Replicating analyses of item response curves using data from the Force and Motion Conceptual Evaluation

Connor J. Richardson,[1] Trevor I. Smith,[1,2,*] and Paul J. Walter[3]

[1]Department of Physics and Astronomy, Rowan University, Glassboro, New Jersey 08028, USA
[2]Department of STEAM Education, Rowan University, Glassboro, New Jersey 08028, USA
[3]Department of Mathematics, St. Edward's University, Austin, Texas 78704, USA

(Received 19 April 2021; accepted 31 August 2021; published 1 October 2021)

Ishimoto, Davenport, and Wittmann have previously reported analyses of data from student responses to the Force and Motion Conceptual Evaluation (FMCE), in which they used item response curves (IRCs) to make claims about American and Japanese students' relative likelihood to choose certain incorrect responses to some questions. We have used an independent dataset of over 6,500 American students' responses to the FMCE to generate IRCs to test their claims. Converting the IRCs to vectors, we used dot product analysis to compare each response item quantitatively. For most questions, our analyses are consistent with Ishimoto, Davenport, and Wittmann, with some results suggesting more minor differences between American and Japanese students than previously reported. We also highlight the pedagogical advantages of using IRCs to determine the differences in response patterns for different populations to better understand student thinking prior to instruction.

DOI: 10.1103/PhysRevPhysEducRes.17.020127

I. INTRODUCTION

Research-based multiple-choice assessment instruments, such as the Force Concept Inventory (FCI) [1] and Force and Motion Conceptual Evaluation (FMCE) [2], are ubiquitous in physics education research. These tools have been used as standard measures for research, instruction, and programmatic assessment purposes for the past three decades [3,4]. Conceptual assessments have been beneficial for illustrating the need for and benefit of research-based instructional materials and strategies [5–8].

Although common practices for collecting and analyzing data from the FCI and FMCE were established over 20 years, significant research persists on how to use these assessments and interpret the results [9–30]. Ishimoto, Davenport, and Wittmann compared preinstruction FMCE student responses from a population of American introductory physics students to preinstruction responses from a population of Japanese introductory physics students using both overall score distributions and item response curve (IRC) analyses [30]. IRCs show the fraction of students who selected each answer choice to a particular item as a function of their total score on the test [26–29]. These analyses can show if different groups of students (e.g., high scoring or low scoring) select different incorrect answer choices and help reveal trends evident in students' patterns of correct and incorrect answer choices.

Our goal is to replicate the work of Ishimoto, Davenport, and Wittmann from Ref. [30], which we will refer to as IDW. Our data are comprised of preinstruction responses to the FMCE from 6584 American introductory physics students from a variety of colleges and universities. We carefully compare the score distribution from our data to those published by IDW, and we use a recently developed method for comparing IRCs quantitatively between the three populations [31]. Our primary goal is to determine how well our results agree with their results from American students, which will indicate whether their results are more generally representative of American students. We also compare our results with their Japanese students' results to determine if we see the same discrepancies that they report between American and Japanese populations.

IDW found that the preinstruction score distributions for American and Japanese students are fairly similar in shape, but that the scores for Japanese students are statistically significantly higher than those of American students (11.22 ± 0.19 vs 9.05 ± 0.15, $t = 8.8$, $p < 0.005$). Using qualitative visual comparisons, IDW concluded that the patterns of correct and incorrect answer choices shown on IRCs did not differ greatly between the American and Japanese student populations for most FMCE items. They reported some isolated cases of discrepancies in the answer patterns shown in the IRCs, which they attribute to cultural differences between the two populations, e.g., children growing up in Japan do not have as much experience with

*smithtr@rowan.edu

flipping coins as children in the USA, and American students typically have more experience driving and riding in cars than Japanese students.

Based on the findings of IDW, and our overall goals for replicating their work, we answer the following research questions:

1. Does the distribution of preinstruction FMCE scores from our dataset match those from the IDW American dataset and support their claim that Japanese students have higher preinstruction scores than American students?
2. Do the IRCs for our dataset match those from the IDW American dataset, indicating that students chose the same responses at similar rates?
3. Are the IRCs for our dataset different from those from the IDW Japanese dataset in the same ways they previously reported for American and Japanese students?
4. What additional information about the similarities or differences between IRCs can we reveal by using rigorous quantitative comparisons?

Without access to independent data from Japanese students, we cannot fully replicate the IDW study. However, we feel confident that using an independent dataset of American students that is more than twice as large as in their original study allows us to make strong claims about the veracity of their results from American students, which will then, in turn, allow us to evaluate the strength of their overall claims.

Ishimoto, Thornton, and Sokoloff have previously provided evidence for the validity of the Japanese translation of the FMCE used by IDW [32]. None of the current study authors is knowledgeable about the Japanese language or an expert in Japanese culture. We do not attempt to evaluate the validity of the translation nor comment on claims made by IDW about cultural differences between Japanese and American students that may be the cause of their observed IRC differences. Our research questions focus solely on interpretations of our analyses of the three datasets, with an ultimate goal of determining whether or not the results from American physics students presented by IDW may be considered representative of American physics students in general.

II. ITEM RESPONSE CURVES

Morris *et al.* introduced IRCs as a simplified form of item response theory (IRT) that uses the total test score as an independent variable, rather than the IRT latent trait of ability level [26,27]. IRC analyses are similar to the IRT nominal response model but require far less computational power [22,23,33,34].

A. Reading an IRC

To illustrate the information shown in IRCs, Fig. 1 shows an example of students' response patterns for item 17 on the FMCE. Item 17 presents students with the description of

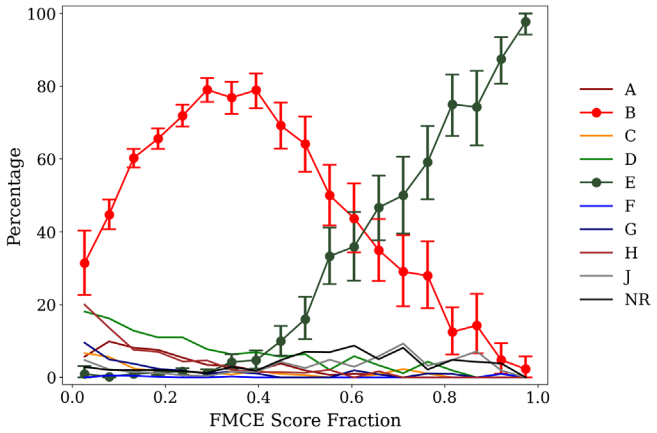

FIG. 1. Item response curves for 6584 American students (the RSW dataset) for item 17 on the FMCE. NR = no response. The error bars on E (the correct answer) and B (the most commonly chosen incorrect answer) are the 95% confidence intervals of 10 000 bootstrapped sample datasets based on the RSW data described in Sec. III C. For visual clarity, specific plot markers and error bars are omitted for less frequently chosen answers.

motion for a toy car—"the car moves toward the left (toward the origin) with a steady (constant) velocity"—and students must choose a graph of force vs time "which could allow the described motion of the car to continue" [2]. In Fig. 1 there is one curve for each answer choice, which shows the percentage of students selecting a particular answer choice as a function of respondents' fractional overall score. The line segments connecting the data points only serve as a visual aid. The correct answer choice is E (constant zero force, shown in dark green); the correct answer on any IRC plot may be seen as having low frequency with low-scoring students and high frequency with high-scoring students. Answer choice B (constant negative force, shown in red) is the most common incorrect answer choice (shown by higher frequencies than all other incorrect answers), and it is consistent with the common idea that the total force on an object is proportional to the object's velocity. Answer choice B also has an intermediate maximum, indicating that students are most likely to choose B when they have a moderate level of understanding (indicated by the fractional score) [35].

An important caveat when reading IRCs for preinstruction FMCE responses is that the score distribution is not uniform. As shown in Fig. 2, the mode of the distribution occurs at a fractional score of ~0.15–0.2, and a large majority of students earn scores < 0.4. In this way, IRCs cannot directly show students' overall likelihood of choosing each answer choice. Additionally, answer choice J is the equivalent of "none of the above" for all items on the FMCE; therefore, we do not consider students who choose J to be a coherent group. We know that those students think that none of the other answers is correct, but we cannot claim that they all agree on what a correct answer would be.

