Determining a hierarchy of correctness through student transitions on the FMCE

Kyle J. Louis,[1,2] Bartholomew J. Ricci,[1,3] and Trevor I. Smith[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 Mathematics, Rowan University, 201 Mullica Hill Rd., Glassboro, NJ 08028, USA*

Using data from over 14,000 student responses, we rank incorrect responses on the Force and Motion Concept Evaluation (FMCE). We develop a hierarchy of responses using item response theory and the McNemar-Bowker chi-square test for asymmetry. We use item response theory (IRT) under the assumption that students who score well have a greater understanding of physics than those who do not; therefore, responses that have a greater likelihood of being selected by those who score well are considered better responses. We use the McNemar-Bowker chi-square test (MB) under the assumption that student understanding is more likely to increase than decrease after an introductory mechanics course. Therefore, more dominant transitions from one answer to another from pretest to posttest indicate that one answer is better than another. We present the results from the IRT and MB analyses, highlighting both agreement and disagreement between the hierarchies of responses generated by each.

I. INTRODUCTION

Research-based assessments instruments (RBAIs) are commonly used to measure growth in students' conceptual understanding of physics [1]. Responses are typically scored as correct or incorrect, and a measure of growth is reported based on the number of questions answered correctly before and after instruction. The most commonly reported measure of growth is the normalized gain, but effect size measures (such as Cohen's d) have become more popular due to evidence that the normalized gain metric is biased against students with less prior exposure to physics [2]. Both of these measures share the shortcoming that they ignore the specific incorrect responses that students choose. The power of RBAIs to measure student understanding comes from the fact that the questions and response choices were developed based on research into students' understanding of relevant topics, and incorrect choices often correspond with intuitive ideas held by many students when they begin a physics course [1, 3–5]. Using dichotomous scoring methods eliminates the possibility to extract information about what students may be thinking when they choose an incorrect answer, and it implicitly assumes that all incorrect answers are equally wrong. We challenge this assumption by seeking to determine a ranking of incorrect responses (from more to less correct) for each question on a common RBAI for introductory mechanics, the Force and Motion Conceptual Evaluation (FMCE) [3].

A unique ranking of incorrect responses would allow us to measure growth in understanding if a student chooses different incorrect responses before and after instruction. In order to rank incorrect responses from better to worse, we must define what makes one incorrect response better than another. In the following sections we propose two different assumptions that allow us to define better incorrect responses, and present the results from independent analyses associated with each.

Our data come from more than 7,000 students' matched pre-/posttest responses to the FMCE. Students come from a variety of institutions and instructional settings. To illustrate the process of identifying a ranking of incorrect responses and combining two sets of results from independent analyses, we focus on question 18 on the FMCE. Questions 14–21 involve a toy car moving horizontally under the influence of an arbitrary force. Question 18 asks students to select a graph of force vs. time that would correspond with the statement "The car moves to the right and is slowing down at a steady rate (constant acceleration)" (Fig. 1). In the following sections we present details of our two analysis methods; we focus on similarities and differences between the ranking produced by each method for question 18 of the FMCE.[1] The same process may be used to determine a ranking of incorrect responses for every question on the FMCE, or any other RBAI.

II. RANKING RESPONSES USING ITEM RESPONSE THEORY

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 consistent with previous work using item response curves to rank incorrect responses [6–8]. In this study we use more sophisticated ranking methods based on Item Response Theory (IRT). IRT simultaneously estimates each student's overall understanding of the material (the latent trait or person parameter, θ) and determines the probability that s/he will be correct on each question given θ [9]. 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) model, the probability of a student answering a question correctly is given by,

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

[1] We omit response *J* (none of the above) from our analyses because it does not have a unique interpretation.

