Comparing pre/post item response curves to identify changes in misconceptions

Paul J. Walter
Mathematics Department, St. Edward's University, 3001 S. Congress Ave., Austin, TX, 78704, USA

Trevor I. Smith
Department of Physics and Astronomy and Department of STEAM Education, Rowan University, Glassboro, NJ 08028, USA

We use quantitative measures and visual inspection to compare the item response curves (IRCs) of matched pre-/post-instruction Force Concept Inventory (FCI) data. We find that the IRCs are not static; the pre-instruction IRCs differ from the post-instruction IRCs by more than can be explained by random chance. This result is also the case for a subpopulation consisting of students who make little or no gains on the FCI, suggesting that learning is taking place even when scores do not change appreciably. We consider three items where students make substantial progress (item 4) or little progress (items 14 and 21) compared to overall changes in FCI scores.

2021 PERC Proceedings edited by Bennett, Frank, and Vieyra; Peer-reviewed, doi.org/10.1119/perc.2021.pr.Walter

I. INTRODUCTION

There are various ways to ascertain changes in understanding that result from instruction. Common practice includes administering a research-based multiple-choice assessment before and after instruction, determining an overall score for each student, and reporting a measure of growth (often normalized gain or effect size) [1–3]. An important but often implicit assumption in these common practices is that any learning will appear as an increase in the number of items answered correctly. In recent years, many researchers have explored methods for reporting student growth using methods that are more informative than comparisons of single-number scores generated from dichotomously scored items. Morris *et al*. analyze pre-/post-instruction concept inventory data to incorporate all of the answer choices on an item to construct transition matrices to observe changes in student understanding [4]. In another example, Pérez-Lemonche *et al*. apply multidimensional item response theory to study the effect of answer choices that highlight particular student misconceptions and act as distractors [5]. They also characterize item response curves (IRCs) based on their shape, such as attractive distractors having a maximum at an intermediate score. A strength of item response theory and the related IRCs is that they highlight the relationship between a student's overall understanding of a topic and their likelihood of choosing each answer on a given item [6].

Similar to Morris *et al*. and Pérez-Lemonche *et al*., we analyze data from the 30-item Force Concept Inventory (FCI) [7]. The novelty of our approach is that we focus on comparing students' IRCs before instruction to those same students after instruction. We do so by showing the IRCs graphically and employing an approach that allows for having a quantifiable measure of how close the pre-instruction IRCs are to the post-instruction IRCs. In this work, we investigate whether IRCs (and the relationships between total score and the probability of choosing each answer choice) are static and if learning appears as a simple increase in the overall score.

We seek to answer the following research questions.

1. Do the pre-instruction and post-instruction IRCs of matched students differ? If so, how pronounced are and what features characterize those differences?
2. Are the pre-/post-instruction IRCs the same for those students who make little to no gains?

II. BACKGROUND

Morris *et al*. first introduced item response curves (IRCs) for each item on the FCI using data from more than 6,000 respondents [6, 8]. As an example, Fig. 1 shows the item response curves for item 4 on the FCI for the pre-instruction and post-instruction results of 9,354 matched students. IRCs plot the percentage of students who selected an answer choice as a function of the overall score. In Fig. 1, the correct answer choice is E (shown in red); students with a perfect score

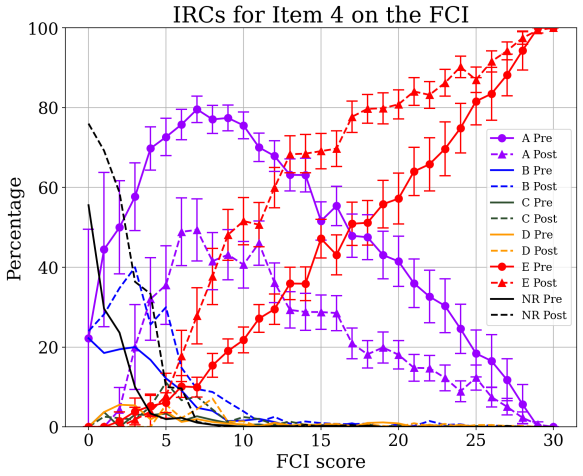

FIG. 1. Pre-instruction and post-instruction item response curves for item 4 on the FCI for the data set of 9,354 students. NR = no response. The error bars are the 95% confidence intervals of 10,000 bootstrapped sample data sets.

