Use of item response curves of the Force and Motion Conceptual Evaluation to compare Japanese and American students' views on force and motion

Michi Ishimoto,[1] Glen Davenport,[2] and Michael C. Wittmann[3]
[1]*Kochi University of Technology, Tosayamada-cho, Kami-shi, Kochi 782-8502, Japan*
[2]*Department of Educational Psychology, Neag School of Education, University of Connecticut, 249 Glenbrook Rd., Storrs, Connecticut 06269, USA*
[3]*Department of Physics and Astronomy, and Center for Research in STEM Education, University of Maine, Orono, Maine 04469, USA*
(Received 24 May 2017; published 30 November 2017)

Student views of force and motion reflect the personal experiences and physics education of the student. With a different language, culture, and educational system, we expect that Japanese students' views on force and motion might be different from those of American students. The Force and Motion Conceptual Evaluation (FMCE) is an instrument used to probe student views on force and motion. It was designed using research on American students, and, as such, the items might function differently for Japanese students. Preliminary results from a translated version indicated that Japanese students had similar misconceptions as those of American students. In this study, we used item response curves (IRCs) to make more detailed item-by-item comparisons. IRCs show the functioning of individual items across all levels of performance by plotting the proportion of each response as a function of the total score. Most of the IRCs showed very similar patterns on both correct and incorrect responses; however, a few of the plots indicate differences between the populations. The similar patterns indicate that students tend to interact with FMCE items similarly, despite differences in culture, language, and education. We speculate about the possible causes for the differences in some of the IRCs. This report is intended to show how IRCs can be used as a part of the validation process when making comparisons across languages and nationalities. Differences in IRCs can help to pinpoint artifacts of translation, contextual effects because of differences in culture, and perhaps intrinsic differences in student understanding of Newtonian motion.

DOI: 10.1103/PhysRevPhysEducRes.13.020135

I. INTRODUCTION

Student views of force and motion develop within the context of personal experience and background characteristics. As such, one might expect to find that students from different nations have different ideas about force and motion. Concept inventories such as the Force Concept Inventory (FCI) [1] and the Force and Motion Conceptual Evaluation (FMCE) [2] are widely used to probe American students' views on force and motion and to compare the effectiveness of different instructional methods. These assessments are distinct because they include distractors that represent common incorrect ideas that have been identified by American physics education researchers. Accurate evaluation of conceptual understanding depends on the effectiveness of these distractors, which are attractive to students with typical, naïve versions of physics concepts.

Concept inventories have crossed national boundaries, with English versions being used in other countries and translated versions appearing in non-English-speaking countries. For example, the FCI is widely used to assess instructional effectiveness and to compare the conceptual understanding of students from different nations [3]. These uses of the FCI are based on the assumption that the response data of non-American students are comparable to those of American students. However, evaluating non-American students' use of American concept inventories requires additional validation, especially if surveyed students are novice learners and if translation is involved because subtle contextual differences in the distractors could decrease data reliability and validity. Therefore, the validation process should investigate the test itself, students' interactions with test items, and the structure of conceptual knowledge in different nations. The validated use of a concept inventory would provide data with comparable reliability to that of American data and an accurate assessment of instructional effectiveness.

A previous study [4] using American and Japanese FMCE data found that the most common incorrect responses were the same across the two samples. The translated version of the FMCE met some of the quality standards of classical test theory, thereby indicating that the translation was appropriate, i.e., Japanese students interacted with the test questions in a manner similar to that of

American students. However, one limitation of the study was that the proportions of item responses of American and Japanese students could not be compared because of the lack of access to the raw data of American students. With access to a large amount of data from American and Japanese students, the present study extends this comparison of student views by comparing all item answer choices and exploring their selection across different levels of student performance. We use item response curves (IRCs) to explore each response to each item across all levels of proficiency. The study is a first look at IRCs for FMCE items of Japanese and American students. Based on our findings, we propose that IRCs can be used as evidence when constructing validity arguments for translated assessments or using existing assessments with new populations.

When one evaluates non-American data taken from a translated American assessment, there are several reasons why one might expect to find differences in item functioning:

(1) The *translation* process involves the translator's interpretation of content, context, and writing style, which could change the way students approach and interpret the item.
(2) The *language* itself, regardless of translation, can cause differences in how ideas are understood and perceived.
(3) Physics assessment items tend to be grounded in real-world *contexts*, but those contexts may not be universal. For example, the American FMCE uses toy cars, coin flipping, and a collision between a truck and a car.
(4) Different *educational systems* may cover topics in varying levels of depth or may use different pedagogical approaches to teaching, leaving students in different nations with different naïve concepts.

Given the number of possible reasons that students across cultures might interact differently with assessment items, we hypothesize that the IRCs of American and Japanese students would differ in some meaningful ways.

