originally consisted of 22 028 algebra- and calculus-based first semester introductory physics student responses, which were matched pre-post. Students with any blank responses pre- or post-test were removed, which left a total of 19 745 student responses.

For this study, it was assumed that as students go through instruction their thinking about Newtonian mechanics becomes more coherent compared to before instruction. Thus, postinstruction student responses were used to create the partial credit grading model, since that data should be the most likely to generate a rigid and coherent model.

III. METHODOLOGY

First a description of the 2PLNRM will be given, followed by a discussion of how the model parameters can be interpreted. Then, the process of how to construct the partial credit model scores is detailed.

A. Two-parameter logistics nominal response model

The two-parameter logistic nominal response model (2PLNRM) was first proposed in 2010 by Suh and Bolt [31]. Their model combines the ability of a traditional two-parameter logistic model (2PLM) to model correct responses [33] and Bock’s nominal response mode (NRM) to model individual response options [34]. The 2PLNRM uses two different equations to model the response options on an assessment. The correct response to a question is modeled by a traditional 2PLM:

$$P(\theta) = P(x = 1|\theta, a, d) = \frac{1}{1 + e^{-(a\theta + d)}},$$

where θ is the latent ability, a is the item discrimination, and d is defined through the item difficulty, which is defined as $\delta = -d/a$. The latent ability, or student ability, is similar to the z scores of the total scores received by the students on the

FIG. 1. Trace lines for question 13. From this plot it can be seen that option D is the correct response. The most common incorrect response is dependent on which part of ability space is being considered. The item parameters used to make this plot were taken from Table III in the Appendix. Below the figure, please find question 13 and the response options that correspond to the trace lines. **FCI 13)** A boy throws a steel ball straight up. Consider the motion of the ball only after it has left the boy's hand but before it touches the ground, and assume that forces exerted by the air are negligible. For these conditions, the force(s) acting on the ball is (are):

- A) a downward force of gravity along with a steadily decreasing upward force.
- B) a steadily decreasing upward force from the moment it leaves the boy's hand until it reaches its highest point; on the way down there is a steadily increasing downward force of gravity as the object gets closer to earth.
- C) an almost constant downward force of gravity along with an upwards force that steadily decreases until the ball reaches its highest point; on the way down there is only a constant downward force of gravity.
- D) an almost constant downward force of gravity only.
- E) none of the above. The ball falls back to ground because of its natural tendency to rest on the surface of the earth.

assessment [33]; technically, they are also parameters of the model that need to be estimated through fitting the model to the data. Item discrimination is related to the slope of the probability curve when $\theta = \delta$. The greater the item discrimination (i.e., slope) for a given item, the more likely that a small increase in student ability could lead to a notably larger chance in getting the correct answer, and vice versa. Item difficulty is the ability associated with a 50% chance in getting the problem correct [19,33]. For an example of the probability line generated from this model see option D (purple, dot-dashed line) in Fig. 1; the item difficulty for option D is -0.194 and the item discrimination is 2.56 . The "0" along the ability axis is set to be the mean of the students' ability scores (i.e., the scale is student centered).

The sum of the probabilities of a student at any ability selecting any response should be equal to 1. As such, the incorrect responses take up the remaining probability left over from the correct response's model. For m incorrect options, a single incorrect response k is modeled as

$$P_k(\theta) = P_k(x = 0|\theta, a, d, \mathbf{ak}, \mathbf{b}) \\ = \left(1 - \frac{1}{1 + e^{-(a\theta + d)}}\right) \frac{e^{(ak_k\theta + b_k)}}{\sum_{i=1}^m e^{(ak_i\theta + b_i)}},$$

where the exponential fraction multiplying $[1 - P(\theta)]$ is the NRM for the distractors. The b_k and ak_k parameters are called the intercept and slope parameters, respectively. An example of the probability lines (also known as trace lines) can be found in Fig. 1 for options A, B, C, and E. All of the item parameters and student ability estimations for this study were performed using likelihood maximization methods in the R software MIRT [35,36]. The details of these parameter estimation techniques are outside of the scope of this study and are left to the numerous articles dedicated to the subject, see Ref. [36] and the references cited within for more details. Since the parameters estimated in this model are generated from student responses, the resulting partial credit model should be thought of as a student-driven model. This implies that it may contain a combination of expert- and novicelike thinking, which is embodied in the assumption presented in the Sec. I.

Within the NRM, slope parameters indicate how related a response option is to the latent trait being measured by the assessment. For example, if $ak_1 > ak_2$ then distractor 1 is said to be more related to the latent trait of the assessment compared [37,38] to distractor 2. However, slope parameters cannot be compared between different questions since the slope parameter scales are unique to each question. Thus, the ordering of the slope parameters implies an ordering of the response options within individual questions, but does not allow for direct comparisons between questions.

This interpretation of the slope parameters assumes that students who select correct options will also tend to select distractors that are more related to the latent trait than students who select fewer correct options [17]. As a result, some partial credit within questions may be allocated based on a student's progression through alternative world views while on a trajectory towards a Newtonian world view (see the discussion of question 17 in Sec. VA). Hence, the partial credit scores built from the slope parameters should be thought of as a growth and learning-based grading model, where scores may not only reflect sophistication in *actual* Newtonian thinking but may also indicate growth *towards* Newtonian thinking via motion through alternative world views.

B. Generation of the partial credit scoring model

Within the 2PLNRM only the incorrect options yield NRM slope parameters because the correct answers are

used to represent the underlying latent trait being measured. In the case of the FCI, the overall latent trait can be said to be “Newtonian-ness,” which is simply the ability for a student to correctly answer questions on the FCI. As a result, only the incorrect responses to a question are numerically comparable in their sophistication toward measuring the latent trait. This lack of slope parameters for numerical comparisons of any correct response to its associated incorrect responses is a shortcoming of the 2PLNRM. However, a slope parameter for all correct responses may be inferred by using an inverted key.