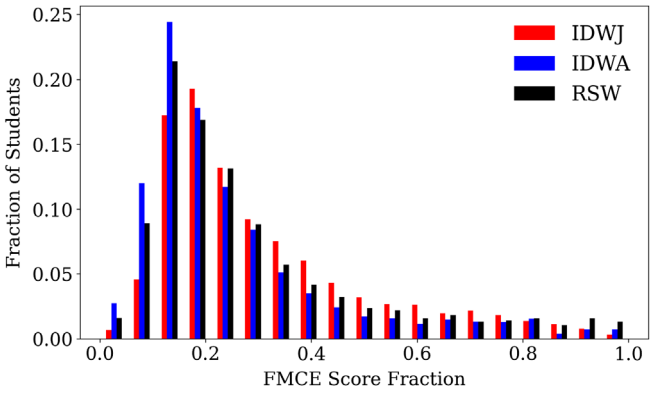

FIG. 2. The fractional score distributions on the FMCE are shown for three separate datasets: 1531 Japanese students (IDWJ) and 2347 American students (IDWA) used in Ishimoto, Davenport, and Wittmann [30]; 6584 American students (RSW) obtained from PhysPort and other sources [7,15,16,37].

B. IRC vs IRT

Item response theory (IRT) has a solid theoretical foundation that has been developed over the past several decades based on the notion that student responses to multiple-choice tests are dependent on a latent trait that cannot be measured directly (in our context, this would be an overall understanding of introductory mechanics) [33,34,38–40]. Additionally, IRT can be generalized to multiple dimensions that represent different latent traits (i.e., types of knowledge or understanding). Previous work has shown how multidimensional IRT analyses can identify items on the FMCE that map onto various facets of understanding (similar to factor analysis) [17]. Recently, a multidimensional analysis using the IRT nominal response model has identified two orthogonal dimensions of understanding based on students' specific responses (both correct and incorrect) to FCI items [21].

In contrast, item response curves have been in use for about 15 years, and they are no more complicated than reporting population-specific conditional probabilities of students choosing a particular response given their score on the test. Students' overall scores substitute for the single IRT latent trait of ability level [26]. As such, IRCs can be viewed as a simplified form of a one-dimensional IRT analysis with less statistical power [41].

Using a rigorous IRT-based analysis would provide more statistical power for any of our claims, and recent work has shown how the nominal response model can be applied to FMCE data to determine a ranking for all incorrect responses to each question [23]; however, there are benefits to using IRC-based analyses instead. IRT analyses require datasets at least 10 times as large as the number of estimated parameters to have robust statistical power [38]: a minimum of 3000 students for the FCI and 7220 for the FMCE. Additionally, performing IRT analyses requires fairly sophisticated computational methods. Software exists to perform these calculations (e.g., the mirt package for R [43,44]), but IRC analyses are much simpler and require only basic spreadsheet operations to perform. As such, IRC analyses are more accessible for researchers with smaller datasets and those who prefer more straightforward computational methods. We have chosen to use IRC analyses for this work because our goal is to replicate the IDW study [30] as closely as possible.

III. METHODS

Many of our methodological choices were taken directly from the IDW study [30]: cleaning the dataset based on rules for removing some incomplete response sets, calculating an overall FMCE score for each student based on the common 37-point scoring scheme, creating two-score bins for generating IRCs, and using statistical methods to compare overall score distributions and qualitative visual comparisons of IRCs. We augmented these methods by performing additional statistical analyses to compare score distributions and utilizing quantitative analyses to compare IRCs.

Our raw data consist of preinstruction item responses to the FMCE gathered from 7325 students at various colleges and universities in the United States known to be disjoint from IDW's data pool. One of our major sources is PhysPort's Data Explorer, a large online database of anonymous data uploaded by instructors [37]. We know little more than that these students were enrolled in an introductory physics course in the USA, and that they were given the FMCE before and after instruction. We do not know their majors, previous levels of physics taken, or demographic information; IDW report having some of this information [30]. Approximately 1000 students in our dataset come from four known data sources: two public state universities, a two-year college, and a selective private liberal arts college. Other analyses of some of these data sources have previously been reported [7,15,16]. Some information is known about the instructional methods used at these four institutions, but no demographic information is available about the students.

A. Cleaning and analyzing the dataset

We mirrored IDW's methods for cleaning the data. As such, we only included response sets that included answers to at least one item in each of the previously defined item clusters (force sled, reversing direction, force graphs, acceleration graphs, velocity graphs, Newton III, and energy [30,42]). We also omitted response sets that had more than 6 unanswered items in total. IDW mentions having the same student appear twice and choosing to use their first responses. We did not have a method to test if a student appeared twice in our data. After cleaning, we removed the responses of 741 students leaving 6584 to be analyzed. We refer to this as the RSW dataset.

To compare students, we generated a score in the same manner as IDW, who followed the recommended scoring

procedure of Thornton and Sokoloff [2,6]: items 5, 6, 15, 33, 35, 37, and 39 are not scored; items 8–10, 11–13, and 27–29 are each treated as sets in which students earn two points for answering all three items in a set correctly, and zero points if any item is answered incorrectly. Thus, we used the same 40 items to calculate a student's score, with a maximum score of 37.

We used a two-score binning method to create our IRCs, so students who scored a 0 or 1, 2, or 3, etc., were plotted together. Binning in this way reduces the standard error of the plots because it increases the sample size per point [30]. We calculated the fraction of students who chose each answer choice for each score bin and plotted these as IRCs. We created IRCs for all items on the FMCE even though the total score does not include each item individually.

B. Comparing datasets

1. Comparing score distributions

Ishimoto, Davenport, and Wittmann reported the score distributions for both of their datasets (as a bar graph in Fig. 4 of IDW [30]), as well as the population size, mean score, median score, and standard deviation for each (in Table II of IDW [30]). Using the image processing software ImageJ (version 1.53) [45], we analyzed the IDW score distribution bar graph to estimate the fraction of students in each dataset that earned each score on the FMCE. Combing this information with the reported descriptive statistics allowed us to determine the number of students that earned each score and recreate the score distributions from IDW [30].

We calculated the same descriptive statistics for our data that were reported in IDW, but due to the nonparametric nature of the score distributions, we chose comparison methods that deviate somewhat from theirs. We compared our score distribution to both of the IDW distributions using a Kruskal-Wallis rank-sum test. We also performed *post hoc* pairwise comparisons with the Wilcoxon rank-sum test using the Benjamini-Hochberg p-value correction method. Additionally, in recognizing that p values alone are not sufficient for determining whether or not a real effect exists [46,47], we chose to compute the effect sizes of the differences between the datasets to get a more complete picture of any differences. We calculate rank epsilon squared (ε^2) as a measure of the main effect between all three score distributions, and the rank-biserial correlation (r) as a pairwise effect size. Analyses used the R statistical software platform [43,48].

2. Treating IRCs as vectors and taking dot products

Walter, Nuhfer, and Suarez provide a means of quantitatively comparing the IRCs of two populations by treating the information in IRCs as multidimensional vectors and computing a dot product as a metric of similarity [31]. They compared the IRCs of groups with differing demographics (e.g., gender, ethnicity) using data collected from the 25-item Science Literacy Concept Inventory [49].

The metric for quantitatively comparing the IRCs of two populations is given by

$$\langle \mathbf{a} \bullet \mathbf{b} \rangle_k = \frac{\sum_j \sum_i n_{a,j} n_{b,j} \hat{a}_{ijk} \hat{b}_{ijk}}{\sum_j n_{a,j} n_{b,j}}, \tag{1}$$

where $\langle \mathbf{a} \bullet \mathbf{b} \rangle_k$ is the IRC dot product for populations a and b on item k. The components of $\hat{a}_{ijk}$ and $\hat{b}_{ijk}$ are the number of respondents who chose answer choice i who had an overall score in score bin j for item k, which are then normalized with respect to the number of answer choices. For example, for population a

$$\hat{a}_{ijk} = \frac{a_{ijk}}{\sqrt{(\sum_i a_{ijk} a_{ijk})}},$$

which enforces the condition that

$$\sum_i \hat{a}_{ijk} \hat{a}_{ijk} = 1 \quad \forall\, j, k.$$

Also, $n_{a,j}$ and $n_{b,j}$ are the number of respondents who were in score bin j for population a and b, respectively.