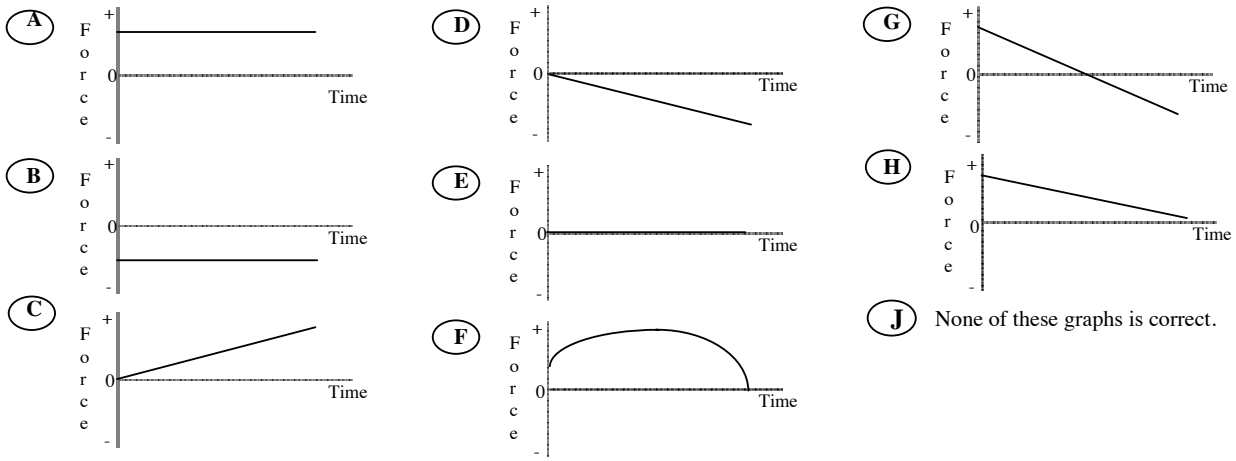

FIG. 1. Answer choices for Q18 on the FMCE. The correct answer for a car moving "to the right" and "slowing down at a steady rate" is *B*.

where a is the discrimination parameter and b is the difficulty parameter. The interpretation of these parameters may best be understood by examining a plot of $P(\theta)$ vs. θ. Consider the B curve in Fig. 2 showing the probability of being correct on question 18. The difficulty b provides the location of the midpoint of the curve: $P(b) = 0.5$, $b \approx 0.75$ in Fig. 2. If the curve is shifted to the right ($b > 0$), the question is considered more difficult because only students with higher values of θ are likely to be correct; if the curve is shifted to the left ($b < 0$), the question is considered to be less difficult. The discrimination a is related to the slope of the curve at the midpoint: the steeper the slope, the easier it is to differentiate between students' latent traits; whereas, the flatter the slope, the more difficult it is to differentiate between students' latent traits. Some previous work has used the three-parameter logistic model to analyze RBAI data [10], 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 [5, 11]. IRT analyses were performed using the mirt package in the R programming language [12, 13].

We use the nested logit model developed by Suh and Bolt to determine a ranking for the incorrect responses [14]. In this model, the probability of a student choosing a specific incorrect response, k, is given by

$$P_k(\theta) = \left(1 - \frac{1}{1+e^{-a(\theta-b)}}\right) \frac{e^{a_k(\theta-b_k)}}{\sum_i e^{a_i(\theta-b_i)}} \qquad (2)$$

where the parenthetical term is the probability of being not correct from the 2PL model, and the second term is Bock's nominal response model (NRM) in which the summation in the denominator is performed over all incorrect responses [15]. 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 [16]. Table I shows the values of a_k for several of the responses to question 18. Graphically, the value of a_k is related to the value of θ at which students are most likely to choose response k: for example, in Fig. 2 students are most likely to choose response A if they have a latent trait value in the range of about $0 < \theta < 1$ ($a_A = 0$), compared to students choosing response F if they have a latent parameter of $\theta < 0$, with lower values of θ being more and more likely to choose F ($a_F = -2.11$).

Table I shows the ranking determined by the 2PL-NRM analyses. We defined a binning threshold such that responses

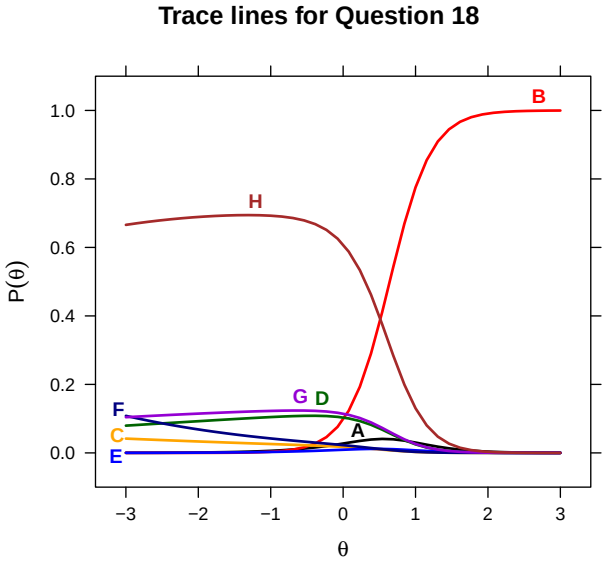

FIG. 2. Question 18 2PL-NRM response plots. The correct answer is B; the most common incorrect answer is H, which is consistent with the intuitive response [5]. $N = 14,012$ (mixed pre- and post-instruction responses)

TABLE I. Item Response Theory Ranking for Question 18

B	>	A	>	D	=	G	=	H	>	C	>	F
a_k:		0	>	−1.43	=	−1.47	=	−1.52	>	−1.87	>	−2.11

with a_k values within a range 0.1 of each other were considered to be equally correct (this corresponds to roughly 5% of the range of all a_k values). The correct answer, B, does not have a corresponding a_k value, but it is automatically considered the best response. We only include responses in our results given by at least 1% of the population.