Recall that when using the ordering proposed by the 2PLNRM under the correct key for the assessment, the response options which are the least related to the latent trait (the incorrect responses) can be quantified. An inverted key can instead be constructed and the analysis rerun using the “least Newtonian” options, the options least related to the original Newtonian latent trait, as given by the slope parameters, to represent the underlying trait being measured.

The latent trait examined through the 2PLNRM using the inverted key can be thought of as representing “non-Newtonian-ness” of students. Ideally, the inverted slope parameter for the correct response option, henceforth expressed as ak_{Inv} , will be identified as the least non-Newtonian option for an individual question (i.e., the most Newtonian option). This should be the case since the correct response was assumed to be the most Newtonian option under the correct key, which suggests that it should be identified as the least non-Newtonian option using the inverted key. In summary, it should be expected that the ordering of the response options under the inverted key should simply be the reverse of the ordering when using the correct key. If this is the case, the subtractive difference between the inverted key slope parameters of the correct option and the most Newtonian distractor can be calculated. This difference can be added to the most Newtonian distractor’s slope parameter under the correct key to give a measure for how much larger the correct response’s slope parameter is compared to the most Newtonian distractor. This process can be summarized by using the following:

$$ak_{Cor} = (ak_{Inv\ 1st\ Dist} - ak_{Inv\ Cor}) + ak_{1st\ Dist},$$

where $ak_{1st\ Dist}$ and ak_{Cor} are the slope parameters for the most Newtonian distractor the correct response option under the correct key, respectively, and $ak_{Inv\ 1st\ Dist}$ and $ak_{Inv\ Cor}$ are the slope parameters for the most Newtonian distractor and the correct response option under the inverted key.

Now, slope parameters can be assigned to all of the response options when using a 2PLNRM. From this ordering the assumption that the correct response option is the most Newtonian can be verified. In the case of this study all instances identified the correct option as the most

Newtonian, and all of the response option orderings found using the inverted key were simply mirrored versions of the correct key’s proposed ordering, as anticipated. It should be noted that some questions’ correct response options were only barely identified as the most Newtonian response option. These questions will be discussed in detail in Sec. VB.

Since the slope parameters can be thought of as being related to the correctness of response options, they can in turn be used to generate a partial credit grading model. Recall, the slope parameters for each question are on their own scales, which are unique to each question. Because of the uniqueness of these scales, the slope parameters between questions cannot be directly compared to one another. Assuming all of the correct responses are all equally Newtonian, the ak scale for each question can be rescaled to range from 0 to 1. A value of 1 would be interpreted as being perfectly Newtonian. Partial credit scores less than 1 suggest how close a student is to a proper Newtonian world view as measured by the question. These new scores may allow for the direct comparison of options between questions, however, this is not explored in this study and will be investigated in the future.

IV. RESULTS

Once the slope parameters were estimated for the entire sample, instances where options had similar or identical slope parameters were observed. To enable the statistical comparison of potentially tied options 1000 uniformly drawn classes of 10 000 students were generated and used to estimate the FCI’s questions’ slope parameters. The means and standard deviations of these estimations can be found in Table III in the Appendix. The root mean squared of the error of approximation (RMSEA) was calculated for each of the 1000 uniformly draw classes. The RMSEA’s upper 95% confidence interval had a range of 0.0205 to 0.034 75, a mean of 0.038 58, and a standard deviation of 0.001 26. Acceptable RMSEA values are said to be below 0.06 [39,40]. From this it can be seen that all 1000 classes had acceptable fit with the 2PLNRM.

From these random classes a distribution of ak values were found. Following Louis, Ricci, and Smith’s suggestion [17], any mean ak values which were closer than 0.1 to each other were treated as being the same in the ordering of the options, see Table I. This criterion is purely *ad hoc*, functioning only to generate a visual guide for the ordering of the response options in the partial credit model and has no impact on any estimations performed. Louis, Ricci, and Smith in Ref. [17] present another method for identifying an ordering of the incorrect options using 2PLNRM slope parameters and contingency tables to establish a robust estimation of a “hierarchy of correctness” between incorrect response options. Since the analysis method presented in this study used only the slope parameters from a

TABLE I. Ordering of response options as given by the 2PLNRM. A difference of less than 0.10 in the mean slope parameter, see Table III in the Appendix, was treated as a tie in the ordering.

Questions	Ordering of response options								
1	B	=	E	<	A	=	D	<	C
2	C	<	D	<	B	=	E	<	A
3	E	=	B	<	D	<	A	<	C
4	C	<	B	<	D	<	A	<	E
5	C	<	E	<	D	<	A	<	B
6	D	<	E	<	C	<	A	<	B
7	D	<	A	<	C	<	E	<	B
8	C	<	A	<	D	=	E	<	B
9	B	=	D	<	A	<	C	<	E
10	D	<	E	<	C	<	B	<	A
11	B	<	C	<	E	<	A	<	D
12	E	<	D	<	C	<	A	<	B
13	E	<	B	<	A	<	C	<	D
14	E	<	A	<	B	<	C	<	D
15	E	<	D	<	B	<	C	<	A
16	D	<	B	<	C	<	E	<	A
17	C	<	E	=	D	<	A	<	B
18	C	<	E	<	A	=	D	<	B
19	B	<	D	=	A	<	C	<	E
20	A	<	E	<	C	<	B	<	D
21	A	<	B	<	D	<	C	<	E
22	E	<	C	<	D	<	A	<	B
23	E	=	A	=	D	<	C	<	B
24	D	=	E	<	B	<	C	<	A
25	B	<	A	=	E	<	D	<	C
26	C	<	A	<	B	<	D	<	E
27	E	<	B	<	A	<	D	<	C
28	B	<	A	<	C	<	D	<	E
29	C	<	A	<	E	<	D	<	B
30	B	<	E	<	D	<	A	<	C

2PLNRM, and not contingency tables, a discussion of contingency tables and their uses is left to Ref. [17].

The results of the 1000 class item parameter estimations can be found in Table III in the Appendix. The proposed ordering from the ordering of question responses as given by their ak values can be found in Table I. The partial credit scores were found using the mean slope parameter values from the 1000 classes and the described methodology, see Table II.

