Showing the dynamics of student thinking as measured by the FMCE

Trevor I. Smith,[1,2] Kerry A. Gray,[3,2] Kyle J. Louis,[1,2] Bartholomew J. Ricci,[1,4] and Nicholas J. Wright[1,2]

[1]*Department of Physics & Astronomy, Rowan University, 201 Mullica Hill Rd., Glassboro, NJ 08028, USA*
[2]*Department of STEAM Education, Rowan University, 201 Mullica Hill Rd., Glassboro, NJ 08028, USA*
[3]*Department of Chemistry & Biochemistry, Rowan University, 201 Mullica Hill Rd., Glassboro, NJ 08028, USA*
[4]*Department of Mathematics, Rowan University, 201 Mullica Hill Rd., Glassboro, NJ 08028, USA*

Using data from over 14,000 student responses we create item response curves, fitted to the polytomous item response theory model for nominal responses, to evaluate the relative "correctness" of various incorrect responses to questions on the Force and Motion Conceptual Evaluation (FMCE). Based on this ranking of incorrect responses, we examine individual students' pairs of responses to FMCE questions, using transition matrices and consistency plots, to show how student ideas develop over the span of an introductory mechanics course. Using data from two different schools ($N \approx 200$ each), we explore how these representations can show student learning even when individuals do not choose the correct answer. Comparing response pairs provides a rich picture of student learning that is unavailable in many traditional analyses.

I. INTRODUCTION

Research-based assessment instruments (RBAI), such as the Force and Motion Conceptual Evaluation (FMCE)[1] and the Force Concept Inventory (FCI)[2], have persisted as standard tools for measuring student learning gains during introductory physics courses [3]. Common practice for instructors and researchers is to administer an RBAI at the beginning and end of a course, determine which questions each student answered correctly, and report a measure of growth (typically normalized gain or effect size). This is a fairly quick and easy way to measure student gains, which has been made even easier by PhysPort.org, which hosts many RBAIs and offers to analyze student responses via its Data Explorer.

Unfortunately, all incorrect answers are treated equally in typical analyses, and little to no attention is paid to why a student would select a particular incorrect answer. These concerns have been addressed by various researchers: Smith, Wittmann, and Carter interpreted the most common incorrect response to each FMCE question through a resources perspective and used Model Analysis to show shifts in modes of student thinking [5, 6], Thornton identified several "student views" based on interpretations of various incorrect responses on a small subset of FMCE questions [7], and more recently Walter and Morris have ranked incorrect responses to FCI questions (from most to least sophisticated) using Item Response Curves (IRCs) and shown how students progress through answer choices using transition matrices [8].

In this paper we present an initial ranking of incorrect responses to questions on the FMCE based on IRCs from over 14,000 student responses. We combine these rankings with our previously reported use of consistency plots to produce a rich picture of the dynamics of student thinking [9]. One exciting and important feature of this analysis is the ability to document a student's growth and learning even if that student never selects the correct answer to a question.

TABLE I. Comparison of parameters from the 3PL IRT model. Average parameter values are reported with the standard error. FCI results from Ref. [4]. Student's t-test was used to compare the distributions; the associated p-values are reported along with Hedges' g as a measure of effect size.

	a	b	c
FCI	1.1 ± 0.1	0.1 ± 0.2	0.14 ± 0.01
FMCE	1.3 ± 0.2	0.3 ± 0.1	0.012 ± 0.004
p-value	0.03	0.23	< 0.001
effect size	0.3	0.2	0.5