For each item in our work, a population has 19 associated IRC vectors, one for each score bin; each vector contains one component for each answer choice and is normalized over all answer choices for that item (5–9 for the FMCE). For example, a normalized IRC vector for population a is $\hat{a}_{ijk}$ for a particular score bin j and item k. The IRC dot product [Eq. (1)] for a particular item is the weighted (based on the number of respondents per score bin) average of dot products of the normalized IRC vectors of two populations. Since all IRC vector components are ≥ 0, the values of the IRC dot product range from 0 (the IRCs are completely different) to 1 (the IRCs are identical) [50].

To compare two populations using IRC dot products, one must know each data point's precise value for each IRC. One author of the IDW study was gracious enough to share their results with us to make these comparisons.

3. Randomized trial confidence interval

To have a means of determining whether a value of an IRC dot product is likely to arise from random chance, Walter, Nuhfer, and Suarez introduce what we will refer to as the "randomized trial confidence interval" [31]. To construct this interval, they determine the range of IRC dot product values from 10 000 randomized trials of simulated students. To create the simulated student populations, we use the sample size and score distribution for population a, the sample size and score distribution for population b, and the probabilities that a simulated student with a particular score in the overall population (a and b

combined) will choose a specific response [51]. The purpose of creating the randomized trial confidence interval is to determine the expected range of IRC dot product values if the two populations (a and b) were arbitrary subsets of the same overall population.

Each randomized trial involves creating two simulated populations, each with the same number of students as their respective corresponding real populations (a and b). Each simulated student is assigned to a scoring bin using weighted random sampling based on the corresponding real population's score distribution. This process ensures each simulated population's score distribution will closely mimic the corresponding real students' score distribution. Each simulated student is assigned an answer choice using weighted random sampling of the overall population's probabilities (a and b combined) for students in the same score bin. The two simulated populations are unbiased in that their simulated students select answer choices using the same probability distribution; the IRC dot product of the simulated populations would be a value of 1 if not for the effects of randomness.

Each randomized trial is completed by taking the simulated populations' IRC dot product. After completing 10 000 randomized trials, a range of outcomes is constructed, capturing 95% of the randomized trials (from the 2.5 to 97.5 percentiles). For any given item, when the real population's IRC dot product falls within the 95% randomized trial confidence interval, then the differences in IRCs for the real populations may result from random chance.

4. Comparing to purely random data

Given that the correct answer IRC for any population must have low frequencies at low scores and high frequencies for high scores, and given that all IRCs for incorrect answers must have low frequencies at high scores, the IRC dot product between any two populations will never be zero. To establish a reasonable lower bound for dot-product comparisons, we created three simulated student populations equal in number to RSW, IDWA, and IDWJ, respectively. Each of these simulated datasets was created under the assumption that respondents answered every item randomly, and then their overall score was determined. By taking the IRC dot product of a simulated student population with its associated real student population, we can estimate a minimum baseline value of the IRC dot product for completely unrelated datasets. The item and population-specific baseline dot product value can depend on the population's size, the number of answer choices, the number of answer choices serving as effective distractors, and the item difficulty.

C. Estimating uncertainty

To estimate the uncertainty in our (RSW) IRCs, we used 10 000 bootstrapped samples to generate a distribution of values for each data point [52,53]. We chose a random selection of student response sets (with replacement) in each instance to create a sample dataset of the same size as the RSW dataset. Response sets were kept intact to ensure that any correlations across items were preserved. We computed the value of each IRC data point for all items and all responses for each sample dataset. We include error bars on our IRCs to indicate the central 95% of each distribution [54].

For each sample dataset and for each item, we computed the IRC dot product between the sample dataset and the IDW American dataset as well as the IRC dot product between the sample dataset and the IDW Japanese dataset [30]. This process provides a distribution of IRC dot product values between the RSW dataset and each IDW dataset for each item. We used these distributions to create error bars representing the central 95% of each distribution, which we will refer to as the "IRC dot product confidence interval." We could not calculate the IRC dot product confidence intervals between the two IDW datasets because we did not have access to the raw data showing each student's responses to each item. Without the complete response set from each student, we could not generate similar sample datasets by randomly selecting students to include.

IV. SCORE DISTRIBUTIONS ARE NOT DIFFERENT ACROSS POPULATIONS

Ishimoto, Davenport, and Wittmann report that their datasets' score distributions are statistically different according to a two-tailed t test ($t = 8.8$, $p < 0.005$) [30]. Figure 2 shows the FMCE preinstruction score distributions for all three datasets: our data (labeled RSW), the IDW American dataset (IDWA), and the IDW Japanese dataset (IDWJ). Table I shows the descriptive statistics for these distributions.

The results from a Kruskal-Wallis rank-sum test show a statistically significant main effect between the three datasets, $\chi^2(2) = 143.66$, $p < 0.001$; moreover, the pairwise Wilcoxon rank-sum tests indicate that each dataset is significantly different from each other dataset for all three pairs, with $p < 0.001$. In contrast, the overall effect size of the comparisons was very small, with a 95% confidence interval of $\varepsilon^2 = [0.01, 0.02]$, and calculations of the 95% confidence intervals of the rank-biserial correlation r show that the effect sizes of these pairwise differences are

TABLE I. Descriptive statistics for all three datasets: our data (RSW), the American data from IDW (IDWA), and the Japanese data from IDW (IDWJ).

Statistic	RSW	IDWA	IDWJ
Sample size	6584	2348	1531
Mean score	10.46	9.05	11.22
Median score	8	7	9
Standard deviation	8.20	7.36	7.56

also small: $r = [0.09, 0.14]$ for RSW–IDWA, $r = [0.08, 0.14]$ for IDWJ–RSW, and $r = [0.19, 0.26]$ for IDWJ–IDWA [48,55]. The results from the pairwise Wilcoxon rank-sum tests agree with those of IDW, but we disagree with the interpretation that American students' score distribution differs meaningfully from that of Japanese students. The fact that the mean and median scores of our RSW dataset fall in between those of the IDW datasets, and the fact that the effect sizes between each pair of distributions are small, suggest that the score distributions of all three datasets are not meaningfully different (despite the statistical significance reported by the Kruskal-Wallis and Wilcoxon rank-sum tests) [56]. The similarities between score distributions may be seen in Fig. 2 with all three largely overlapping each other and showing similar shapes.

V. ITEM RESPONSE CURVE RESULTS

A. Results reported by Ishimoto, Davenport, and Wittmann

Based on qualitative comparisons of their IRCs, Ishimoto, Davenport, and Wittmann reported that most IRCs are quite similar for their American and Japanese datasets [30]; however, they reported several notable exceptions.

- On items 3 and 7 (a sled being pushed and slowing down), the most common incorrect response for both datasets is that the force is in the direction of motion and decreasing, but Japanese students are more likely than American students to choose an accompanying force that is opposite the direction of motion and increasing.
- Items 11, 12, 13 (force on a coin flipped vertically in the air) appeared more difficult for the Japanese students than the American students. In contrast, both groups did equally well on isomorphic items 8, 9, and 10, which use a toy car moving up and down a ramp.
- Items 16 and 18 "appear to be slightly easier for the American students" [30]. These items require students to choose a graph of force vs time for a toy car moving to the right and either speeding up or slowing down, respectively.
- Item 22 (choosing a graph of acceleration vs time for a toy car moving to the right and speeding up) was easier for American students than Japanese students.
- Items 30–32 and 34 (interaction forces during collisions) differed between American and Japanese students. American students selecting an incorrect answer are likely to use mass-dependence reasoning on item 30 (larger truck and smaller car moving at the same speed toward each other) and action-dependence reasoning on items 32 and 34 (a moving car hits a stationary truck of either greater or equal mass, respectively) [42]. By contrast, Japanese students are much more likely to answer these questions correctly. One exception is that American students are similarly likely to answer item 31 correctly (a small fast car collides with a large slow truck); this may relate to a tendency of some students to see the opposite effects of mass dependence and action dependence as balancing to result in equal forces [42].
- Items 44–47 (selecting velocity or kinetic energy of a sled after going down a hill) are easier for Japanese students than American students.