V. DISCUSSION

The following is a discussion of three questions which had the greatest subtractive difference between their average partial credits and dichotomous scores, and questions on the FCI which the partial credit model suggests may be malfunctioning. The criterion for identifying the individual questions for these discussions are elaborated upon within their respective subsections. Lastly, a discussion of the

TABLE II. The partial credit scoring model generated from the ak values given in Table III in the Appendix. Questions in bold font are ones that are potentially malfunctioning with distracting options having partial credit scores greater than 0.75, and are discussed in Sec. V B.

Questions	A	B	C	D	E
1	0.227	0	1	0.253	0.002
2	1	0.271	0	0.164	0.308
3	0.434	0.004	1	0.084	0
4	0.662	0.320	0	0.368	1
5	0.539	1	0	0.352	0.173
6	0.433	1	0.272	0	0.155
7	0.305	1	0.469	0	0.574
8	0.565	1	0	0.613	0.616
9	0.155	0	0.428	0.05	1
10	1	0.356	0.295	0	0.098
11	0.772	0	0.261	1	0.642
12	0.548	1	0.491	0.059	0
13	0.332	0.242	0.459	1	0
14	0.575	0.686	0.776	1	0
15	1	0.529	0.859	0.116	0
16	1	0.119	0.335	0	0.947
17	0.623	1	0	0.456	0.435
18	0.258	1	0	0.276	0.119
19	0.237	0	0.330	0.199	1
20	0	0.460	0.205	1	0.089
21	0	0.208	0.579	0.394	1
22	0.562	1	0.139	0.421	0
23	0.003	1	0.268	0.008	0
24	1	0.198	0.264	0	0.013
25	0.172	0	1	0.526	0.202
26	0.293	0.371	0	0.425	1
27	0.417	0.253	1	0.546	0
28	0.159	0	0.278	0.446	1
29	0.177	1	0	0.983	0.394
30	0.496	0	1	0.339	0.165

affects the partial credit model has on the total scores of students and classes are outlined.

A. Three most changed questions

The proposed partial credit ordering of the questions, see Table I, may now be qualitatively analyzed to check for consistency. For brevity only the three questions whose partial credit and dichotomous scores have the greatest subtractive difference from one another are discussed. This difference was obtained by generating 1000 classes of 500 uniformly sampled students and comparing the mean average item scores. From this, items 14, 15, and 17 were identified as having the greatest subtractive difference between their partial credit and dichotomous scores, with Cohen's d values of 17.1, 25.2, and 21.3, respectively, on the post-test. Plots of the average items scores can be found in Figs. 7 and 8 in the Appendix.

The remainder of this section is a discussion of these three items and how their option ordering under the partial credit model aligns well with the idea of increasing Newtonian sophistication. Potential student reasoning behind the ordering of a question's options presented below is purely expert speculation. Written responses or student interviews could be used to verify that this explanation aligns with actual student thinking, which is suggested for a future study.

1. Question 14

Question 14, see Fig. 2, involves an airplane which is traveling to the right while carrying a bowling ball. The bowling ball is suddenly released, and the students are asked to identify the path the ball is most likely to take from five options. From the partial credit model, these options can be ordered from least to most Newtonian as E, A, B, C, and D. Option E is identified as the least Newtonian and is a "Wile E. Coyote" like path with the ball traveling to the right without falling, then falling essentially straight down. This is the least physical option, as it is effectively impossible under any circumstance. Thus its placement as the least Newtonian option is appropriate. The remainder of the options progress as A, B, C, and then D, which from Fig. 2 appears to be an advancement of the ball's landing spot from behind (A) to underneath (B) and finally in front of the point of release (C and D). For a student to think that the bowling ball will travel backwards indicates a severe lack of Newtonian understanding; similarly, for option B. Option C is nonphysical, however, it is still more Newtonian than options A and B as the ball actually moves forwards from its point of release. Thus, the ordering of these options does agree with increasing amount of Newtonian sophistication.

2. Question 15

Probing Newton's third law, question 15 asks students about the relation between the force a car exerts on a truck and vice versa, see Fig. 3. Since both the car and truck are accelerating to the right in this problem, students may have issues with this question by inappropriately applying

FIG. 2. FCI question 14 [1]. A plane travels to the right with a velocity of $\vec{v}$ and releases a bowling ball, which it was carrying. Options (A)–(E) indicate potential paths the bowling ball could take while it fall from the plane. The proposed Newtonian ordering for this question is given as $E < A < B < C < D$.

FIG. 3. FCI Question 15 [1]. A car is pushing on a truck and both accelerate to the right. This questions asks students about how the force from the car on the truck relates to the force from the truck on the car. The proposed Newtonian ordering for this question is given as $E < D < B < C < A$.

- A) the amount of force with which the car pushes on the truck is equal to that with which the truck pushes back on the car.
- B) the amount of force with which the car pushes on the truck is smaller than that with which the truck pushes back on the car.
- C) the amount of force with which the car pushes on the truck is greater than that with which the truck pushes back on the car.
- D) the car's engine is running so the car pushes against the truck, but the truck's engine is not running so the truck cannot push back against the car. The truck is pushed forward simply because it is in the way of the car.
- E) neither the car nor the truck exert any force on the other. The truck is pushed forward simply because it is in the way of the car.

Newton's second law when they should be applying Newton's third law. The ordering of the options is given as E, D, B, C, and then A, where A is the correct response. Options E and D suggest that the truck is pushed forward (to the right) simply because it is in the way of the car, and no forces are actually applied. This is clearly not the case and their positions as the least Newtonian responses is understandable. The progression from options B to C to A may be rationalized in the following manner. First, since the truck is bigger (in the figure) than the car, it will apply more force (option B). However, a discerning student may realize that the system is traveling to the right, which means the car "has" to push harder on the truck for the car and truck to accelerate to the right (option C). Finally, to get to option A the student will need to realize the forces in question are third law pairs and must be equal in magnitude. Thus, this ordering of response options makes clear conceptual sense.