(i.e., an FCI score of 30) selected that answer choice. Answer choice A (shown in purple) is an example of a distractor, which attracts students with a particular misconception (i.e., the heavier object in a collision exerts a larger force).

IRCs are quite similar to item characteristic curves created using item response theory (IRT) analyses. In dichotomous IRT models, item characteristic curves show the probability of getting an item correct as a function of the latent characteristic of ability level [9]. In contrast, IRCs substitute the overall score in place of IRT's ability level. IRCs contain more information than IRT's item characteristic curves by showing the percentage of students who select each answer choice, not only whether they selected the correct answer. IRT nominal response models also provide information about the probability of students choosing each response option, but these analyses typically require very large data sets, which limits their utility in many cases [10, 11].

Ishimoto, Davenport, and Wittmann (hereafter, IDW) used IRC analysis to compare Japanese and American students' pre-instruction performance on the Force and Motion Conceptual Evaluation (FMCE) [12, 13]. They found from visual inspection that the IRCs for the two populations were highly similar for most items. They attributed differences to contextual differences resulting from the translation from English to Japanese, and to cultural differences between Japanese and American students; e.g., American students typically have more experience driving and riding in automobiles.

Walter, Nuhfer, and Suarez (hereafter, WNS) introduced a metric for quantifiably comparing the IRCs of two populations [14]. WNS used more than 12,000 students' responses to the 25-item Science Literacy Concept Inventory (SLCI), which assesses respondents' understanding of citizen-level science literacy, to compare the IRCs of different demographic populations [14, 15].

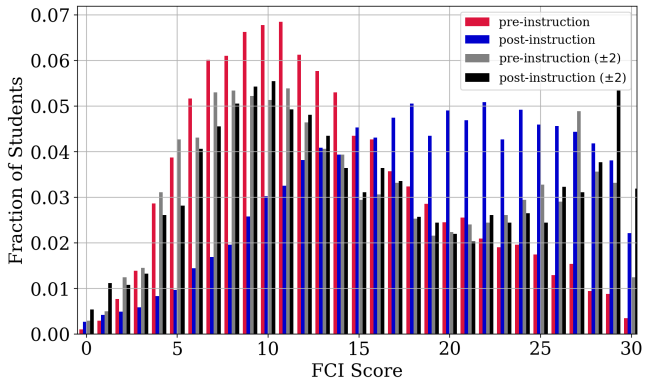

FIG. 2. The fraction of students with each score on the 30-item FCI for 9,354 matched pre-/post-instruction respondents, as well as for 2,414 matched students who scored within ±2 on the pre-instruction and post-instruction FCIs. The data was provided by PhysPort.

Richardson, Smith, and Walter (hereafter, RSW) replicated IDW's work using a more extensive data set of American students while expanding upon the same quantitative approach employed by WNS to compare IRCs [11, 12, 14].

III. DATA

PhysPort provided our data set consisting of the matched pre-/post-instruction FCI results of 9,636 students [16]. Students who left 10 or more items unanswered on either the pre-instruction or post-instruction FCI were not included in the original data set. Students who selected the same answer choice for all 30 items on either the pre-instruction or post-instruction FCI were removed from the sample. In almost all such cases (99 of 101), students repeatedly chose answer choice A. Students who scored lower on the post-instruction FCI than on the pre-instruction FCI by 6 or more out of 30 were removed from the sample, which removed an additional 181 students [17]. This left a sample of 9,354 students with matched pre-/post-instruction FCI data. The data set does not include demographic information or the mode of instruction for most participants, and such information is not considered in this work. Figure 2 shows the fraction of students with each overall score on the pre-/post-instruction FCI.