II. ITEM RESPONSE THEORY

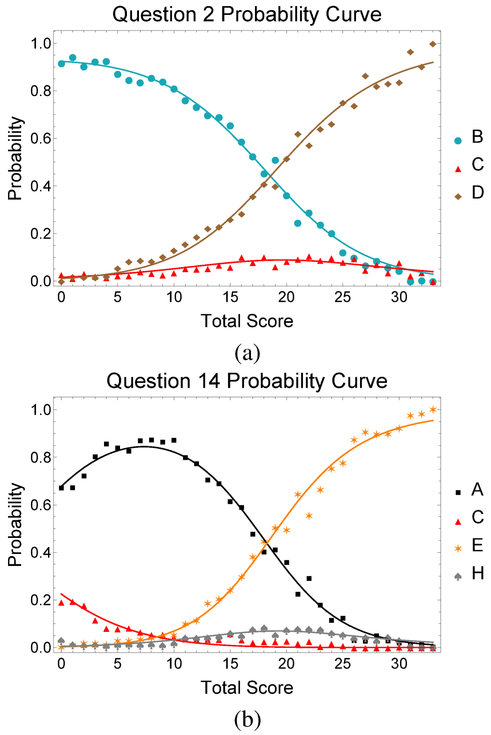
Item response theory (IRT) assumes that student responses to individual questions depend on a latent trait (often referred to as "ability" or "proficiency") [10]. Wang and Bao used the three-parameter logistic (3PL) IRT model to examine student responses to the FCI [4]. In the 3PL model the probability of a student answering a question correctly is

$$P(\theta) = c + \frac{1-c}{1+e^{-1.7a(\theta-b)}}, \tag{1}$$

where θ is the student parameter (or latent trait). The a parameter indicates the discrimination of the question, b represents the difficulty, and c is the probability of a student guessing the correct answer.

Table I shows a comparison between the FCI results found by Wang and Bao and our results on the FMCE. Data come from over 7,000 students from multiple institutions. Following standard practice in IRT analyses, pretest and post-test data were combined, resulting in over 14,000 response sets [11]. As shown in Table I, our discrimination (a) and difficulty (b) parameters for the FMCE are fairly similar to those found by Wang and Bao for the FCI: parameter distributions overlap significantly, and effect sizes are small [12]. One may also see that the guessing parameter (c) on the FMCE is significantly lower than that on the FCI, with a moderate effect size. This supports previous claims that the FMCE is more difficult than the FCI for low-scoring students [13].

2017 PERC Proceedings, edited by Ding, Traxler, and Cao; Peer-reviewed, doi:10.1119/perc.2017.pr.090



(a)

(b)

FIG. 1. IRC with PIRT fits for Questions 2 (a) and 14 (b). Total score measured out of 33 points; the vertical coordinate is the conditional probability of choosing a particular answer given that total score.

III. RANKING FMCE RESPONSES

Morris *et al.* went beyond the correct/incorrect binary by creating Item Response Curves (IRCs) showing the conditional probability that a student earning a particular total score on the FCI will choose each answer to a specific question [14]. Given the strong correlation between the student parameter and the total score on the FCI [4], Morris *et al.* use the total score as the independent variable rather than estimations of a latent trait. Walter and Morris expanded on this work by using IRCs to rank incorrect responses on the FCI from more to less sophisticated [8]. Their main claim is that an incorrect response that is more popular among higher-scoring students represents a higher level of understanding than a response that is most popular among lower-scoring students.

Following the methods of Morris *et al.* we created IRCs for each question on the FMCE by plotting the percentage of students who selected each answer based on their total score on the FMCE. We determined students' total scores out of 33 possible points based on typical scoring recommendations for the first 43 questions of the FMCE [13]. Unfortunately, we were unable to use the method described by Walter and Morris to rank incorrect responses to FMCE questions [8]. The inclusion of up to four additional answer choices (nine on some FMCE questions compared to the FCI's five) made it impossible to distinguish between many of the options. To overcome this challenge we fitted the IRCs using the polytomous IRT (PIRT) model for nominal responses. The PIRT model provides the conditional probability that a student with total score s will choose answer k on a particular question:

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

where a_k and b_k are parameters for each answer choice curve, and N is the number of possible answer choices [15]. We used the `NMinimize` function in Wolfram Mathematica to perform a global fit for all N curves to determine the $2N$ parameters for each question. As an illustrative example, Fig. 1 shows the IRCs and PIRT fits for questions 2 and 14. Both questions ask students about an object moving to the right at a steady (constant) velocity, which we define as Case 2 [9, 16]. All answer choices were included in PIRT analyses, but only responses with three data points above 10% are displayed.