III. RANKING RESPONSES USING THE MCNEMAR-BOWKER CHI-SQUARE TEST

Assumption: Students are more likely to choose better responses after instruction than before instruction.

This assumption is justified by the fact that students' scores on the FMCE are generally higher after instruction [11]. The McNemar-Bowker (MB) chi-square test for asymmetry aligns with this assumption by using students' matched pre-/posttest responses to identify asymmetric transitions between response pairs [17, 18]. If a student gives answer choice A on the pretest and answer choice B on the posttest, the transition is $A \to B$. We use a threshold for statistical significance of $p \leq 0.05$ (using the False Discovery Rate (FDR) correction). MB analyses were performed using the rcompanion package in the R programming language [19].

The values in Table II represent the number of students who underwent a particular pre/post transition. Values that are bold represent the dominant direction of transitions that were statistically significant. For example, 26 students transitioned $A \to H$, while 92 students transitioned $H \to A$; therefore, A is a better response than H. Values that are gray represent statistically insignificant transitions that may be ignored for our analyses. The italicized values represent statistically insignificant transitions that are relevant to our analyses. These transitions resulted in an adjusted p-value of 1.0, meaning that the transition has the same statistical significance, regardless of the direction. For example, 227 students made the transition $H \to D$ and 236 students made the transition $D \to H$. Table III(a) shows all significant transitions as well as the percentage of the population who made each transition (in either direction). Table III(b) shows the percentage of the population involved in each of several transitions between equivalent responses.

TABLE II. Question 18 MB transition table. Bolded values indicate the larger transition value (i.e., better on the posttest than the pretest). Gray values are not statistically significant and may be ignored. Italicized values indicate transitions that are not statistically significant, but are important for our interpretations. $N = 6{,}757$ (matched pre/post responses)

		Posttest							
	Q18	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Pretest	**A**	4	28	2	9	0	2	4	26
	B	14	717	6	33	5	4	12	87
	C	7	**37**	6	10	1	4	12	46
	D	16	**208**	13	87	10	14	*63*	*236*
	E	1	11	1	4	2	1	0	8
	F	10	**45**	3	16	2	16	22	72
	G	**22**	**250**	13	*59*	7	21	93	*281*
	H	**92**	**1420**	52	*227*	**26**	72	*281*	1904

TABLE III. Question 18 MB results: (a) statistically significant and (b) notable but statistically insignificant transitions. The pre-/posttest response comparison shows which transition is more favorable for that particular answer pair (i.e., $B > H$ means more students transitioned $H \to B$ than $B \to H$).

(a) Statistically significant transitions

Response Comparison	Adjusted p-value	Percent of Population
$B > H$	< 0.001	22.3%
$B > G$	< 0.001	3.9%
$B > D$	< 0.001	3.6%
$A > H$	< 0.001	1.7%
$B > F$	< 0.001	0.7%
$B > C$	< 0.001	0.6%
$E > H$	0.01	0.5%
$A > G$	0.002	0.4%

(b) Notable but statistically insignificant transitions

Response Comparison	Percent of Population
$G = H$	8.3%
$D = H$	6.9%
$D = G$	1.8%

IV. PUTTING IT ALL TOGETHER

Our goal is to define a unified ranking of incorrect responses for question 18 to be able to more accurately measure student understanding by recognizing productive elements in incorrect responses. Comparing the results from Tables I and III one may see both similarities and differences. Response B is correct for this question, and Table III shows that all other answer choices transition to it. The next best answer from IRT results is A. This is supported by several aspects of the MB results:

- A and B are statistically indistinguishable; and
- A is significantly better than both G and H.

It is interesting to note that the only two responses in Table II that are indistinguishable from the correct response are also the only other responses that appear on the greater-than side in Table III: A and E. These are also the only incorrect responses that are chosen by more students after instruction than before (the sum of the column in Table II is greater than

the sum of the row). Additionally, these are the only incorrect responses relating to graphs in which force is constant over time. These results may indicate that both of these responses are better than any of the other incorrect responses; however, we omit E from our IRT analyses because fewer than 1% of the student population selects E either before or after instruction, indicating that the result for E may not be robust. This is consistent with Fig. 2, which shows response E as a flat line with near zero probability and a very slight bump at $\theta \approx 0.75$ (about the same as A).