IV. METHODS

In addition to comparing pre-instruction and post-instruction IRCs by visual inspection, we use the quantitative approach of calculating IRC dot products to compare two populations, as described in WNS [14]. In this approach, each population is represented by 31 vectors (one for each score bin), each with five dimensions (one for each response option). The components of each vector are the response frequencies for that score bin (i.e., the IRC data points). The vectors are normalized to have a magnitude of unity, and the IRC dot product between two populations is the weighted average of the 31 individual dot products. Given that all values of each IRC are positive (or zero), the theoretical limits of the IRC dot product range from 0 (completely different IRCs—not realized in practice) to 1 (identical IRCs) for each item.

To interpret the value of the IRC dot products in meaningful ways, we incorporate two forms of uncertainty. WNS ascertain whether an IRC dot product value for an item is potentially the result of random chance by pooling all data together and repeatedly assigning each response set to one of the comparison groups at random. Following the convention of RSW, we will refer to the central 95% range of randomized trials described in WNS as the randomized trial confidence interval (RTCI) [11]. The RTCI provides an expected range of IRC dot product values between two identical populations. As was introduced in RSW, we also include an IRC dot product confidence interval (DPCI) using the central 95% distribution of 10,000 bootstrapped simulations. The DPCI is based on the inherent uncertainty in the data points of the IRCs themselves. Large gaps between the RTCI and DPCI suggest that the low value of the IRC dot product represents meaningful differences between the IRCs. We compute the dot product effect size (DES) as described in RSW to quantify this difference [11].

To help answer Research Question 2, we created a subset of the 9,354 students in our sample whose pre-instruction and post-instruction scores differed by no more than 2 out of 30. This group consisted of 2,414 students (26% of our data set); their score distribution (labeled as ±2) is shown in Fig. 2.

V. RESULTS COMPARING PRE/POST IRCS

Figure 3 shows the IRC dot products comparing the pre-instruction and post-instruction IRCs (data points shown in red) for each item. The light red error bars are the RTCIs, and the dark red error bars are the IRC DPCIs. The same is shown in blue for the subset of students whose pre-instruction scores were within ±2 of their post-instruction scores. An interesting feature, which is the subject of future work, is that the IRC dot products of the ±2 population are above the IRC dot product confidence intervals for many items [18]. Some potential factors include the sample size, the bimodal score distribution (Fig. 2), and ceiling effects related to the IRC dot products being close to 1.

Pérez-Lemonche *et al*. identify 12 incorrect answer choices acting as effective distractors and having a maximum at intermediate scores: 4A, 5D, 5E, 11C, 13C, 15C, 17A, 18D, 18E, 25D, 28D, 30E [5]. While item 4 has the lowest IRC dot product, the next two lowest values are for items 5 and 18, which are the only items that Pérez-Lemonche *et al*. identify as having two such intermediate distractors.

For the overall population, the IRC dot product for every item is below its corresponding RTCI. Item 4 has the lowest IRC dot product value of 0.88, indicating that the item has the

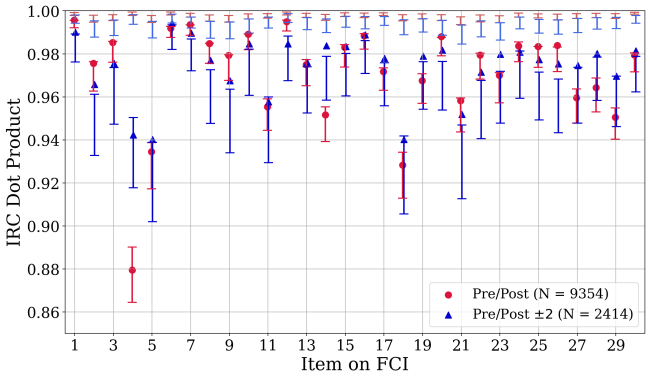

FIG. 3. The IRC dot product of the pre-instruction and post-instruction FCI IRCs of 9,354 matched students is shown in red for each item. The light red error bars represent the randomized trial confidence intervals for the 9,354 students. The dark red error bars represent the IRC dot product confidence intervals. The data points and error bars in blue are the same for the subset of students who score within ±2 on the pre-instruction and post-instruction FCIs.