To test these claims, we report the results from our dot product analyses for pairwise comparisons of the three datasets and the dot product analyses for each of the datasets compared to a dataset of entirely random responses. For each pairwise comparison of datasets, we also show the range of IRC dot product values expected if the differences between the datasets resulted from random chance. We also examine the nature of the differences between datasets for the individual items listed in IDW as notably different across datasets.

B. IRC dot product results and comparisons to randomly selected responses

Figure 3 shows the IRC dot products for each item for each of the three pairs of populations: RSW-IDWJ; RSW-IDWA; and IDWJ-IDWA. Based on the similarities of the score distributions, one might expect that the RSW dataset would produce IRCs that are equally similar to those of the IDWA and IDWJ datasets, but Fig. 3 shows us that this is not the case. The IRC dot product values for each item for the RSW-IDWA comparison (black circles) show that the two American populations' IRCs are highly similar to each other, with all values within the range [0.9771, 0.9996]. Consequently, we find that values of the IRC dot product for each item for the RSW-IDWJ and the IDWJ-IDWA pairs are quite close to each other, both with much broader distributions of IRC dot product values: RSW-IDWJ values

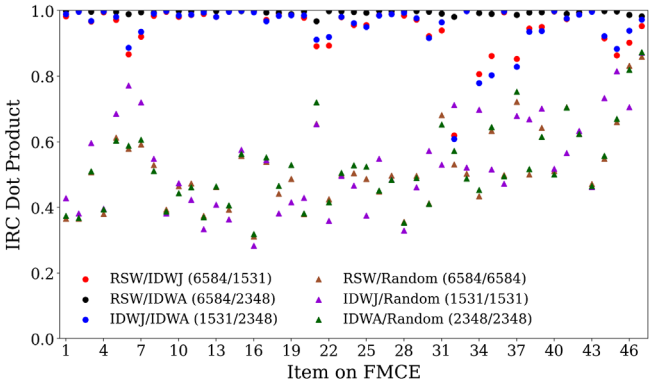

FIG. 3. The IRC dot products for all three pairs of populations shown in Figs. 5, 6, and 7. The IRCs of the American students (RSW and IDWA) are quite similar, and thus the IRC dot product values of the Japanese students (IDWJ) with either set of American students (RSW or IDWA) are similar for each item. IRC dot products are also shown for each population compared to a simulated set of random responses.

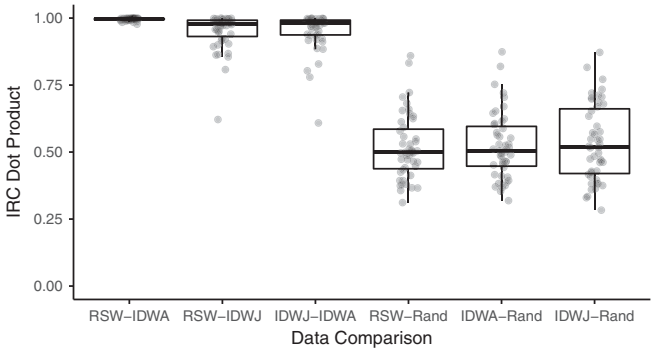

FIG. 4. Distributions of the IRC dot product values for each of the comparisons shown in Fig. 3. The RSW-IDWA comparison has a very narrow distribution near unity. The RSW-IDWJ and IDWJ-IDWA distributions are very similar and lower than RSW-IDWA. Each comparison to a simulated dataset of random responses has a distribution that is noticeably lower than any of the comparisons between real datasets.

fall within the range [0.6214, 0.9997], and IDWJ-IDWA values are within the range [0.6087, 0.9993]. The box plots in Fig. 4 show the distribution of IRC dot product values for each comparison.

More detailed analyses of several specific IRCs are included in Sec. V D. We include IRCs for all items in the Supplemental Material [57].

Figures 3 and 4 also show the results for each population (RSW, IDWA, and IDWJ) paired with an equal number of simulated students who are randomly guessing. For the vast majority of items, the comparisons within the three experimental datasets have considerably higher IRC dot product values than comparisons between any experimental datasets and the random dataset; i.e., they are much closer to each other than with a random dataset. One major exception to this trend is item 32, for which the IRC dot products between IDWJ and either RSW or IDWA are around 0.61, well within the range of the IRC dot product values involving random data. Some other items, such as 45 and 46, have IRC dot product values between datasets that approach those involving random data, but they are still distinct.

The comparisons to simulated random datasets in Figs. 3 and 4 provide essential information and context when interpreting the meaning of the IRC dot product values. Theoretically, the value of a dot product must fall within the range [0, 1]; however, the lowest values seen in Fig. 3 are slightly below 0.3 (item 16). Also, the IRC dot products' values may be affected by the structure of the test itself. Items 1–29 have eight or nine possible answer choices for each item, but items 44–47 have only five answer choices. As such, simulated random responses are more likely to match up with actual data for items 44–47, and we see larger values (in general) of the IRC dot products comparing random responses for those items (none is below 0.5). The number of answer choices, however, is not the only factor related to the IRC dot products with random data: items 1–4 have lower values than items 5–7, even though they all have eight possible answer choices and involve the same physical scenario (a sled being pushed across an icy lake). The values of the IRC dot product of a real population with purely random data depend on the shape of the real populations' IRCs and the real populations' score distribution. For example, purely random guessing will not mimic the IRC for an answer choice that acts as a distractor with an intermediate maximum.

C. IRC dot products compared to uncertainty ranges

1. Reexamining the comparisons of Ishimoto, Davenport, and Wittmann

Figure 5 shows the IRC dot product for each item comparing the Japanese (IDWJ) and American (IDWA) student populations used in the IDW study [30]. The black error bars in Fig. 5 show the randomized trial confidence interval for each item. The IRC dot product values for 44 of the 47 items fall outside the range of their respective randomized trial confidence intervals, suggesting that the differences between the datasets are likely not due to random chance and represent meaningful differences in student response patterns. The only items with IRC dot product values within their randomized trial confidence interval are items 4, 15, and 40. These items are notable in that students respond to all three of these by overwhelmingly selecting either the correct response or a single incorrect response (out of eight or nine options), but they are not unique. Other items with similar response patterns do not have dot products within their randomized trial confidence intervals. Similar to item 4, more than 95% of students choose either the correct response or a single incorrect response for items 1 and 16. Similar to items 15 and 40, more than 85% of students choose the correct response for items 33 and 43 before instruction.

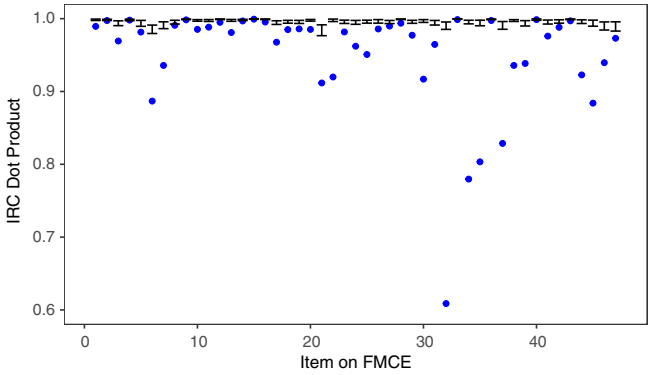

FIG. 5. The IRC dot product for each item on the FMCE is shown in blue for the IDW American (IDWA) and Japanese (IDWJ) datasets [30]. The black error bars represent the randomized trial confidence intervals described in Sec. III B 3.