IRCs are plots of proportions of responses as a function of the students' total raw score, which allow researchers to evaluate how each item functions across the range of ability levels and for all possible responses. Morris *et al.* [5,6] brought IRCs from the field of item response theory (IRT) to physics education research. They showed how the plots could be used as tools to describe qualitatively how items on physics assessments are structured. Figure 1 shows an example IRC, from item 8 of the FMCE.

The shape and position of the plots can be used as rough estimates of item difficulty and discrimination. The *difficulty* of an item is lower when the correct answer curve is shifted horizontally to the left. When the curve is closer to the origin, students at lower proficiency levels are more likely to answer the item correctly. Item *discrimination* can be estimated by the shape of the correct answer curve, giving a sense of the item's ability to distinguish between low- and high-scoring students. More discriminating items

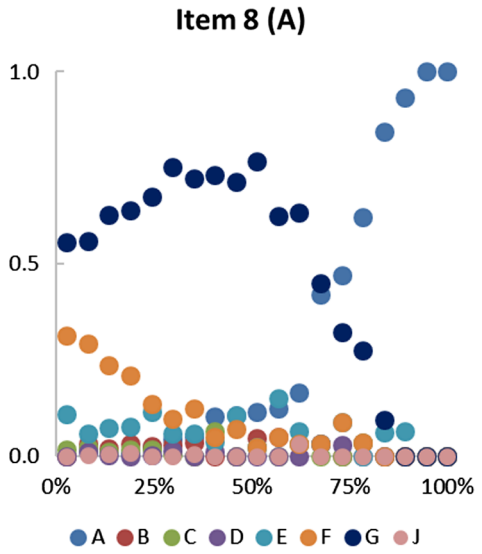

FIG. 1. IRCs of item 8 of the FMCE American data. A, American students.

have steeper slopes at the midpoint of the curve, where the proportion of correct responses is one half. A steep slope indicates that low-ability students are unlikely to answer correctly and that high-ability students are likely to answer correctly, thus discriminating between the two.

When the correct response curve differs in difficulty or discrimination across demographic groups, the result is called differential item functioning (DIF). This is a separate concept from the idea of item bias, where members of one demographic group are more likely to answer the item correctly. DIF analysis answers the question of whether students of different demographic groups have the same probability of answering the item correctly *given that they are equal in terms of overall performance*. Psychometricians working on large-scale tests spend much of their time performing DIF analyses to ensure that tests will generate meaningful, comparable scores across subpopulations. They use very sensitive statistical tests that detect small differences in difficulty or discrimination. Figure 2 illustrates the kinds of small differences in (a) difficulty and (b) discrimination that must be identified and accounted for when calibrating large-scale assessments.

Because we want to examine item response structure across two substantially different samples, it is not necessary to perform high-precision tests for DIF in this study. If we use an analogy to physics lab activities, then conventional DIF tests are high-precision instruments that show whether two objects are the same size; however, in our case, we are trying to determine whether the objects are even the same shape. Given that the students were raised in different cultures, taught in different school systems, and took the test in two different languages, we would not expect the IRCs to be identical. The question that remains, however, is whether the items are similar at all. Previous research

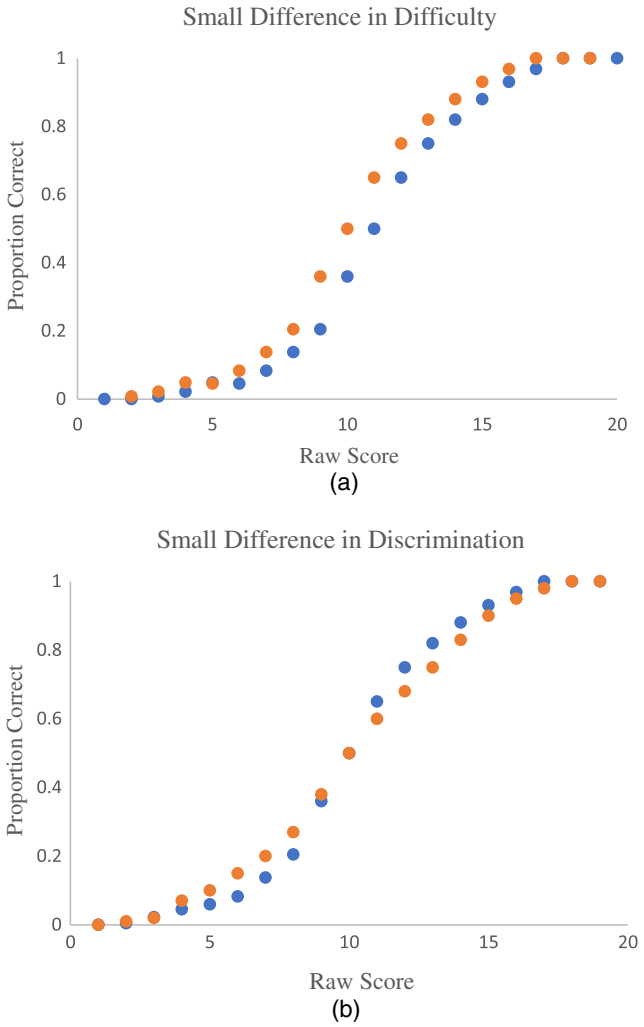

FIG. 2. Plots of the correct IRC of a single item, using groups for which the item has a different (a) difficulty and (b) discrimination. Red dots and blue dots represent the groups in comparison.

showed that the most common incorrect responses were the same across groups [4], providing one piece of information about item functioning. However, that analysis does not provide a complete picture of how students across proficiency levels interact with the test items.