3. Question 17

Question 17 nicely exemplifies why the partial credit model is referred to as a growth-based model and not as a correctness model. This question, see Fig. 4, shows an elevator being lifted up at a constant speed by a cable, and asks students about how the forces acting on the elevator relate to one another. The proposed ordering for this question's response options is C, E, D, A, and then B, with C being the least Newtonian. From a correctness point of view this ordering does not make sense, since options C and

FIG. 4. FCI Question 17 [1]. An elevator is lifted up at a constant speed by a cable. This question asks the students about how the force exerted by the cable on the elevator is related to the force of gravity on the elevator and/or the force due to the air. The proposed Newtonian ordering for this question is given as $C < E = D < A < B$.

- A) the upward force by the cable is greater than the downward force of gravity.
- B) the upward force by the cable is equal to the downward force of gravity.
- C) the upward force by the cable is smaller than the downward force of gravity.
- D) the upward force by the cable is greater than the sum of the downward force of gravity and a downward force due to the air.
- E) none of the above. (The elevator goes up because the cable is being shortened, not because an upward force is exerted on the elevator by the cable).

A seem equally incorrect (they both generate a net acceleration). However, when considering this model as representing growth towards a Newtonian mindset this ordering is practical.

The option ordering for question 17 suggests that when students progress from knowing little about mechanics to a Newtonian mindset they will likely move through an impetus misconception ($\vec{F} \propto \vec{v}$). For example, if option C were true then a net downward force would be generated on the elevator, which is incorrect since the elevator is moving upwards at a constant speed. However, from an impetus perspective a net force should be directed upwards as to align itself with the direction of motion regardless if the speed is constant or not, option A. Since both of these options result in a non-zero acceleration, a purely Newtonian mindset will determine these options to be equally incorrect. However, a growth-based mindset suggests that selecting option A over C indicates the student is closer to Newtonian thinking compared to selecting C over A.

This growth is also reflected in this question 17's partial credit trace lines, see Fig. 5. Observing the trace lines of

FIG. 5. Trace lines for question 17. From this plot it can be seen that option B is the correct response, with option A being a strong distractor. The item parameters used to make this plot were taken from Table III in the Appendix.

this question in Fig. 5 shows that when a student with little Newtonian ability, say $\theta = -3$, progresses towards a higher ability score, the most likely selected response option changes from option C to option A. Around $\theta = 0.5$, option B (the correct option) overtakes option A and its probability of being selected grows monotonically for the rest of the ability space. Thus, changing from a response of C to a response of A can be interpreted as advancing toward a higher student ability score, and can thus indicate a growth towards proper Newtonian thinking.

Overall, question 17 nicely demonstrates the growth-based learning behavior of the partial credit model. Response C is being treated as the most unsophisticated as it does not represent Newtonian views nor does it align with students who still possess the impetus misconception. Response E represents students who may now begin to reflect an impetus mindset, but do not display much if any Newtonian thinking. Options D and A would both generate an upwards acceleration, which ignores the fact that the elevator is moving with a constant speed, but does make sense when viewed through an impetus lens. Option D compared to option A contains an additional incorrect force (downward force of air), and thus can be interpreted as being less sophisticated than option A based on “Newtonian-ness.” Finally, the correct response B is more likely to be chosen once impetus views are overcome and Newtonian thinking is achieved.

B. Potentially malfunctioning questions

In addition to discussing the most affected questions due to the partial credit model, it is worth discussing questions which the model suggests may be malfunctioning. In this context, malfunctioning implies that the question is not functioning as originally intended. For example, a question

could be malfunctioning if students are not interpreting the written statement properly, are not selecting response options based on the expert intent reasoning, etc. Potentially malfunctioning questions can be identified by looking for large partial credit scores [38]. For instance, if all options on a question were equally spread out in terms of Newtonian-ness then the partial credit scores would come out as 0, 0.25, 0.5, 0.75, and 1. So, if a distractor has a partial credit score above 0.75, then it may be an indication that the associated question is not being interpreted properly or is malfunctioning in some other manner. Looking over the partial credit scores in Table II shows that questions 11, 14, 15, 16, and 29 all have distractors with scores greater than 0.75. It should be noted that this method is not a “catch all” and that some items which pass this investigation may still be malfunctioning. This is simply a tool that should be used in conjunction with the many other quantitative and qualitative tools used to analyze assessments.

1. Question 11

Question 11 asks about the forces acting on a puck which is sliding at a constant speed on frictionless ground. The response options given are lists of potential forces that may be acting on the puck. For example, the correct answer, option D, has the downward force of gravity as well as the normal force from the ground. However, option A is given a large partial credit score of 0.772. This option offers only the downward force of gravity and does not include the normal force from the ground. Since this option excludes the normal force and does not include needless forces like options B and C, a score of 0.772 appears reasonable. As a result, question 11 is not believed to be malfunctioning, and simply contains well-separated response options in terms of their relative Newtonian-ness.

2. Question 14

A similar discussion can be held for option C on question 14, which received a score of 0.776. Recalling the previous discussion about this question, Figure 2 shows that option C is the only reasonable distractor since it gets the landing spot of the bowling ball correct. When compared to options A and B, the fact that path C has the ball continue in the same direction as the plane after release, and lands in a reasonable location, justifies the large partial credit score this option received. Thus, question 14 is determined to not be malfunctioning.

3. Question 15

Question 15 was previously discussed and has been identified as potentially malfunctioning, with option C being awarded a partial credit score of 0.859, which is generous for an incorrect response. Recall this question has a car pushing a truck to the right while both are accelerating

to the right. Option C suggests that the force from the car on the truck should be greater than the force of the truck onto the car. This misconception is typically identified as either active agent (the object doing the action applied a greater force) or impetus ($\vec{F} \propto \vec{v}$) [1]. However, due to the high score of this question a more Newtonian explanation could be the issue. For example, students could be thinking in terms of Newton’s second law when they should be thinking on terms of Newton’s third [41]. This could lead semi-Newtonian students to select option C over the others since it appears to generate a net force in the direction of acceleration. So, option C could be being awarded points based on inappropriate utilization of Newton’s second law. This suggests that this question is not malfunctioning, and is just a difficult question for students in general. However, one could argue that it is malfunctioning, since students may not be selecting a distractor based on the intended interpretation of the distractor (i.e., using Newton’s second law inappropriately vs using impetus).