As mentioned above, IDW reported 16 items that they noticed having different IRCs between their two datasets: 3, 7, 11–13, 16, 18, 22, 30–32, 34, and 44–47 [30]. All of these items have IRC dot product values below their randomized trial confidence intervals in Fig. 5, and many of them are well outside the range (e.g., item 34 has an IRC dot product value of 0.780, and a randomized trial confidence interval of [0.993, 0.998]). We find some consistencies when looking at specific items.

- Items 3 and 7 have the lowest IRC dot product values of the force sled item cluster (items 1–4, 7).
- Items 11–13 have lower IRC dot product values than items 8–10. The second item has the highest value for each of these groups, and the third item has the lowest value.
- Item 22 has a lower IRC dot product value than any of the other items in the acceleration graphs cluster (items 22–26).
- The items involving Newton's third law in the context of cars and trucks (30–38) include those with the lowest IRC dot product values. Items 32 and 34 have the lowest IRC dot product values, indicating the largest differences between the IRCs, and item 31 has a much higher value, indicating better agreement.

IDW report that items 16 and 18 may be easier for American students than Japanese students. Both of these force graphs cluster items have IRC dot product values below their respective randomized trial confidence intervals. Item 21 has the lowest IRC dot product value in the force graphs cluster (items 14, 16–21) and thus the most pronounced differences in its IRCs in the region where most students scored. Visual inspection of the item 21 IDWJ-IDWA IRCs, which are highly similar to the item 21 RSW-IDWJ IRCs shown in Fig. 8, show that item 21 is easier for both the RSW and IDW American students than the IDW Japanese students.

Figure 5 also reveals additional items that seem to have significantly different IRCs between the two datasets. For example, item 38 has a fairly low IRC dot product of 0.936, which is well outside the random trial confidence interval [0.996, 0.999], but item 38 is not highlighted in the IDW study as being particularly different between American and Japanese students. In fact, IDW mention items 36 and 38 together as showing similarities between their two populations; our IRC dot product analyses support this conclusion for item 36 (IRC dot product of 0.997), but not for item 38. This suggests that some differences between IRCs may be less salient upon visual inspection.

2. Comparing two American datasets

Figure 6 shows the IRC dot products comparing the IRCs of both of the American datasets: our RSW dataset and the IDWA dataset from Ref. [30]. The vertical range of Fig. 6(a) was selected to allow easy comparisons with Figs. 5 and 7, even though the minimum IRC dot product value is 0.977 (item 21). Figure 6(b) shows the same information as Fig. 6(a), but with a smaller vertical range to allow a more detailed examination of the results. The IRC dot product values of 24 of the 47 items fall within their randomized trial confidence intervals, indicating that any differences between them are potentially the result of random chance. The IRC dot product values of 22 items are below their randomized trial confidence intervals. One item (item 26) has an IRC dot product value above its randomized trial confidence interval, indicating a high similarity between those IRCs.

To quantitatively compare the IRCs of any two populations, we define a comparison metric that is similar to an effect size. The "dot product effect size" (DES) for item i is defined as

$$\mathrm{DES}_i \equiv \frac{\mathrm{RT}_{\mathrm{mid},i} - \langle \mathbf{a} \cdot \mathbf{b} \rangle_i}{\mathrm{CI}_{\mathrm{pooled},i}}, \tag{2}$$

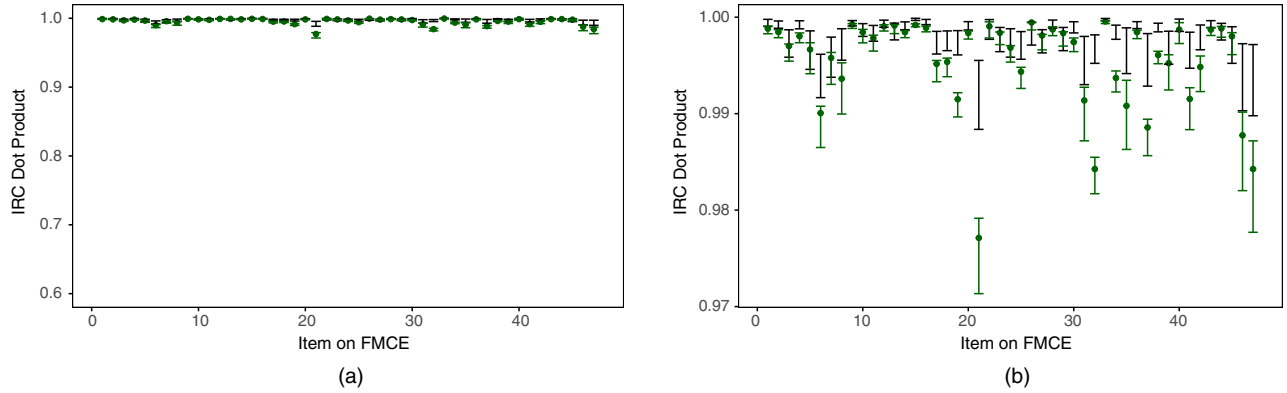

FIG. 6. The IRC dot product for each item on the FMCE is shown in green for our American dataset (RSW) and the IDW American dataset (IDWA). The green error bars around the data points represent the IRC dot product confidence intervals described in Sec. III C. The black error bars represent the randomized trial confidence intervals described in Sec. III B 3. The vertical range of (a) matches Figs. 5 and 7. Plot (b) contains the same information as (a), but a smaller vertical range is chosen to allow better visualization.

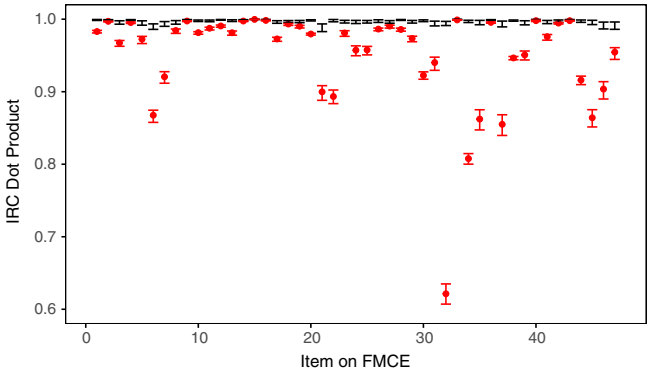

FIG. 7. The IRC dot product for each item on the FMCE is shown in red for our American dataset (RSW) and the IDW Japanese dataset (IDWJ). The red error bars around the data points represent the IRC dot product confidence intervals described in Sec. III C. The black error bars represent the randomized trial confidence intervals described in Sec. III B 3.

where $\mathrm{RT}_{\mathrm{mid},i}$ is the midpoint of the randomized trial confidence interval of two populations for item i, and $\langle \mathbf{a} \cdot \mathbf{b} \rangle_i$ is the IRC dot product between populations a and b for item i. The pooled confidence interval $\mathrm{CI}_{\mathrm{pooled},i}$ is defined as

$$\mathrm{CI}_{\mathrm{pooled},i} \equiv \sqrt{\mathrm{CI}_{\mathrm{RT},i}^2 + \mathrm{CI}_{\mathrm{dot},i}^2}, \tag{3}$$

where $\mathrm{CI}_{\mathrm{RT}}$ is the size of the randomized trial confidence interval (e.g., the black error bars in Fig. 6) and $\mathrm{CI}_{\mathrm{dot}}$ is the size of the IRC dot product confidence interval (e.g., the green error bars in Fig. 6) [58].

Table II shows the DES value for each item for the comparison between our RSW dataset and the IDW American dataset (RSW vs IDWA columns). Comparing Table II with Fig. 6 provides a visual interpretation for our DES values: items 8, 31, and 46 all have error bars that are just barely touching, and they all have DES values of 0.6 [59]. Over half of the items (28 out of 47) have DES values at or below 0.6, indicating that the error bars overlap, and there is a reasonable likelihood that the value of the IRC dot product is high enough for the RSW and IDWA IRCs to be considered the same. Of the remaining 19 items, only one (item 32) has a DES value above 2. Additionally, nine items have negative DES values, indicating that the IRC dot products for those items are higher than the midpoint of the respective randomized trial confidence intervals, suggesting that those IRCs are highly similar.