On a similar note, conventional DIF analyses are designed for dichotomous items or those that have been dichotomized into "correct" and "incorrect." Conventional psychometric approaches tend to break down under conditions where multiple distractors are meaningful [7]. Sadler [7] showed that the correct answer curves for concept inventory items are not always the clean, monotonically increasing curves assumed in the IRT framework.

Given these restraints, our analysis of the IRCs for the FMCE is a qualitative one. Following the suggestions of Morris *et al.* [5], we examine the IRC plots visually and make inferences about item functioning. In making comparisons between Japanese and American samples, we look for items where the shapes of the curves are noticeably different. This qualitative approach has the advantage of looking for DIF across all responses, not just the correct curves.

IRCs are essentially representations of student behavior, charting student responses to the stimuli of the items. If the IRCs differ noticeably between the samples, we can infer that there was some difference that caused different behaviors. The cause might be educational, contextual, linguistic, or cultural, i.e., students raised in a particular culture may have fundamentally different ideas about force and motion. Interacting with the item in the same way does not necessarily prove that the underlying issues are the same for the two groups. It is possible for two different items or populations to generate the same IRCs. However, the probability of doing so randomly is extremely small. Although similar curves are not proof of identical thought processes, we consider similar curves as supporting evidence in the process of validation in this study. When validating a new version of an instrument for a new population, we believe that IRCs are useful for highlighting problem areas.

If successful, the valid translation of materials from one nation to another could benefit both tremendously. If the process shows that Japanese and American students enter physics classrooms with similar incorrect ideas—and that valid translations are possible—then educators can begin to share their resources and expertise. Common ground between American and Japanese students' learning of physics would allow the sharing of research, instructional strategies, instructional materials, and assessments. However, if item contexts prove to be specific to one culture, or if certain misconceptions appear in only one population, then research must be applied to samples individually to obtain country-specific results.

II. METHODS

In this section, we describe the FMCE, the Japanese translation of the FMCE, and the sampled students and show how we generated the IRCs.

A. The FMCE

The first major concept inventory was the FCI, which used conceptual items with no required computation to measure the extent to which students choose common incorrect ideas about forces [1]. Multiple-choice items included attractive distractors, i.e., responses that were likely to be chosen by students with misconceptions. The FCI had a large impact on physics education because it demonstrated that students did not understand fundamental physics concepts—even after instruction and even those students who could solve difficult physics problems [8]. The instrument became popular among high school and college instructors and is still widely used, with translated versions in 27 languages as of September 2017 [9].

Following the impact and popularity of the FCI, researchers developed concept inventories to assess student understanding in many content areas (e.g., the Astronomy and

Space Science Concept Inventory [10]). Thornton and Sokoloff [2] designed the FMCE to cover similar content to the FCI but with an emphasis on varied representations of physical phenomena. The instrument was more closely aligned to their RealTime Physics curriculum, which used demonstrations and sensors to help students understand physics concepts by engaging them with interactive data and graphs [11].

The FMCE can be analyzed in terms of seven groups of items, commonly referred to as item clusters. Each cluster uses a single-item stem, which provides context for the items, and a common group of responses. The clusters target specific misunderstandings about physics that were identified by previous qualitative research. Table I provides a brief description of each item cluster.

When we compare the performances of Japanese and American students on the FMCE, we must consider the translation of the survey from English into Japanese, a substantially different language. Sentence translation involves interpretation of both word choice and writing style, which can result in subtle differences in readers' comprehension of the content. In addition, the difference in grammar requires some adjustment so that translated sentences seem natural to the reader. Ishimoto, Thornton, and Sokoloff [4] used a colloquial style of Japanese language, aiming to match the casual writing style of the English FMCE. The result is somewhat different from Japanese textbooks and exams, which use a more formal, concise writing style. The goal of the casual style, in both languages, is to obtain "natural" responses rather than to send students into "school mode," where students may answer the way they think the teacher wants rather than to respond honestly. The translation process, from one colloquial style to another, involves interpretation and guesswork about reader interpretation. Ishimoto *et al.* [4] presented results showing that the Japanese students' common incorrect responses were similar to those of American students, as reported by Smith and Wittmann [12]. Comparing Japanese and American IRCs in this study is a further step in evaluating the validity of this translation.