4. Question 16

Question 16 uses the same set up and response options as question 15, however, now the car and truck are cruising at a constant speed. Because of this slight modification in the situation, option E is now awarded a partial credit score of 0.947. This question has been identified in the past as malfunctioning [4,28,29]. It has been observed that students are able to answer this question correctly through the use of Newton’s first law, not through Newton’s third [4]. If the car and truck are moving at a constant speed, and there are no resistive forces, they would not apply forces onto one another. The large score for option E corroborates these results. This question could potentially be fixed by including a statement that friction still acts on the car and truck. However, until this question is modified and reexamined it must be concluded that question 16 is malfunctioning.

5. Question 29

The last of the potentially malfunctioning questions to be examined is question 29, which gives option D a score of 0.983. Question 29 is a force identification question which asks about a chair sitting on the floor and offers choices for the possible forces which could be acting on the chair. These forces include gravity, a normal force from the floor, and a downward force due to the air. Option D includes all of these forces, whereas the correct response (option C) only contains gravity and the normal force from the floor. This question was commented on by Morris *et al.* in Ref. [12] where they pointed out that question 29 “... extends beyond the Newtonian thinking dimension that the FCI is designed to measure ...”. Further, in an item response theory analysis of the FCI, this question was identified as being the least discriminating question on the FCI [19]. As a result of students’ potential lack of familiarity with some of the content probed by this item,

option D is awarded an unreasonable amount of partial credit. It is suggested that question 29 be examined in more depth by following the procedure used in Ref. [28], and by conducting student interviews or giving free response questionnaires for this question. Until a study of this nature is done, question 29 can only be concluded to be potentially malfunctioning for currently unknown reasons.

C. Comparing the dichotomous and proposed partial credits scoring models

The affect of the partial credit model on the total scores on the FCI can be found in Fig. 6; the red lines are one-to-one correspondence lines. From (a) and (b) in Fig. 6, it can be seen that all of the scores were improved as a result of the partial credit model. This is to be expected since the partial credit model is assigning points for responses which would have originally been given zero points. Additionally, the poorer performing students benefit more from the partial credit compared to the top performing students. Because of the increase in students' pretest and post-test scores, concerns may be raised that an individual student's gain will change appreciably.

One gain that has been seen to be correlated to the pretest scores is the Hake gain [42,43]. From (c) in Fig. 6 it can be seen that there is no drastic change to the Hake gain's values. A linear regression of the partial credit Hake gain to the dichotomous Hake gain gives a slope of 1.02 and an

intercept of 0.02 with an adjusted r -squared value of 0.91. So, the average Hake gain is effectively unchanged with partial credit scoring. Similar conclusions can be made with the normalized gains [42] and can be found in (d) of Fig. 6. Therefore, researchers and instructors could use the partial credit scoring presented in place of dichotomous scoring, and should expect only small differences between the class gains found for each scoring method.

However, since a unidimensional score for a multi-concept assessment is likely to be misleading for students and instructors [44], the aggregate partial credit total score should not be presented to the students. In fact, under no means should students be given their scores for an individual question in any conceptual assessment used in PER. If this kind of information became widespread these assessments would be invalidated as usable research tools. We suggest supplying the students their normalized gain using the partial credit scores, since this will be a better measure of their Newtonian growth over the course of instruction.

VI. LIMITATIONS AND FUTURE WORKS

There exist multiple methods which can be used to obtain orderings for response options, such as contingency tables [45], 2PLNRM slope parameter values, student interviews, expert orderings, etc. This study only used one of these possible methods. The other methods mentioned make different assumptions about how the ordering of response options can be obtained, which may generate slightly different orderings as a result. These other methods could be used to refine this partial credit scoring model. An example of a more rigorous treatment of finding the ordering of response options can be found in Ref. [46] for the force and motions conceptual evaluation. Something similar to this study could be performed for the FCI, and other multiple choice conceptual assessments.

Since this study used data which came from both algebra- and calculus-based physics classes, the results may be an average of these two populations. Applying the method presented in this study to samples of just algebra- and calculus-based classes would allow for comparisons to be made between the populations which make up those courses. Any differences found in the partial credit models may illuminate differences in thought processes between the samples. Similar studies could be done separating based on gender, ethnicity, etc.

It is worth noting that this partial credit model does not fix all of the known issues the FCI has been determined to have from previous research. However, the validity of the proposed partial credit methodology is independent of the functionality of the FCI. Also, given the sensibility of the results found in this study, and the limited number of potentially malfunctioning items, we believe the presented results are reliable. If a significant

FIG. 6. Plots comparing the results for the FCI using dichotomous scoring and the proposed partial credit scoring. (a),(b) Partial credit total scores vs the dichotomous total scores for students on the FCI pretest and post-test, respectively. (c),(d) Partial credit's vs the dichotomous Hake gain and normalized gains, respectively. The red lines in all of the plots are a slope equals 1 line to be used for reference.

number of the items had been found to be malfunctioning, the generated model's reliability would need to be called into question.

Future work could be done by applying the methodology of this study to generate partial credit scoring models for other commonly used assessments in physics education research.

VII. CONCLUSIONS

RQ 1: What ordering for the response options on the FCI can be generated when using a 2-parameter logistic nominal response model, and as a result what partial credit grading criteria can be proposed and what information might a partial credit model reveal about the questions on the assessment?

This study sought to create a partial credit grading model for the Force Concept Inventory. To do this a two-parameter logistic nominal response model was used to estimate the slope parameters for all of the distractors on the assessment. Using an inverted key (a key with all of the least correct responses given by the model being treated as the correct answers) gave a measure for the slope parameter for the correct responses. Under the assumption that students who get more questions correct will tend to select "more correct" distractors compared to students who got few items correct, an ordering for the response options was obtained, see Table I. The slope parameters were then adjusted to make the partial credit model by translating the slope parameter values to make the least Newtonian distractor's slope parameter zero, then rescaling so that the maximum value on the scale was 1 for all of the questions. These values can be found in Table II.