item	overall	±2	item	overall	±2
1	0.7	0.4	16	1.3	0.3
2	1.5	0.9	17	2.5	0.9
3	1.1	0.6	18	3.2	1.4
4	4.6	1.6	19	2.1	0.6
5	2.8	1.4	20	1.0	0.4
6	1.3	0.2	21	2.3	1.0
7	0.8	0.3	22	1.4	0.7
8	1.2	0.5	23	1.8	0.4
9	1.4	0.8	24	1.5	0.6
10	1.1	0.4	25	1.4	0.7
11	2.9	1.2	26	1.1	0.7
12	0.7	0.6	27	2.4	0.7
13	1.9	0.8	28	2.1	0.7
14	2.8	0.4	29	3.2	1.0
15	1.4	0.6	30	2.2	0.9

TABLE I. The dot product effect size (DES) values for pre-/post-instruction comparisons for the overall data set (N = 9,354) and for the subset of students whose scores on the pre-instruction and post-instruction FCIs differ by no more than ±2 (N = 2,414).

most pronounced differences in its pre- and post-instruction IRCs. The IRC dot product values in Fig. 3 and the DES values shown in Table I allow us to compare the amount of difference in pre/post-IRCs for the overall population and the ±2 subpopulation for each item. RSW define a threshold for similar IRCs as having a DES value of 0.6 or less [11]. Table I shows that none of the comparisons of the overall data set are within this range, and only 13 items are similar for the ±2 data set. This suggests that, even though the IRC dot products seem high (only item 4 is below 0.92), they are not as high would be expected if the differences between the IRCs were the result of random fluctuations in student responses. To further explore these differences we present a careful analysis of three items: one that shows significant learning, and two that show minimal learning.

A. Example of Significant Gains in Understanding

Based on both the IRC dot product values in Fig. 3 and the DES values in Table I, the IRCs' changes from pre-instruction to post-instruction are more pronounced on item 4 than for any item on the FCI. Item 4 asks students to compare the forces a car and a truck exert on each other during a collision. Figure 1 (on p. 1) shows the pre- and post-instruction IRCs for item 4 on the FCI. There are considerable differences in the pre-/post-instruction IRCs for the correct answer choice, E (shown in red), and the distractor, answer choice A (shown in purple). The error bars on these IRCs were generated from the bootstrapped samples used to generate the DPCIs. The leftward shift of the choice E IRC shows that lower-scoring students are more likely to choose the correct answer to item 4 post-instruction than pre-instruction. The downward shift of the choice A IRC shows that lower-scoring students are less

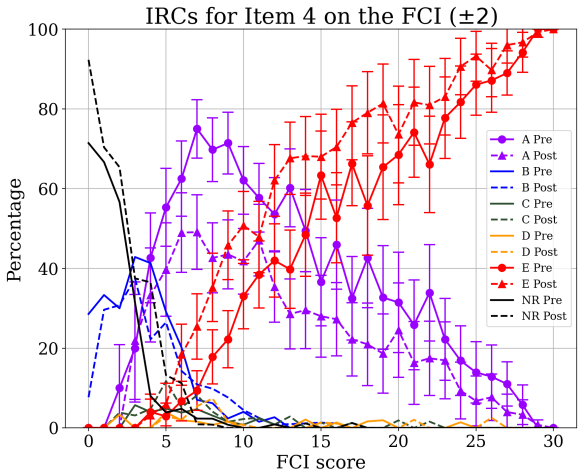

FIG. 4. Pre-instruction and post-instruction item response curves for item 4 of the FCI for the 2,414 students who scored within ±2 on the pre-instruction and post-instruction FCIs. NR = no response.

likely to choose the most common distractor post-instruction. On the far left of the plot, we also see an inversion of choices A and B, with B being more likely post-instruction for the lowest-scoring students. Together these results suggest that students are more likely to learn the content tested by item 4 than that of other items on the FCI. The pronounced changes to the IRCs in Fig. 1 reflect that students substantially improve their understanding of objects in a collision exerting equal and opposite forces on each other.