As expected, Fig. 1(a) shows that the probability of choosing the correct answer (D, no net force) increases with higher scores; and the probability of choosing the most common incorrect response (B, constant velocity requires constant force [5]) is highest at the lowest scores. The really interesting feature of this plot is the curve for answer C (constant velocity requires decreasing force in the direction of motion); this curve has a peak around a score of 20, indicating that moderate to high scoring students are more likely to choose C than low-scoring students. This suggests that answer C may be a more sophisticated choice than answer B. We rank the answer choices based on the location (score) at which the PIRT curve reaches its maximum value. The B curve peaks at $s = -2.6$, and the C curve at $s = 19.6$; therefore, D>C>B. This negative peak score indicates that the probability of choosing response B gets continually larger as total score gets lower.

Figure 1(b) shows another example of IRCs with PIRT fits. Unlike question 2, question 14 asks students to choose a graph of force vs. time that is associated with this motion. One can certainly see similarities between Figs. 1(a) and 1(b): the correct answer (E) monatonically increases, the most common incorrect answer (A) is most probable at low scores, and the answer suggesting that a constant velocity requires a decreasing force in the direction of motion (H) has

TABLE II. Question 2 Transition Matrices: percentage of students who gave each pre/post answer pair. The best (correct) answer is D, and the worst answer is B. Bold numbers indicate students who chose a better response on the post-test, and italicized numbers indicate students who chose a worse response. Data come from algebra-based introductory mechanics courses at two different schools.

		School 1					School 2		
		Post-test					Post-test		
		D	C	B			D	C	B
Pretest	D	2.3	*0.5*	*0.0*	Pretest	D	5.7	*0.0*	*1.6*
	C	**0.5**	1.4	*1.4*		C	**1.6**	0.0	*0.5*
	B	**20.0**	**4.2**	69.3		B	**46.4**	**4.7**	34.9

a peak probability around $s \approx 20$. The major difference in Fig. 1(b) is the existence of a fourth prominent response (C) whose PIRT fit has a distinctly negative slope at $s = 0$: this suggests that C is a more naïve response than the common incorrect answer (A). In fact, answer C is consistent with a graph of the position of the object as it moves to the right with a constant velocity. Answers A, C, and H have peak probabilities at $s = 7.3$, -11.2, and 19.3, respectively; thus, we rank these responses as, E>H>A>C.

IV. REPRESENTING STUDENT LEARNING

Rankings of incorrect responses allow us to examine students' answers on the pre- and post-tests to gauge how much learning occurred during the semester. Walter and Morris suggest using transition matrices that show the percentage of students who gave each pre/post answer pair for a particular question [8]. Tables II and III show the transition matrices from two different data sets (from different schools) for questions 2 and 14, respectively. The row indicates the pretest response, and the column indicates the post-test response. In order to match other representations we have modified the format of the transition matrices: the correct answers are in the top row and left column, students who choose a better answer on the post-test are shown in bold, and those who choose a worse answer are shown in italics.

TABLE III. Question 14 Transition Matrices: percentage of students who gave each pre/post answer pair. The best (correct) answer is E, and the worst answer is C.

		School 1					School 2			
		Post-test					Post-test			
		E	H	A	C		E	H	A	C
Pretest	E	1.4	*0.0*	*0.9*	*0.0*	E	6.3	*0.0*	*0.5*	*0.0*
	H	**0.9**	0.5	*0.5*	*0.0*	H	**0.5**	0.0	*0.0*	*0.0*
	A	**20.0**	**1.4**	57.7	*5.5*	A	**55.6**	**2.1**	18.5	*1.6*
	C	**1.4**	**0.0**	**3.2**	1.4	C	**6.3**	**0.0**	**1.6**	1.1

Table II shows that about 20% of students at School 1 and 48% of students at School 2 improve to the correct answer on question 2 after instruction. We can also see that about 4–5% of students at each school improve from answer B to answer C, indicating that these students are learning, just not as much as one might hope. Table III shows similar results with about 23% (School 1) and 62% (School 2) of students improving to the correct answer on question 14; and again, 3–5% of students choose a better incorrect response after instruction at each school. Table III also shows that no students at either school who chose either E (correct) or H on the pretest selected C on the post-test for question 14; this supports our claim that H is a more sophisticated response than A or C.