B. Study samples and student background

Japanese schools have much less cultural and racial diversity compared with American schools. Most students grow up in urban settings and use public transportation. They do not have as much experience driving automobiles or snow sledding, activities that are relevant to the clusters of the FMCE. They enjoy American culture through electronic media such as movies, TV news, and the Internet. For

TABLE I. Description of items in clusters.

Cluster	Context	Task	Common incorrect responses
Force sled	A person is moving a frictionless sled across an icy surface.	Items describe the motion of the sled. Students are asked to select the force that matches the motion.	Students choose a force that is identical to the velocity, not change in velocity.
Reverse direction	A toy car travels up a ramp, slows, and comes back down the ramp. A coin is tossed into the air, slows, and falls back down.	Students asked to select the net force (acceleration) acting on the object at each point in the trajectory.	Students choose a force identical to the velocity rather than the single downward force of gravity on the coin or the net down-ramp force on the car.
Force graphs	A frictionless toy car moves to the left or right on a frictionless track.	Items describe different motions of the toy car. Students are asked to select the matching force vs time graph.	Students choose the graph of the velocity rather than the graph of the force.
Acceleration graphs	A frictionless toy car moves to the left or right on a frictionless track.	Items describe different motions of the toy car. Students are asked to select the matching acceleration vs time graph.	Students choose the graph of the velocity rather than the graph of the acceleration.
Newton III	Two vehicles, cars or trucks, collide with each other or push each other.	Each item describes a collision or push under different circumstances. Students are asked to identify which vehicle exerts the greater force.	Students choose the larger or faster vehicle rather than equal forces.
Velocity graphs	A frictionless toy car moves to the left or right on a frictionless track.	Items describe different motions of the toy car. Students are asked to select the matching velocity vs time graph.	Students rarely have difficulty with these items.
Energy	A child rides a sled down a frictionless hill.	Each item describes a hill that is steeper or higher. Students are asked to identify whether the end velocity or kinetic energy is greater.	Students vary in the way they approach these items.

example, Japanese students know about coin tossing through the media, but they do not practice it themselves. Japanese middle schools have a mandatory science curriculum, which teaches Newton's third law and mechanical energy conservation qualitatively in simple settings [13]. Most high schools did not offer mandatory physics courses during the period when the students who participated in this study were in high school, and a large majority of these students took biology and/or chemistry instead. About half of Japanese high school graduates attend college or university.

The American FMCE data were collected between 1999 and 2002, so all students in the sample had completed their elementary and secondary education before the No Child Left Behind Act passed into law and many years before the Next Generation Science Standards were written. As such, American science curricula varied widely across states and districts. Although it seems likely that a majority of American students were exposed to Newtonian motion during middle or high school, assessment results show that most students enter college with non-Newtonian views on force and motion. As in Japan, few American schools require that students take physics, and most students take earth science, biology, and chemistry instead [14].

The Japanese data sample came from first-year engineering students from various prefectures in Japan. All students were enrolled in an introductory mechanics course between 2004 and 2014. Half of the sample had taken an algebra-based high school physics course.

The American data sample came from six institutions of different types, including a state university, a community college, and a military academy. All students were enrolled in an introductory college-level mechanics course. An unknown proportion of students in the sample had taken high school physics courses, likely varying across institutions. The American FMCE data sample was collected as a part of the evaluation of the national dissemination of the RealTime Physics curriculum.

C. Data collection, cleaning, and scoring

Both samples of FMCE results were collected by instructors or researchers at the beginning of a physics course, before students encountered any relevant content. The data sets were cleaned using the same set of criteria. Students were included only if they answered at least one question on each of the item clusters. The majority of students eliminated from the data sets were those who did not answer questions on the energy cluster, probably because of time constraints on testing. A very small number of students appeared in the data sample twice, in different semesters, likely because they chose to retake the class. In each case, we selected only the first set of FMCE results. After cleaning, the data sets contained less than 0.1% missing responses, which were marked as incorrect.

IRCs require an estimate of each student's ability level, which we call a proficiency score. Thornton *et al.* [15] said that "the FMCE was not originally designed to have results analyzed with a single-number score," but they also explained that it is necessary to generate such scores for some research purposes. We used a scoring template, which