An interesting result from the presented analysis was its ability to detect potentially malfunctioning questions [38]. Questions 11, 14, 15, 16, and 29 were all found to be potentially malfunctioning, but upon investigation only questions 16 and 29 were determined as likely malfunctioning, which is in agreement with previous research results [4,12,28].

RQ 2: How do student gains on the FCI compare between using dichotomous scoring vs the proposed partial credit scoring?

As one would expect, the effect of the partial credit scoring resulted in all of the students receiving higher scores on the FCI pretest and post-test compared to their dichotomous scores, with the expectation of perfect scores, see Fig. 6. However, the average Hake and normalized gains were relatively unaffected. So, individual students can be given a better indication of how much they have progressed towards Newtonian thinking, while not overtly affecting the gains measured for an entire class.

As a result of this study, educators who do not know polytomous item response theory (PIRT) can now use results derived from PIRT without having to know the ins and outs of PIRT. This gives instructors the ability to assign model driven scores to distractors to give students a better sense of where they are on their path towards proper Newtonian thinking, as measured by the FCI. This would be useful in classrooms as students will be able to see their individual growth towards a more Newtonian world view through their different selections of the distracting options from pre- to post-test on the FCI.

ACKNOWLEDGMENTS

The authors would like to thank PhysPort for allowing us to use their data. This project was funded by the Montana State University Physics Department.

APPENDIX: TABLES AND FIGURES

The following are complementary tables and figures for readers interested in the fine details of the results of the 2PLNRM and how the average score on items changed due to the partial credit scoring criteria proposed. Table III displays the 2PLNRM estimated parameter results for all of the items on the FCI. Figures 7 and 8 contain plots of the average scores for each item post and preinstruction (in that order); red represents the average score using dichotomous scoring and blue using partial credit scoring. These figures show how much the average performance of each item is affected due to the introduction of the partial scoring criteria.

TABLE III. Two-parameter logistics nominal response model's item parameters given as mean (standard deviation) for the 1000 uniformly sampled classes of 10 000 students. The ak_1 and b_1 values are taken to be zero in the model to set the measurement scale of the other parameters. The last three columns give the relative distance of the correct option from the nearest incorrect option in ak space, as given by the inverted key item estimations. The standard deviations for $\Delta ak_{inv\ cor}$ were calculated using $\sigma = \sqrt{\sigma_{correct\ option}^2 + \sigma_{closest\ distractor}^2}$, where σ is the standard deviation.