Fig. 4 shows the IRCs for item 4 of the 2,414 students who scored within ±2 on the pre-instruction and post-instruction

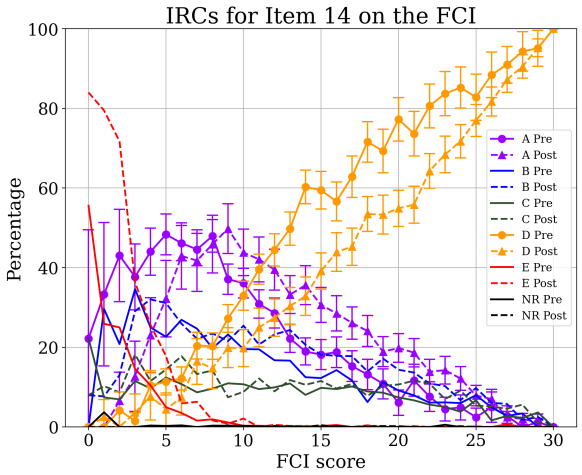

FIG. 5. Pre-instruction and post-instruction item response curves for item 14 for the data set of 9,354 students. NR = no response.

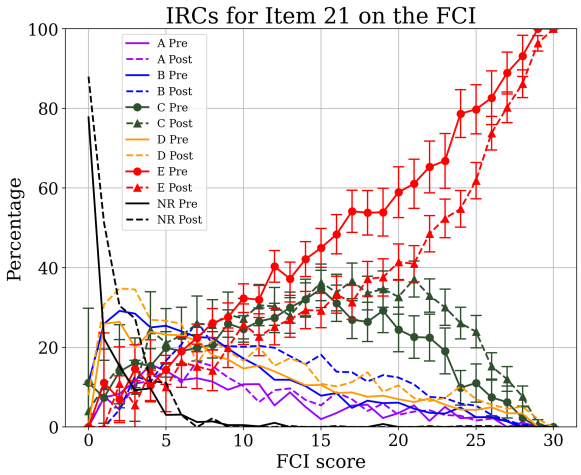

FIG. 6. Pre-instruction and post-instruction item response curves for item 21 for the data set of 9,354 students. NR = no response.

FCIs. While the differences in the IRCs are less pronounced than for Fig. 1, even students who made little or no overall gains made progress in understanding Newton's third law.

B. Examples of Little Gains in Understanding

Comparing pre-instruction and post-instruction IRCs allow for identifying cases where students do not learn particular concepts even while making gains on the concept inventory overall. Figure 5 shows the IRCs for item 14 on the FCI, which asks students to select the path a bowling ball would take if it fell from an airplane. For a given overall score, students select the correct answer (D) at higher rates on the pre-test than the post-test. The IRC for the distractor, choice A, shifts to the right from pre-test to post-test. The changes in the IRCs indicate that, while students are making gains on the FCI overall, they are not making commensurate progress on the concept involved in this item.

Figure 6 shows the IRCs for item 21 on the FCI, which asks students to select the best path for a rocket that has a constant speed in one direction (to the right) while accelerating in another (upward). As was the case for the IRCs on item 14, the IRCs on item 21 show how for a given overall score, students are more likely to get the item correct (choice E) on the pre-test. We also see how the distractor (choice C), which has the rocket moving in a straight path in a direction up and to the right, shifts to the right from pre-test to post-test. Items 14 and 21 involve the same concept of selecting the best path to describe an object having a constant speed in one direction while accelerating in another. The rightward shifts (for both the correct and incorrect answer choices) suggest that students may be increasing their scores overall but not changing the ways that they interact with these items. We observe the same effect on items 19 and 27 (IRCs not shown).