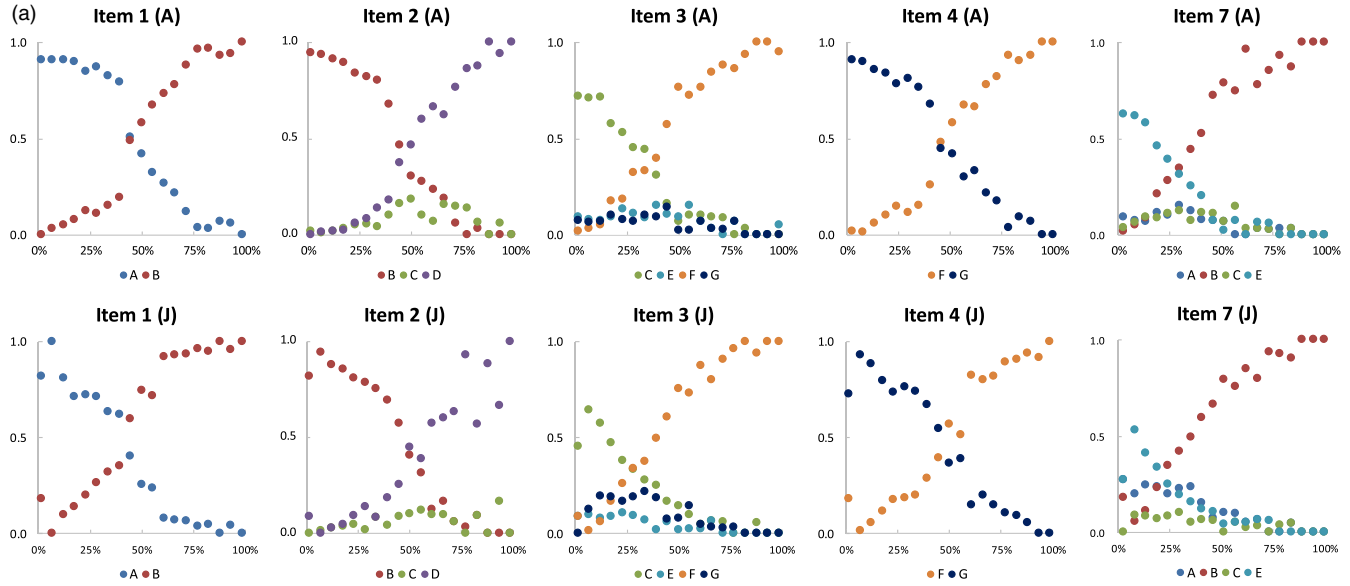

FIG. 3. (a). IRCs for the force sled item cluster. A, American students; J, Japanese students. (b). IRCs for the reverse direction item cluster. A, American students; J, Japanese students. (c). IRCs of the force graphs item cluster. A, American students; J, Japanese students. (d). IRCs of the acceleration graphs item cluster. A, American students; J, Japanese students. (e). IRCs of the Newton III item cluster. A, American students; J, Japanese students. (f). IRCs of the velocity graphs item cluster. A, American students; J, Japanese students. (g). IRCs of the Energy item cluster. A, American students; J, Japanese students. (h). IRCs of unscored items. A, American students; J, Japanese students.