Q	a	d	ak ₁	ak ₂	ak ₃	ak ₄	b ₁	b ₂	b ₃	b ₄	ak _{inv cor}	ak _{inv 1stdis}	Δak _{inv cor}
1	1.137(0.027)	1.905(0.026)	0(0)	-0.319(0.061)	0.037(0.056)	-0.316(0.103)	0(0)	-0.863(0.066)	-0.188(0.055)	-1.937(0.112)	-1.098(0.038)	-0.047(0.057)	1.051(0.069)
2	0.937(0.019)	0.028(0.017)	0(0)	-0.33(0.06)	-0.131(0.028)	0.045(0.061)	0(0)	-1.678(0.048)	-0.116(0.024)	-1.961(0.048)	0(0)	0.842(0.054)	0.842(0.054)
3	1.161(0.023)	0.801(0.019)	0(0)	-0.699(0.042)	-0.569(0.064)	-0.706(0.073)	0(0)	-1.087(0.042)	-2.222(0.064)	-2.208(0.071)	-0.921(0.025)	0(0)	0.921(0.025)
4	0.963(0.023)	1.037(0.02)	0(0)	-0.923(0.131)	-1.788(0.194)	-0.795(0.172)	0(0)	-3.935(0.152)	-5.258(0.279)	-4.121(0.182)	-0.911(0.023)	0(0)	0.911(0.023)
5	1.426(0.027)	-0.343(0.02)	0(0)	-1.116(0.066)	-0.387(0.056)	-0.757(0.062)	0(0)	0.419(0.046)	1.547(0.035)	0.619(0.041)	-0.956(0.045)	0(0)	0.956(0.045)
6	1.516(0.039)	2.518(0.037)	0(0)	-0.401(0.093)	-1.083(0.151)	-0.694(0.238)	0(0)	-1.812(0.105)	-3.654(0.211)	-3.804(0.297)	-1.421(0.041)	0(0)	1.421(0.041)
7	1.16(0.028)	1.727(0.026)	0(0)	0.332(0.058)	-0.617(0.098)	0.545(0.057)	0(0)	-0.069(0.058)	-1.957(0.121)	0.134(0.056)	-1.405(0.047)	-0.542(0.057)	0.863(0.074)
8	1.299(0.027)	0.988(0.021)	0(0)	-1.788(0.17)	0.153(0.05)	0.16(0.042)	0(0)	-4.47(0.251)	-0.364(0.045)	-0.058(0.04)	-1.381(0.035)	-0.165(0.05)	1.216(0.061)
9	0.98(0.021)	0.179(0.018)	0(0)	-0.197(0.054)	0.347(0.051)	-0.134(0.07)	0(0)	1.309(0.046)	1.678(0.042)	-0.181(0.061)	-1.026(0.045)	-0.299(0.045)	0.727(0.064)
10	1.969(0.039)	1.904(0.032)	0(0)	-0.144(0.093)	-0.845(0.093)	-0.612(0.094)	0(0)	0.634(0.082)	0.338(0.087)	-0.092(0.085)	0(0)	1.53(0.077)	1.53(0.077)
11	1.711(0.032)	0.327(0.023)	0(0)	-2.331(0.095)	-1.544(0.065)	-0.394(0.082)	0(0)	-1.648(0.079)	0.756(0.034)	-0.995(0.052)	-0.687(0.046)	0(0)	0.687(0.046)
12	1.261(0.032)	2.028(0.03)	0(0)	-0.129(0.113)	-1.115(0.155)	-1.249(0.172)	0(0)	2.361(0.1)	-0.782(0.171)	-1.755(0.206)	-1.031(0.12)	0(0)	1.031(0.12)
13	2.561(0.05)	0.498(0.029)	0(0)	-0.4(0.087)	0.57(0.081)	-1.481(0.328)	0(0)	0.343(0.086)	2.236(0.073)	-4.097(0.45)	-2.933(0.09)	-0.52(0.078)	2.413(0.119)
14	1.116(0.021)	0.285(0.018)	0(0)	0.373(0.034)	0.672(0.04)	-1.923(0.246)	0(0)	-0.271(0.03)	-0.442(0.032)	-6.367(0.402)	-1.406(0.029)	-0.656(0.037)	0.75(0.047)
15	0.48(0.017)	-0.024(0.014)	0(0)	0.908(0.058)	-1.138(0.143)	-1.457(0.348)	0(0)	2.748(0.055)	-2.372(0.194)	-4.562(0.542)	0(0)	0.389(0.018)	0.389(0.018)
16	1.335(0.031)	1.963(0.028)	0(0)	0.584(0.102)	-0.322(0.147)	2.237(0.145)	0(0)	2.045(0.123)	-0.871(0.191)	1.554(0.131)	0(0)	0.143(0.047)	0.143(0.047)
17	1.212(0.023)	-0.64(0.019)	0(0)	-1.875(0.126)	-0.504(0.044)	-0.566(0.078)	0(0)	-4.955(0.165)	-2.108(0.037)	-3.301(0.066)	-1.134(0.023)	0(0)	1.134(0.023)
18	1.674(0.03)	0.155(0.021)	0(0)	-0.563(0.12)	0.038(0.105)	-0.304(0.108)	0(0)	0.327(0.093)	2.518(0.073)	1.614(0.078)	-1.577(0.1)	0(0)	1.577(0.1)
19	1.27(0.025)	0.774(0.019)	0(0)	-0.386(0.08)	0.151(0.06)	-0.063(0.04)	0(0)	-1.257(0.078)	-0.506(0.053)	0.963(0.038)	-1.243(0.04)	-0.153(0.06)	1.09(0.072)
20	1.17(0.024)	0.334(0.018)	0(0)	0.647(0.053)	0.288(0.035)	0.125(0.076)	0(0)	-0.343(0.043)	0.901(0.034)	-1.52(0.067)	-0.759(0.048)	0(0)	0.759(0.048)
21	0.991(0.021)	-0.253(0.017)	0(0)	0.38(0.055)	1.057(0.052)	0.72(0.057)	0(0)	0.89(0.055)	2.073(0.049)	1.026(0.054)	-1.394(0.036)	-0.625(0.033)	0.769(0.049)
22	1.313(0.024)	0.439(0.02)	0(0)	-1.116(0.093)	-0.373(0.039)	-1.482(0.116)	0(0)	-2.867(0.096)	-0.669(0.032)	-3.841(0.14)	-1.155(0.025)	0(0)	1.155(0.025)
23	1.38(0.026)	0.466(0.02)	0(0)	0.421(0.044)	0.007(0.046)	-0.005(0.073)	0(0)	0.958(0.04)	0.503(0.043)	-0.952(0.066)	-1.58(0.043)	-0.419(0.044)	1.161(0.062)
24	1.703(0.037)	2.034(0.033)	0(0)	0.144(0.093)	-0.431(0.132)	-0.403(0.107)	0(0)	1.734(0.093)	-0.673(0.147)	-0.185(0.114)	0(0)	1.605(0.039)	1.605(0.039)
25	1.399(0.027)	-0.584(0.021)	0(0)	-0.442(0.081)	0.911(0.07)	0.078(0.077)	0(0)	0.177(0.079)	2.769(0.06)	0.826(0.066)	-1.977(0.068)	-0.758(0.063)	1.219(0.093)
26	1.898(0.036)	-1.393(0.027)	0(0)	0.225(0.027)	-0.844(0.075)	0.379(0.037)	0(0)	-0.024(0.021)	-2.763(0.078)	-0.763(0.027)	-2.054(0.038)	-0.397(0.036)	1.657(0.052)
27	1.043(0.023)	0.833(0.019)	0(0)	-0.274(0.038)	0.216(0.072)	-0.696(0.165)	0(0)	-0.925(0.037)	-1.997(0.054)	-3.726(0.171)	-0.992(0.025)	-0.235(0.069)	0.757(0.073)
28	1.437(0.029)	1.084(0.023)	0(0)	-0.38(0.123)	0.285(0.111)	0.687(0.103)	0(0)	-0.005(0.138)	1.116(0.12)	3.144(0.111)	-2.039(0.112)	-0.712(0.111)	1.327(0.158)
29	0.274(0.02)	1.222(0.017)	0(0)	-0.28(0.116)	1.277(0.071)	0.344(0.11)	0(0)	-0.674(0.147)	2.683(0.078)	-0.451(0.119)	-1.553(0.086)	-1.526(0.088)	0.027(0.123)
30	1.495(0.028)	-0.219(0.02)	0(0)	-0.936(0.067)	-0.297(0.077)	-0.624(0.055)	0(0)	0.01(0.052)	-0.394(0.054)	1.807(0.035)	-0.951(0.051)	0(0)	0.951(0.051)

FIG. 7. Average item scores for questions 1–15 pre and post-test. Averages from dichotomous scoring are in red and the partial credit scoring item averages are in blue.

FIG. 8. Average item scores for questions 16–30 pre and post-test. Averages from dichotomous scoring are in red and the partial credit scoring item averages are in blue.