3. Comparing a new American dataset to the previous Japanese dataset

Figure 7 shows the IRC dot products comparing the IRCs of our American dataset (RSW) and the IDW Japanese dataset. These are highly similar to the IRC dot products shown in Fig. 5 comparing the IDW Japanese and American students. This result is to be expected given that the IRC dot products shown in Figs. 3 and 6 show the IRCs of both American populations are highly similar. For the datasets in Fig. 7, 46 out of the 47 items have IRC dot

TABLE II. The dot product effect size values for all comparisons between our RSW dataset, and the IDW American and Japanese datasets. Bolded entries indicate items that were highlighted by IDW as being particularly different between their American and Japanese datasets.

Item	RSW vs IDWA	RSW vs IDWJ
1	0.6	4.2
2	0.7	1.0
3	0.1	**3.0**
4	0.9	1.4
5	0.0	1.9
6	0.6	6.6
7	0.0	**4.2**
8	0.6	1.4
9	0.1	1.1
10	0.1	4.1
11	0.2	**2.3**
12	0.3	**2.7**
13	−0.3	**2.7**
14	0.6	0.8
15	1.5	0.0
16	0.8	**0.7**
17	0.7	3.5
18	0.8	**0.9**
19	1.6	1.4
20	0.6	6.2
21	1.4	3.9
22	−0.1	**5.5**
23	−0.2	1.8
24	0.1	2.7
25	0.8	3.2
26	−0.5	2.1
27	−0.2	1.1
28	0.4	3.0
29	−0.2	2.7
30	0.8	**7.2**
31	0.6	**2.8**
32	2.6	**13.1**
33	0.5	1.2
34	1.8	**12.7**
35	0.7	4.7
36	0.7	1.7
37	1.1	4.7
38	1.8	10.9
39	0.3	3.3
40	0.2	0.4
41	0.9	2.3
42	0.7	0.7
43	0.5	1.0
44	−0.2	**6.7**
45	−0.2	**5.4**
46	0.6	**3.4**
47	0.8	**1.9**

product values below the range of their respective randomized trial confidence intervals (only item 15 is within the confidence interval).

Table II also shows the DES value for each item for the comparison between the RSW and IDW Japanese (IDWJ) datasets. The DES values in the RSW vs IDWJ column are generally larger than the DES values in the RSW vs IDWA column. Only three items have smaller DES values, and two of these (items 16 and 19) are within 0.2 of the RSW vs IDWA value [60]. The median DES value is 2.7 for the RSW–IDWJ comparison (compared to 0.6 for RSW–IDWA), and the highest value is 13.1 (again for item 32). Over half of the items (29 out of 47) have DES values above 2, compared to only one item for the RSW vs IDWA results. Only two items (15 and 40) have DES values less than 0.6 (interpreted above as having a reasonable expectation of having similar IRCs) [61].

The items highlighted as being notably different by IDW (mentioned above) are shown in bold in Table II. These include the two highest DES values (items 32 and 34), but they do not include all of the highest values (e.g., item 38 is omitted from their discussion). Quantitatively comparing IRCs by computing the IRC dot product, combined with our methods for generating confidence intervals, can reveal different patterns than qualitative visual comparisons. Of particular interest are items 16 and 18, which IDW highlighted as seeming easier for American students. Similar to what we mentioned above with the IDWA vs IDWJ comparison, we find other items in the force graphs cluster (items 14, 16–21) to have greater differences than items 16 and 18; moreover, Table II also shows that the RSW vs IDWJ comparisons for items 16 and 18 are very similar to the RSW vs IDWA comparisons. The differences reported by IDW for these two items may not be generalizable for all American students. Visual inspection of Fig. 8 shows that item 21 seems to be somewhat easier for the RSW American students than the IDW Japanese students: the RSW IRC for the correct response E is slightly to the left of the IDWJ IRC, and more RSW students select E in the most populated region of the plot (fractional scores between about 0.1 and 0.3). The absence of a single dominant incorrect response and the differences in selected answer choices in the populated region lead to a higher DES value than either item 16 or 18 (see Table II).

We cannot compare the DES results between RSW and IDWJ to those between IDWA and IDWJ because we cannot calculate the DES values for the IDWA vs IDWJ comparison. The calculation of the DES depends on the value of the IRC dot product confidence interval, which depends on being able to generate sample datasets from at least one of the datasets involved in the comparison. As mentioned in Sec. III C, we could not sample datasets from either the IDWA or IDWJ datasets because we do not have the raw data containing student responses. For this reason, we do not include blue error bars in Fig. 5.

D. Examining specific items

The IRC dot product for item 32 between an American dataset and the Japanese dataset is the lowest value of any item on the FMCE (0.621 for RSW vs IDWJ and 0.609 for IDWA vs IDWJ), suggesting the IRCs differ more for this item than any other. Figure 9 shows us that the American students in the RSW dataset are more likely to choose B (consistent with the IDWA results) while the Japanese students (IDWJ) are more likely to choose F. Item 32 involves a small car colliding with a large stationary truck. Answer B indicates that the car exerts a larger force on the truck than the truck on the car, which is consistent with the idea that faster, or more active, objects exert more force (a.k.a. action dependence [42]). Answer F indicates that there is not enough information to determine the relative

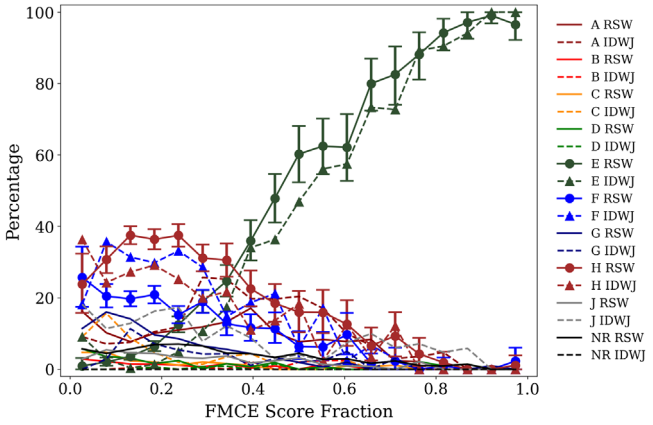

FIG. 8. Item 21 IRCs: the solid lines show the IRCs for the RSW dataset, and the dashed lines show the IRCs for the IDWJ dataset. NR = no response. The error bars on E, F, and H RSW IRCs are the 95% confidence intervals of 10 000 bootstrapped sample datasets based on RSW data described in Sec. III C.

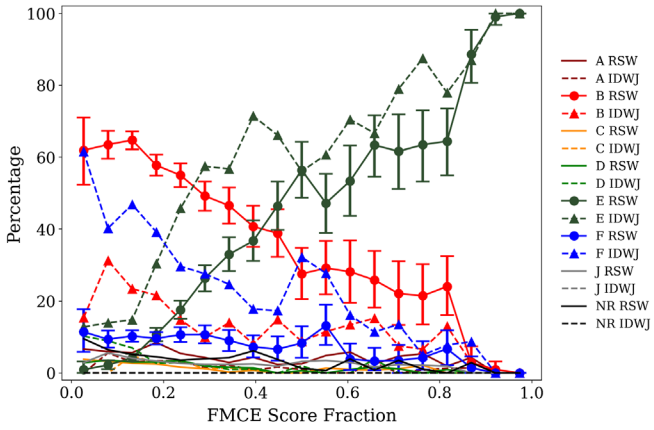

FIG. 9. Item 32 IRCs: the solid lines show the IRCs for the RSW dataset, and the dashed lines show the IRCs for the IDWJ dataset. NR = no response. The error bars on B, E, and F RSW IRCs are the 95% confidence intervals of 10 000 bootstrapped sample datasets based on RSW data described in Sec. III C.