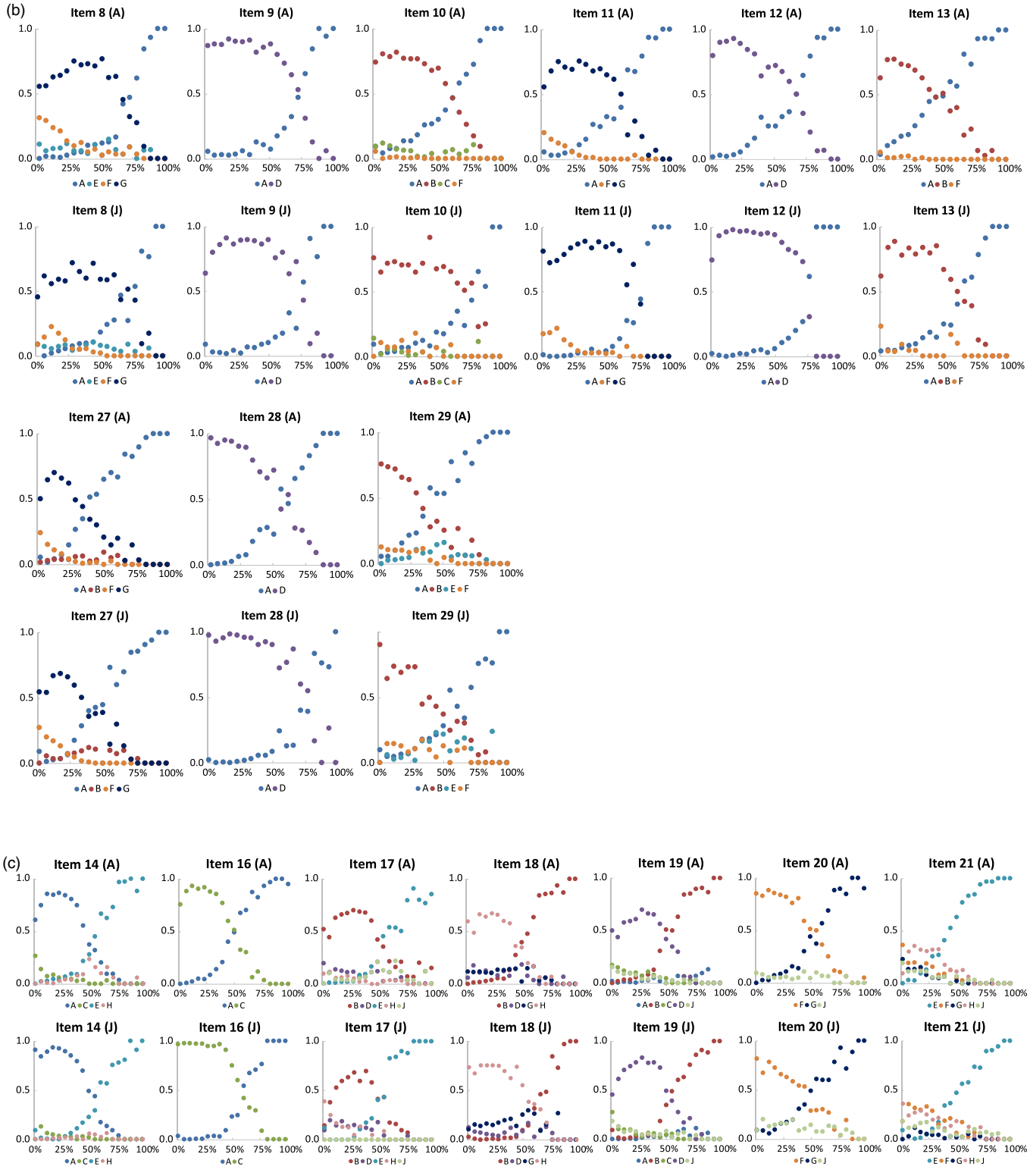

FIG. 3. *Continued*

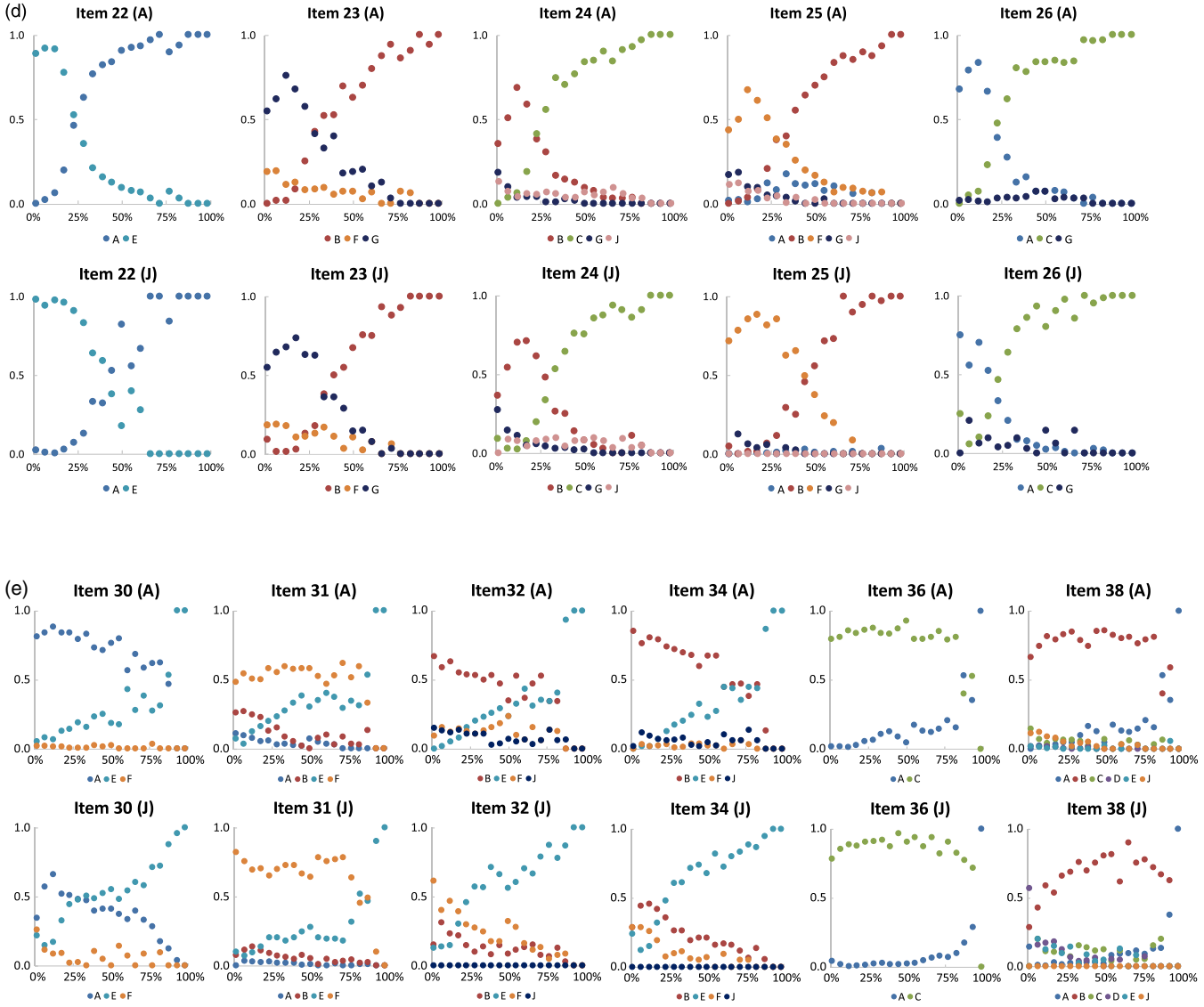

FIG. 3. *Continued*

we call the single-number score (SNS), based on the recommendations of Thornton and Sokoloff [2]. First, seven problematic items (5, 6, 15, 33, 35, 37, and 39) are excluded from scoring. These are mostly false positives, items that students tend to answer correctly but not always for the correct reason. Next, each set of three reverse direction items is scored as a group such that each set receives two points if, and only if, all three items are answered correctly. For example, a student would receive two points for correctly identifying the force on a coin as it moves upward, as it changes direction, and as it moves downward. If any item is answered incorrectly, the student receives zero points for the set. Finally, researchers consider the Energy cluster an optional part of FMCE scoring because its content is somewhat different from the rest of the instrument. We decided to include the energy cluster in the SNS score, allowing us to generate IRCs for the energy items. In the end, our scoring template used 40 of the 47 items and had a maximum possible SNS of 37.

D. Generating IRCs

IRCs are plots of the proportions of each answer choice of a single item as a function of proficiency [5,6]. This representation allows a simultaneous comparison of both correct and incorrect responses and an exploration of responses at each level of proficiency. In a sense, IRCs disentangle response choices from proficiency. Thus, although the Japanese students outperformed American students (with a mean difference of two points), the curves can be used to compare the groups because they