- [1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, *Phys. Teach.* **30**, 141 (1992).
- [2] S. B. McKagan, *PhysPort* (American Association of Physics Teachers, 2011).
- [3] E. Mazur and R. C. Hilborn, Peer Instruction: A user's manual, *Phys. Today* **50**, No. 4, 68 (1997).
- [4] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, *Phys. Rev. ST Phys. Educ. Res.* **8**, 020105 (2012).
- [5] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, *Phys. Rev. ST Phys. Educ. Res.* **11**, 020134 (2015).
- [6] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, *Phys. Rev. Phys. Educ. Res.* **14**, 010124 (2018).
- [7] D. Huffman and P. Heller, What does the Force Concept Inventory actually measure?, *Phys. Teach.* **33**, 138 (1995).
- [8] D. Hestenes and I. Halloun, Interpreting the Force Concept Inventory: A response to march 1995 critique by Huffman and Heller, *Phys. Teach.* **33**, 502 (1995).
- [9] P. Heller and D. Huffman, Interpreting the Force Concept Inventory: A reply to Hestenes and Halloun, *Phys. Teach.* **33**, 503 (1995).
- [10] T. F. Scott and D. Schumayer, Conceptual coherence of non-Newtonian world views in Force Concept Inventory data, *Phys. Rev. Phys. Educ. Res.* **13**, 010126 (2017).
- [11] T. F. Scott and D. Schumayer, Central distractors in Force Concept Inventory data, *Phys. Rev. Phys. Educ. Res.* **14**, 010106 (2018).
- [12] 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).
- [13] P. Eaton, K. Vavruska, and S. Willoughby, Exploring the preinstruction and postinstruction non-Newtonian world views as measured by the Force Concept Inventory, *Phys. Rev. Phys. Educ. Res.* **15**, 010123 (2019).
- [14] E. Brewé, J. Bruun, and I. G. Bearden, Using module analysis for multiple choice responses: A new method applied to Force Concept Inventory data, *Phys. Rev. Phys. Educ. Res.* **12**, 020131 (2016).
- [15] M. C. Wittmann and K. E. Black, Visualizing changes in pretest and post-test student responses with consistency plots, *Phys. Rev. Phys. ST Educ. Res.* **10**, 010114 (2014).
- [16] G. A. Morris, P. Walter, S. Skees, and S. Schwartz, Transition matrices: A tool to assess student learning and improve instruction, *Phys. Teach.* **55**, 166 (2017).
- [17] 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 2018, Washington, DC* (AIP, New York, 2018).
- [18] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, *Phys. Rev. ST Phys. Educ. Res.* **6**, 010103 (2010).
- [19] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, *Am. J. Phys.* **78**, 1064 (2010).
- [20] 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).
- [21] 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).
- [22] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef, The puzzling reliability of the Force Concept Inventory, *Am. J. Phys.* **79**, 909 (2011).
- [23] C. Henderson, Common concerns about the Force Concept Inventory, *Phys. Teach.* **40**, 542 (2002).
- [24] M. R. Semak, R. D. Dietz, R. H. Pearson, and C. W. Willis, Examining evolving performance on the Force Concept Inventory using factor analysis, *Phys. Rev. Phys. Educ. Res.* **13**, 010103 (2017).
- [25] 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).
- [26] R. P. Springuel, M. C. Wittmann, and J. R. Thompson, Applying clustering to statistical analysis of student reasoning about two-dimensional kinematics, *Phys. Rev. Phys. ST Educ. Res.* **3**, 020107 (2007).
- [27] J. Stewart, M. Miller, C. Audo, and G. Stewart, Using cluster analysis to identify patterns in students' responses to contextually different conceptual problems, *Phys. Rev. Phys. ST Educ. Res.* **8**, 020112 (2012).
- [28] J.-i. Yasuda and M.-a. Taniguchi, Validating two questions in the Force Concept Inventory with subquestions, *Phys. Rev. Phys. ST Educ. Res.* **9**, 010113 (2013).
- [29] J.-i. Yasuda, H. Uematsu, and H. Nitta, Validating a Japanese version of the Force Concept Inventory, *Lat. Am. J. Phys. Educ.* **1**, 6 (2012).
- [30] M. L. Nering and R. Ostini, *Handbook of Polytomous Item Response Theory Models* (Taylor & Francis, London, 2011).
- [31] Y. Suh and D. M. Bolt, Nested logit models for multiple-choice item response data, *Psychometrika* **75**, 454 (2010).
- [32] S. Ramlo, Validity and reliability of the Force and Motion Conceptual Evaluation, *Am. J. Phys.* **76**, 882 (2008).
- [33] R. J. de Ayala, *The Theory and Practice of Item Response Theory* (Guilford Publications, New York, NY, 2013).
- [34] R. D. Bock, Estimating item parameters and latent ability when responses are scored in two or more nominal categories, *Psychometrika* **37**, 29 (1972).
- [35] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2014), <http://www.R-project.org/>.
- [36] R. P. Chalmers *et al.*, MIRT: A multidimensional item response theory package for the R environment, *J. Stat. Softw.* **48**, 1 (2012).
- [37] R. D. Bock and I. Moustaki, 15 item response theory in a general framework, *Psychometrika* **26**, 469 (2006).
- [38] D. Thissen, L. Steinberg, and A. R. Fitzpatrick, Multiple-choice models: The distractors are also part of the item, *J. Educ. Measure.* **26**, 161 (1989).
- [39] L. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Struct. Equ. Modeling* **6**, 1 (1999).

- [40] T. A. Brown, *Confirmatory Factor Analysis for Applied Research*, 2nd ed. (Guilford Publications, New York, NY, 2014).
- [41] D. P. Maloney, Rule-governed approaches to physics-Newton's third law, *Phys. Educ.* **19**, 37 (1984).
- [42] J. D. Marx and K. Cummings, Normalized change, *Am. J. Phys.* **75**, 87 (2007).
- [43] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, *Am. J. Phys.* **73**, 1172 (2005).
- [44] P. Ramsden, *Learning to Teach in Higher Education* (Routledge, London, 2003).
- [45] A. H. Bowker, A test for symmetry in contingency tables, *J. Am. Stat. Assoc.* **43**, 572 (1948).
- [46] T. I. Smith, K. J. Louis, B. J. Ricci IV, and N. Bendjilali, Quantitatively ranking incorrect responses to multiple-choice questions using item response theory, [arXiv:1906.00521](https://arxiv.org/abs/1906.00521).