The results in Table III(b) are also consistent with our IRT results in Table I: D, G, and H are considered equally incorrect. The transitions between these answer choices have no dominant direction and they can be interpreted as being similar responses. Looking at the answer choices in Fig. 1, we notice that D, G, and H are all force vs. time graphs with constant, negative slopes. Table II shows that H is much more popular that D or G; it is, in fact, the most common incorrect response [5, 11]. Figure 2 also shows that all three of these responses have similarly shaped probability distributions with H being highest. These analyses suggest that choosing one of these answers over another may not have anything to do with a student's overall understanding of physics. This makes it interesting that nearly 17% of the population switched from one of these responses to another (Table III(b)).

Table I shows responses C and F as being worse than any of the others, but they only show up in Table III(a) as being worse than response B; essentially this only indicates that they are incorrect. Looking at the raw numbers in Table II, we can also see that responses C and F are fairly rare, each only chosen by about 2% of the population. As such, the relative correctness of responses C and F may not be robustly determined by our analyses; however, we may be confident in the ranking of a subset of responses for question 18:

$$\boxed{B > A > D = G = H.} \tag{3}$$

V. SUMMARY AND FUTURE DIRECTIONS

We have shown that two different analyses of students' responses to a question on the FMCE, based on independent assumptions of what makes one response better than another, can yield a consistent ranking of incorrect responses. We have also shown that some of the rankings among less common answers show up differently in the two analyses, but no results from one analysis contradict the results from the other.

We have several future plans to further clarify our results. We will expand our analyses by considering a third definition of what makes one answer better than another: students who choose better responses on the pretest are more likely to be correct on the posttest than students who choose worse responses; this is consistent with Thornton's work on conceptual dynamics [20]. We will also explore data selection techniques by removing rarely chosen responses (such as E on question 18) to try to achieve greater consistency between ranking results, as well as methods to adapt the MB analyses to account for the fact that responses are often concentrated around a single dominant response. Additionally, we plan to conduct interviews with students who select less common (e.g. A or F) or equivalent responses (e.g. D, G, or H) to gain insight into their thought processes while choosing a response. Our ultimate goal is to produce a unique robust ranking of incorrect responses for every question on the FMCE.

ACKNOWLEDGMENTS

We thank Sam McKagan and Ellie Sayre for providing access to data from PhysPort's Data Explorer. We also thank Kerry Gray, Nicholas Wright, Ian Griffin, and Ryan Moyer for their previous contributions as members of the research team. This project was supported by the National Science Foundation through a PhysTEC comprehensive site grant.

[1] A. Madsen, S. B. McKagan, and E. C. Sayre, Am. J. Phys. **85**, 245 (2017).
[2] J. M. Nissen, R. M. Talbot, A. N. Thompson, and B. Van Dusen, Phys. Rev. Phys. Educ. Res. **14**, 010115 (2018).
[3] R. K. Thornton and D. R. Sokoloff, Am. J. Phys. **66**, 338 (1998).
[4] D. Hestenes, M. Wells, and G. Swackhamer, Phys. Teach. **30**, 141 (1992).
[5] T. I. Smith and M. C. Wittmann, Phys. Rev. ST Phys. Educ. Res. **4**, 020101 (2008).
[6] T. I. Smith, K. A. Gray, K. J. Louis, B. J. Ricci, and N. J. Wright, 2017 PERC Proc. , 380 (2017).
[7] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, Am. J. Phys. **80**, 825 (2012).
[8] P. J. Walter and G. Morris, 2016 PERC Proc. , 376 (2016).
[9] F. B. Baker, *The Basics of Item Response Theory* (ERIC Clearinghouse on Assessment and Evaluation, 2001).
[10] J. Wang and L. Bao, Am. J. Phys. **78**, 1064 (2010).
[11] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Phys. Rev. ST Phys. Educ. Res. **5**, 010105 (2009).
[12] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2018).
[13] R. P. Chalmers, Journal of Statistical Software **48**, 1 (2012).
[14] Y. Suh and D. M. Bolt, Psychometrika **75**, 454 (2010).
[15] R. D. Bock, Psychometrika **37**, 29 (1972).
[16] R. D. Bock and I. Moustaki, in *Handbook of Statistics*, Vol. 26, edited by C. Rao and S. Sinharay (Elsevier, 2007) pp. 469–514.
[17] Q. McNemar, Psychometrika **12**, 153 (1947).
[18] A. H. Bowker, Journal of the American Statistical Association **48**, 572 (1948).
[19] S. Mangiafico, *rcompanion: Functions to Support Extension Education Program Evaluation* (2018), R package version 2.0.0.
[20] R. K. Thornton, AIP Conf. Proc. **399**, 241 (1997).