We use consistency plots to simultaneously show students' transitions on both questions (see Fig. 2(a) and 2(b)) [17]. In this representation the row indicates a response to question 2, the column is a response to question 14, the location of a circle represents a given response pair on the pretest, and a triangle shows a response pair on the post-test. Lines connect circles with triangles to form transition arrows from a particular pretest response pair to a post-test response pair. The numbers indicate the percentage of students in a given class who made that particular transition. Squares show percentages of students who gave the same response pair on both the pre- and post-tests. As with our transition matrices, the correct responses are in the top row and left column, and the worst responses are in the bottom row and right column.

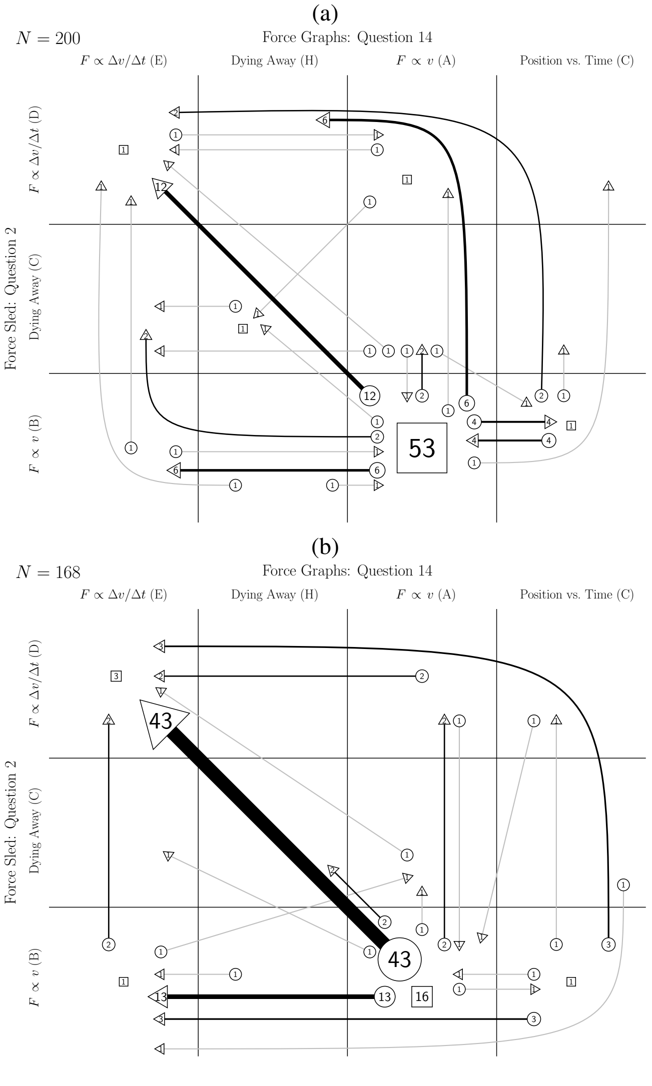


FIG. 2. Case 2 Consistency Plots for Schools 1 (a) and 2 (b).

Many of the observable trends from the transition matrices are visually apparent on the consistency plots: most students at School 1 do not change responses (squares Fig. 2(a)), the most dominant transition at School 2 is becoming correct on both questions (Fig. 2(b)), and getting better is more common than getting worse (more/thicker arrows up and left than down and right on both plots). Consistency plots also make many trends more salient. Carefully examining the transition matrices from School 2 reveals that more students increased to correct on question 14 than question 2; this is readily apparent in Fig. 2(b) with a large 13% arrow going left from cell BA, but only a small 2% arrow going up from the same starting cell. As mentioned above, 20% of students at School 1 transitioned from the most common incorrect answer to correct on each question, but Fig. 2(a) shows that only 12% did so on both questions; one can also see that students at School 1 are about equally likely to become correct on either question 2 or question 14. Figure 2(a) also shows one of the most important features of consistency plots: a cycle in the lower right corner involving 8% of students in which half of them go from C to A on question 14, and half of them make the exact opposite transition. The dynamics of student understanding visible on these consistency plots provide a rich picture of student learning that is unavailable in other representations, and utilizing the rankings from the IRCs ensures that "motion" left and/or up always indicates improvement.