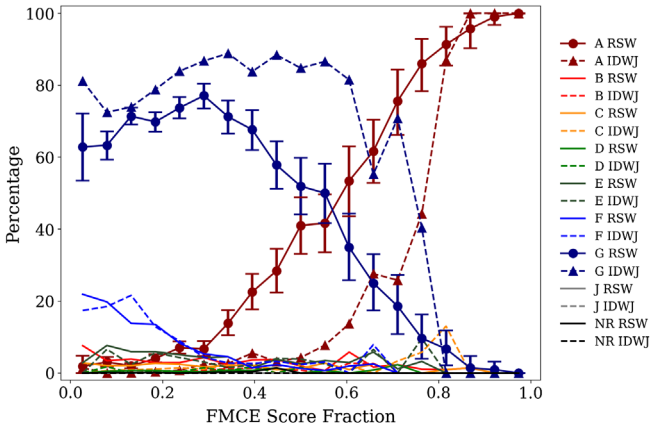

FIG. 10. Item 11 IRCs: the solid lines show the IRCs for the RSW dataset, and the dashed lines show the IRCs for the IDWJ dataset. NR = no response. The error bars on A and G RSW IRCs are the 95% confidence intervals of 10 000 bootstrapped sample datasets based on RSW data described in Sec. III C.

magnitudes of the interaction forces between the car and the truck. IDW claim that this is most likely due to different life experiences between American and Japanese students, where Americans are more likely to grow up using automobiles for commuting. Our results confirm that large portions of American students choose answer B. The most considerable differences between Japanese and American students' responses occur at low scores, which contain most of the students (see Fig. 2).

Figure 10 shows the IRCs of our American (RSW) and the IDW Japanese (IDWJ) datasets for item 11, which examines the force acting on a coin tossed upward while still ascending. From visual inspection, we see that the IRCs that differ the most are for answer choices A (the correct answer choice) and G (the most common incorrect answer) over the fractional score range of 0.3 to 0.8; however, the IRC dot product for this pair of populations on item 11 is 0.988, suggesting that the IRCs are much more similar than for item 32. While the large differences in IRC dot product values between item 11 and 32 may not be evident from viewing the IRCs, the IRC dot product value depends on each population's score distributions. The score distributions' peaks are below 0.2 for all three datasets, with over half of the students in the fractional score range 0.1 to 0.3. The large differences between the item 32 IRCs in this highly-populated range lead to a much lower IRC dot product value.

Similar comparisons can be made between the IRCs for item 11 (Fig. 10) and item 21 (Fig. 8). It would be reasonable to look at these two plots and conclude that the two populations are more different on item 11 than they are for item 21: the difference between the RSW and IDWJ IRCs for the dominant (and correct) answer choice for item 21 (E) appear much smaller than the differences between the two dominant answer choices for item 11; however, Table II shows that item 21 has a higher DES value, and is, therefore, less likely to be attributed to random chance. To resolve this apparent discrepancy, we must again focus on the plot's most populous region: scores from 0.1 to 0.3. In this region item 11 shows three IRCs with non-negligible percentage of being chosen: correct (A), dominant incorrect (G), and one other incorrect (F), which is mostly chosen by the lowest-scoring students. In contrast, item 21 seems to have five IRCs showing answers chosen by more than 10% of students in each dataset in this score range, several with notable differences between the datasets. In the highly populated score range of 0.1 to 0.3, we see the biggest differences between RSW and IDWJ on item 32, followed by item 21, and the most similarity (of these three items) with item 11; this is consistent with both the IRC dot product values in Fig. 7 (0.621, 0.900, and 0.988 for items 32, 21, and 11, respectively) and the DES results in Table II (13.1, 3.9, and 2.3 for items 32, 21, and 11, respectively).

VI. CONCLUSIONS

Our analyses support the results reported by Ishimoto, Davenport, and Wittmann [30]: American physics students' and Japanese students' preinstruction responses to the Force and Motion Conceptual Evaluation differ for many items in nontrivial ways. The majority of these differences are related to specific choices of incorrect answers rather than differences in numbers of correct answers. Because we do not have access to a disjoint dataset of Japanese physics students' responses to the FMCE, we could not fully replicate IDW's work by performing completely independent comparisons between American and Japanese students. However, our work strongly suggests that the IDW American dataset is representative of physics students in the U.S. or at least representative of the student populations typically sampled for research purposes. One caveat to these results is the well-documented tendency for physics education research studies to oversample from large research universities, which tend to have students with more previous exposure to formal physics instruction than is typical for American college and university physics students in general [62]. Because of the anonymous nature of the majority of our dataset (provided by the PhysPort Data Explorer), we are unable to determine whether our student population is truly representative of all American physics students; however, the similarity between our results and the IDW study are quite suggestive.

In terms of score distributions, our dataset fell between the IDW American and Japanese datasets. Similar to IDW, we find these differences to be statistically significant ($p < 0.001$ for both the Kruskal-Wallace, and the pairwise Wilcoxon rank-sum tests), but we do not consider these differences to be pedagogically meaningful. The differences in the mean and median scores between all three datasets are within about 2 out of 37 points, and the effect sizes are small. Additionally, looking at these score distributions alone would likely not impact an instructor's approach to teaching

without having more context highlighting the differences between the populations.

IRCs can reveal pedagogically important differences between student populations. The RSW IRCs are very similar to the IRCs for the IDW American dataset for the vast majority of items. The IRC dot product values are very high overall, and random chance can account for the differences between IRCs for many of the items. Conversely, the IRC dot product values between the RSW and the IDW Japanese datasets are much lower for many items. Further, the dot product effect size (DES) values suggest that these differences are not random fluctuations but are likely due to actual differences in how the two student populations select answer choices for specific FMCE items. Additionally, the IRC dot product values comparing the RSW and IDW Japanese datasets are very similar to the IRC dot product values comparing the IDW American and Japanese datasets, providing additional support to our claim that our results largely support those reported by IDW.

Of items in the force graphs cluster, IDW reported that items 16 and 18 may be slightly easier for American students based on visual qualitative comparisons of the IRCs [30]. We find that item 21 is slightly easier for American students more so than items 16 and 18. When visually comparing IRCs, we focus primarily on the region where the majority of students scored—the vast majority earn fractional scores less than 0.4 (15 out of 37). Differences between IRCs are more meaningful when they are representative of more students, and a benefit of computing the IRC dot product is that each score bin is weighted by the number of students. Using qualitative visual comparisons and quantitative IRC dot products can provide a robust way to compare IRCs and gain valuable insights into similarities and differences between how different groups select answers to the FMCE and other multiple-choice test items.

VII. IMPLICATIONS FOR INSTRUCTION

Our dataset from many American colleges and universities is similar to the disjoint IDW dataset from many American colleges and universities. This is not to say that every class of physics students at an American college or university will select the same proportions of specific answers to the FMCE as shown in our data. One of the strongest results we see from comparing Japanese and American students is that students with different backgrounds answer the FMCE differently, and these differences do not necessarily show up in score distributions that focus only on the number of items answered correctly.

As mentioned above, slight differences in preinstruction score distributions may have little impact on instructional methods, but large differences in the incorrect responses students select could (and should) inform an instructor's approach to a class. Figure 9 illustrates this wonderfully: the RSW American students mostly choose one incorrect answer for item 32 (B, consistent with the idea that faster objects exert more force than slower objects), but the IDW Japanese students mainly select a different choice (F, indicating that more information is needed to determine the forces). Both of these answers are incorrect, but the difference could represent vastly different ways of thinking about the forces involved in two-body collisions; moreover, the differences between these incorrect IRCs are more pronounced than the correct IRCs (answer E), suggesting that an inspection of correct answers alone may hide the magnitude of the differences between these populations. An essential feature of good physics instruction is acknowledging students' initial ideas and using them to facilitate growth in understanding. Looking only at correct answers cannot accomplish this, but IRCs can show if different groups of students choose different answers (e.g., those with low, middle, and high scores).

ACKNOWLEDGMENTS

We thank Sam McKagan and Ellie Sayre for providing access to data from PhysPort's Data Explorer and all of the instructors who were willing to share their students' FMCE responses. We also thank Nicholas Baltera, Paul Kelly, Maria Lentini, and Mitchell Nussenbaum for their previous contributions as research team members. We are deeply grateful to Glen Davenport for his assistance on this project. The National Science Foundation supported this project through Grant No. DUE-1836470.