VI. CONCLUSION

Quantitative comparisons of pre-instruction and post-instruction IRCs show that the relationship between students' overall scores and their likelihood to choose a particular answer choice are not static. As students' scores increase, they do not progress evenly across all items. Some items show larger-than-average gains (e.g., item 4), while others show smaller-than-average gains (e.g., items 14 and 21).

Using IRCs to make these comparisons allows us to see changes to incorrect answer choices as well as the correct choices. For item 4, we see the likelihood of students choosing incorrect response A decreasing dramatically, even for students whose total scores do not change much (Fig. 4). Conversely, for items 14 and 21, students choose the same incorrect responses at the same rates, even when their scores increase. Instructors can use this information to identify topics that may require additional (or different) instruction.

The large size of our data set, and the fact that the data come from many different institutions across the country, suggests that these trends may exist across different forms of instruction. Since we are using the same population in this work and comparing the pre-instruction IRCs to the post-instruction IRCs, bias present in the items (e.g., gender bias) is likely to remain consistent for the pre-instruction and post-instruction data. Any changes in the IRCs are thus a result of the changes in the understanding of the population. Future work will involve looking for additional similarities in the content of the items with IRCs that shift in similar ways.

ACKNOWLEDGMENTS

We thank Eleanor Sayre and PhysPort for providing data used in this work. Supported by NSF award DUE-1836470.

[1] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource Letter RBAI-1: Research-Based Assessment Instruments in Physics and Astronomy, American Journal of Physics **85**, 245 (2017).

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

[3] J. M. Nissen, R. M. Talbot, A. N. Thompson, and B. Van Dusen, Comparison of normalized gain and Cohen's d for analyzing gains on concept inventories, Phys. Rev. Phys. Educ. Res. **14**, 10115 (2018).

[4] G. A. Morris, P. J. Walter, S. Skees, and S. Schwartz, Transition matrices: A tool to assess student learning and improve instruction, The Physics Teacher **55**, 166 (2017).

[5] Á. Pérez-Lemonche, J. Stewart, B. Drury, R. Henderson, A. Shvonski, and D. E. Pritchard, Mining students pre-instruction beliefs for improved learning, in *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale* (2019) pp. 1–10.

[6] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley, Testing the test: Item response curves and test quality, American Journal of Physics **74**, 449 (2006).

[7] D. Hestenes, M. Wells, G. Swackhamer, *et al.*, Force concept inventory, The Physics Teacher **30**, 141 (1992).

[8] 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, American Journal of Physics **80**, 825 (2012).

[9] S. E. Embretson and S. P. Reise, *Item response theory* (Psychology Press, 2013).

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

[11] C. J. Richardson, T. I. Smith, and P. J. Walter, Replicating analyses of item response curves using data from the force and motion conceptual evaluation (2021), arXiv:2104.08552 [physics.ed-ph].

[12] M. Ishimoto, G. Davenport, and M. C. Wittmann, Use of item response curves of the force and motion conceptual evaluation to compare japanese and american students' views on force and motion, Physical Review Physics Education Research **13**, 020135 (2017).

[13] R. K. Thornton and D. R. Sokoloff, Assessing student learning of newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, American Journal of Physics **66**, 338 (1998).

[14] P. J. Walter, E. Nuhfer, and C. Suarez, Probing for bias: Comparing populations using item response curves, Numeracy **14**, 2 (2021).

[15] E. B. Nuhfer, C. B. Cogan, C. Kloock, G. G. Wood, A. Goodman, N. Z. Delgado, and C. W. Wheeler, Using a concept inventory to assess the reasoning component of citizen-level science literacy: Results from a 17,000-student study, Journal of microbiology & biology education **17**, 143 (2016).

[16] S. B. McKagan, Physport, http://www.physport.org (2011).

[17] We assume that students whose scores decrease by 6 out of 30 or more after instruction are likely to have not taken the post-instruction FCI seriously.

[18] While in such cases the IRC dot products do not fall within the IRC dot product confidence intervals of the ± 2 population are the central 95% distribution of 10,000 bootstrapped samples, they do fall within the range of 100% of the distribution.