V. SUMMARY AND FUTURE DIRECTIONS

Results from the 3PL IRT model show that the discrimination (a) and difficulty (b) parameters on FMCE questions are similar to those reported for the FCI, but the guessing parameter (c) is notably lower on the FMCE, supporting previous claims that the FMCE is more difficult than the FCI for low-scoring students [4, 13]. Creating IRCs and fitting the data to the PIRT model allows us to uniquely rank incorrect responses to each FMCE question based on the location of the peak probability [14, 15]. Considering pairs of responses provides a rich picture of how individuals change over a course either on a single question (transition matrices [8]) or on matched question pairs (consistency plots [17]). These representations allow us to quickly examine dominant transition patterns. This goes beyond the typical correct/incorrect binary and allows us to show that student learning is occurring even if students do not select the correct answer.

We still have two open questions: 1) How can we consider student transitions on more than two questions simultaneously? We have previously defined question cases that are each comprised of one question from each of the Force Sled, Force Graphs, and Acceleration Graphs question clusters (e.g., questions 2, 14, and 26 are defined as Case 2 [9, 16]), but consistency plots can only show two of these at a time. Latent transition analysis (LTA) may be used to create transition matrices showing the probabilities of students changing from a particular answer triad on the pretest to another triad on the post-test [18]. Preliminary results of LTA show that the dominant transitions are consistent with those identified by our consistency plots. 2) What makes one incorrect response "better" than another? We have described one way to answer our second question based on identifying the score at which the probability of choosing each answer is maximized. Another method for ranking incorrect responses could involve using the opinions of expert physics educators and physics education researchers. A third method could examine the likelihood that students would transition to the correct answer after choosing each of the incorrect answers (as suggested in Ref. [7]). A combination of these may be required to determine a robust ranking for each question, which is necessary for the claim of being able to show student learning and improvement without selecting the correct response.

ACKNOWLEDGMENTS

We are deeply indebted to Sam McKagan and Ellie Sayre for creating the PhysPort Data Explorer and allowing us access to over 6,000 anonymous students' FMCE responses. This project was partially supported by the National Science Foundation through a PhysTEC comprehensive site grant.

[1] R. K. Thornton and D. R. Sokoloff, Am. J. Phys. **66**, 338 (1998).
[2] D. Hestenes, M. Wells, and G. Swackhamer, Phys. Teach. **30**, 141 (1992).
[3] A. Madsen, S. B. McKagan, and E. C. Sayre, Am. J. Phys. **85**, 245 (2017).
[4] J. Wang and L. Bao, Am. J. Phys. **78**, 1064 (2010).
[5] T. I. Smith and M. C. Wittmann, Phys. Rev. ST Phys. Educ. Res. **4**, 020101 (2008).
[6] T. I. Smith, M. C. Wittmann, and T. Carter, Phys. Rev. ST Phys. Educ. Res. **10**, 020102 (2014).
[7] R. K. Thornton, AIP Conf. Proc. **399**, 241 (1997).
[8] P. J. Walter and G. Morris, in *2016 PERC Proc.*, edited by D. L. Jones, L. Ding, and A. Traxler (2016) p. 376.
[9] I. T. Griffin, K. J. Louis, R. Moyer, N. J. Wright, and T. I. Smith, in *2016 PERC Proc.*, edited by D. L. Jones, L. Ding, and A. Traxler (2016) p. 132.
[10] F. B. Baker, *The Basics of Item Response Theory* (ERIC Clearinghouse on Assessment and Evaluation, 2001).
[11] To compare our findings to those of Wang and Bao, we adjusted our parameters to align with student scores having an average value of zero and a standard deviation of unity.
[12] The $p < 0.05$ for a does not seem meaningful in context.
[13] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Phys. Rev. ST Phys. Educ. Res. **5**, 010105 (2009).
[14] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, Am. J. Phys. **80**, 825 (2012).
[15] D. Thissen, L. Cai, and R. D. Bock, in *Handbook of polytomous item response theory models*, edited by M. L. Nering and R. Ostini (Routledge, New York, 2010) p. 43.
[16] T. I. Smith, in *2015 PERC Proc.*, edited by A. D. Churukian, D. L. Jones, and L. Ding (2015) p. 315.
[17] M. C. Wittmann and K. E. Black, Phys. Rev. ST Phys. Educ. Res. **10**, 010114 (2014).
[18] G. Davenport, *Detecting Conceptual Change with Latent Transition Analysis*, Ph.D. thesis, University of Connecticut (2016).