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

[2] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula, Am. J. Phys. **66**, 338 (1998).

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

[4] A. Madsen, S. B. McKagan, E. C. Sayre, and C. A. Paul, Resource Letter RBAI-2: Research-based assessment instruments: Beyond physics topics, Am. J. Phys. **87**, 350 (2019).

[5] 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).

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

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

[8] J. Von Korff, B. Archibeque, K. Alison 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, Am. J. Phys. **84**, 969 (2016).

[9] 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).

[10] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010124 (2018).

[11] 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).

[12] E. Brewe, 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).

[13] J. Wells, R. Henderson, J. Stewart, G. Stewart, J. Yang, and A. Traxler, Exploring the structure of misconceptions in the Force Concept Inventory with modified module analysis, Phys. Rev. Phys. Educ. Res. **15**, 020122 (2019).

[14] J. Yang, J. Wells, R. Henderson, E. Christman, G. Stewart, and J. Stewart, Extending modified module analysis to include correct responses: Analysis of the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **16**, 010124 (2020).

[15] I. T. Griffin, K. J. Louis, R. Moyer, N. J. Wright, and T. I. Smith, A multi-faceted approach to measuring student understanding, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA*, edited by D. L. Jones, L. Ding, and A. Traxler (2016), pp. 132–135, https://doi.org/10.1119/perc.2016.pr.028.

[16] T. I. Smith, K. A. Gray, K. J. Louis, B. J. Ricci, and N. J. Wright, Showing the dynamics of student thinking as measured by the FMCE, in *Proceedings of the 2017 Physics Education Research Conference*, Cincinnati, OH, edited by L. Ding, A. Traxler, and Y. Cao (2017), pp. 380–383, https://doi.org/10.1119/perc.2017.pr.090.

[17] J. Yang, C. Zabriskie, and J. Stewart, Multidimensional item response theory and the force and motion conceptual evaluation, Phys. Rev. Phys. Educ. Res. **15**, 020141 (2019).

[18] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).

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

[20] 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).

[21] J. Stewart, B. Drury, J. Wells, A. Adair, R. Henderson, Y. Ma, Á. Pérez-Lemonche, and D. Pritchard, Examining the relation of correct knowledge and misconceptions using the nominal response model, Phys. Rev. Phys. Educ. Res. **17**, 010122 (2021).

[22] P. Eaton, K. Johnson, and S. Willoughby, Generating a growth-oriented partial credit grading model for the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **15**, 020151 (2019).

[23] 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).

[24] 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).

[25] P. Eaton, Evidence of measurement invariance across gender for the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **17**, 010130 (2021).

[26] 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, Am. J. Phys. **74**, 449 (2006).

[27] 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).

[28] P. J Walter and G. Morris, Assessing student learning and improving instruction with transition matrices, in *Proceedings of the 2016 Physics Education Research Conference*, Sacramento, CA edited by D. L Jones, L. Ding, and A. Traxler (2016), pp. 376–379, https://doi.org/10.1119/perc.2016.pr.089.

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

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

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

[32] M. Ishimoto, R. K. Thornton, and D. R. Sokoloff, Validating the Japanese translation of the Force and Motion Conceptual Evaluation and comparing performance levels of American and Japanese students, Phys. Rev. ST Phys. Educ. Res. **10**, 020114 (2014).

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

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

[35] Some researchers have suggested that an intermediate maximum may indicate that students with a particular misconception are attracted to that answer choice [21,36].
[36] 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* (Association for Computing Machinery, New York, NY, 2019), pp. 1–10, https://doi.org/10.1145/3330430.3333637.
[37] PhysPort, Data explorer (2017).
[38] R. J. de Ayala, *The Theory and Practice of Item Response Theory* (Guilford Press, New York, NY, 2008), ISBN 978-1593858698.
[39] D. Thissen, L. Cai, and R. Darrell Bock, The nominal categories item response model. in Handbook of Polytomous Item Response Theory Models, edited by M. L. Nering and R. Ostini (Routledge/Taylor & Francis Group, New York, 2010), Chap. 3, pp. 43–75.
[40] R. Darrell Bock and I. Moustaki, Item response theory in a general framework, in *Handbook of Statistics* edited by C. R. Rao and S. Sinharay (Elsevier, 2007), Vol. 26, Chap. 15, pp. 469–514.
[41] In principle, IRCs could be created to mimic a multidimensional IRT analysis by identifying subsets of items that could be used to calculate subscores associated with a particular topic on the test (e.g., Newton's third law). These IRCs would be of limited utility because the score axis would be limited by the number of items in each subset, probably in a range of 4 to 9 items based on item clusters previously identified [6,42].
[42] T. I. Smith and M. C. Wittmann, Applying a resources framework to analysis of the Force and Motion Conceptual Evaluation, Phys. Rev. ST Phys. Educ. Res. **4**, 020101 (2008).
[43] R Core Team, R: A Language and Environment for Statistical Computing (2020).
[44] R. Philip Chalmers, mirt: A multidimensional item response theory package for the R environment, J. Stat. Softw. **48**, 1 (2012).
[45] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis, Nat. Methods **9**, 671 (2012).
[46] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar, Moving to a World Beyond "$p < 0.05$", Am. Statistician **73**, 1 (2019).
[47] R. L. Wasserstein and N. A. Lazar, The ASA's statement on p-values: Context, process, and purpose, Am. Statistician **70**, 129 (2016).
[48] M. S. Ben-Shachar, D. Lüdecke, and D. Makowski, effectsize: Estimation of effect size indices and standardized parameters, J. Open Source Software **5**, 2815 (2020).
[49] E. B. Nuhfer, C. B. Cogan, C. Kloock, G. G. Wood, A. Goodman, N. Zayas 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, J. Microbiol. Biol. Educ. **17**, 143 (2016).
[50] The actual minimum value of an IRC dot product is significantly higher than 0, which we discuss below.
[51] For a simulated student in the overall population (a and b combined) with a score in score bin j, the probability of choosing response i on item k is $(a_{ijk} + b_{ijk})/(n_{a,j} + n_{b,j})$.
[52] A. Canty and B. Ripley, boot: Bootstrap R (S-Plus) Functions (2020).
[53] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Applications* (Cambridge University Press, Cambridge, England, 1997).
[54] For visual clarity, we only include error bars on curves with values above 25% beyond the two lowest score bins. See Fig. 1 for an example.
[55] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Lawrence Erlbaum Associates, Hillsdale, NJ, 1988).
[56] We believe the statistical significance of these differences may be the result of having very large datasets. Small differences may be categorized as statistically significant, even though they may not be particularly meaningful in the context of the measured outcome [46,47].
[57] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.17.020127 for two sets of IRC plots: one that shows the IRCs for the RSW and IDWA datasets for each item, and one that shows the IRCs for the RSW and IDWJ datasets for each item.
[58] Our "dot product effect size" is not an effect size by Cohen's traditional definition because he uses standard deviation as the scaling factor in the denominator [55]. To get a traditional effect size, one must multiply the DES by a factor of 3.92.
[59] When the confidence intervals are just barely touching, the DES value would be 1 if the IRC dot product value was at the center of the IRC dot product confidence interval. When the IRC dot product value is close to 1, it will be higher than the midpoint of the IRC dot product confidence interval due to ceiling effects.
[60] Item 15 seems to be an outlier in that the DES for the IDWJ comparison is much smaller than the IDWA comparison. We believe this to be attributable to a ceiling effect: the randomized trial confidence interval for item 15 in Fig. 7 has a range of [0.9996, 0.9999]. Item 15 is one of the easiest on the FMCE, with most students answering the item correctly before instruction, which is why item 15 is typically omitted from calculations of an overall score.
[61] Items 15 and 40 are two of the easiest items on the FMCE, with most students answering them correctly before instruction: 94% correct on item 15 and 86% correct on item 40 in our RSW dataset.
[62] S. Kanim and X. C. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16**, 020106 (2020).