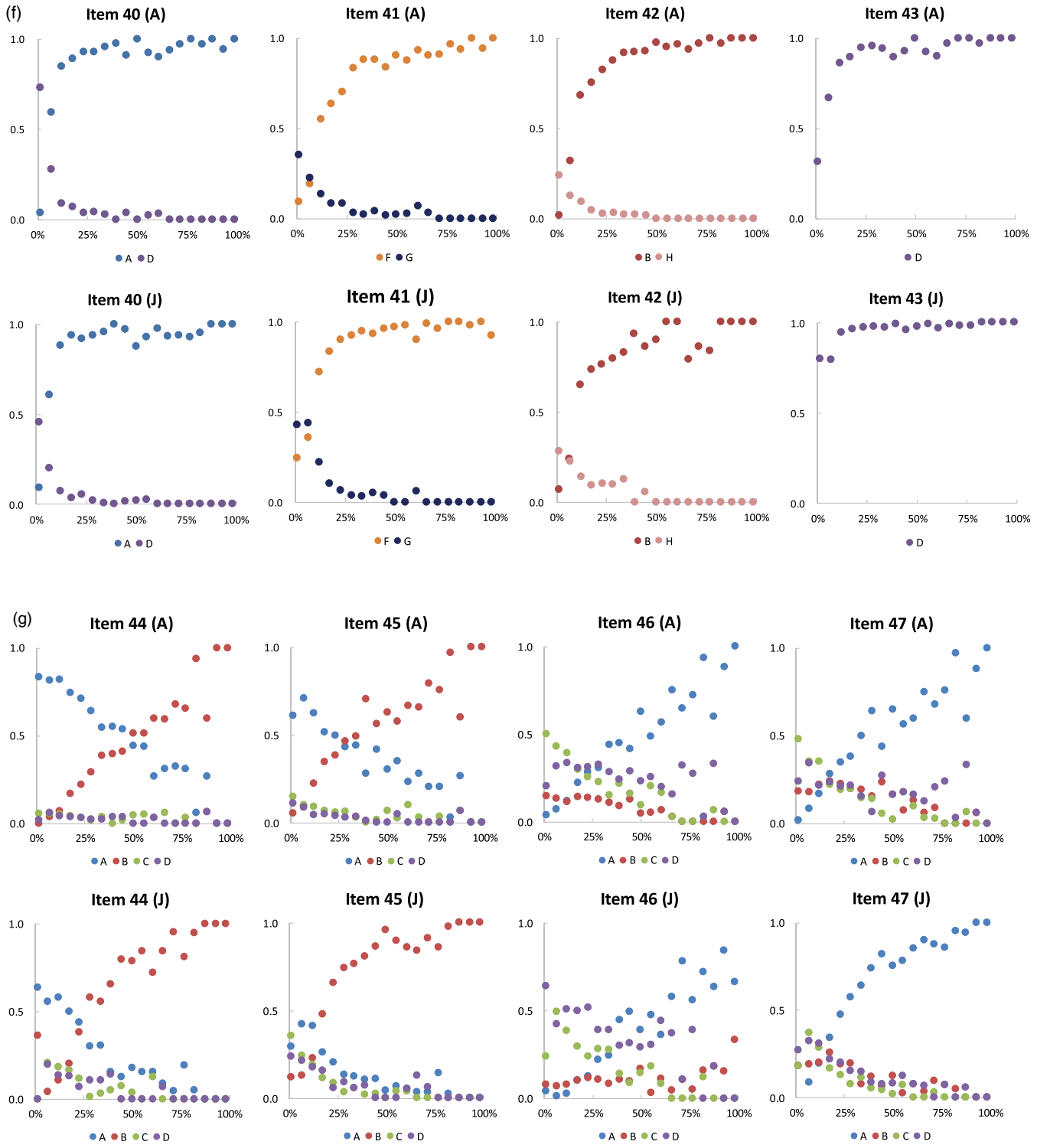

FIG. 3. *Continued*

cover students at all levels of performance. IRCs are closely related to IRT, which uses simultaneous estimation of student proficiency and item characteristics. IRT has the advantage of examining proficiency level and item difficulty on a common scale, generating parameters that are sample independent and item independent [16,17]. That is, if there is no DIF, item characteristics will be the same across samples, and sample characteristics will be the same across instruments.

Raw scores, such as the SNS, do not satisfy the assumptions of IRT but are used as proxy variables for proficiency. Morris *et al.* [6] used the total score of the FCI as a proxy for a

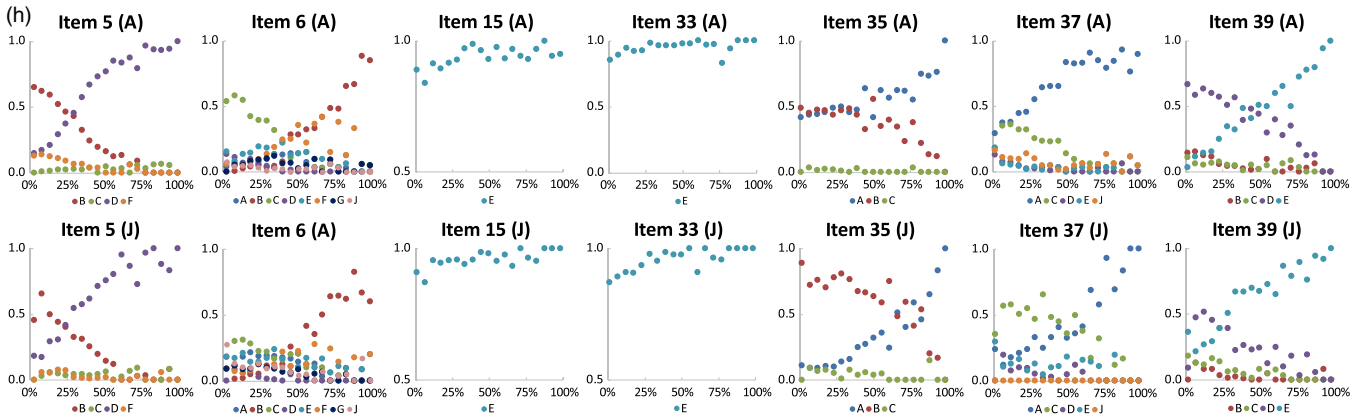

FIG. 3. *Continued*

proficiency, justifying their decision by the very strong correlation ($r = 0.893$) between IRT proficiency estimates and raw scores. In the present study, we followed their example and used the SNS as a proxy for proficiency. The correlation between two-parameter IRT scores and raw scores in the American FMCE sample was greater than 0.95 [18].

IRCs were generated by calculating the proportion of students giving each response at each SNS level and plotting the curves on the same graph. We smoothed the IRCs by creating two-point bins for the SNS; thus, students with scores 0–1, 2–3, 3–4, and so on are presented as single points. Bins with more students have smaller standard errors because uncertainty in proportions tends to decrease with larger sample sizes. Given this, the American IRCs tend to be more precise in terms of sampling and inference because the American sample was larger than the Japanese sample. In addition, the most precise estimation of each curve occurs between SNS scores of 4 and 12, i.e., between 10% and 32%, because this was the range with the most student scores. IRC points at the high end should be interpreted as though they have larger error bars because few students achieve very high scores prior to instruction. The Japanese sample included only nine students with scores of 0 or 1, a number small enough that the data points would not be interpretable; therefore, this bin was excluded from each plot in Fig. 3(a)–3(h). Lastly, we did not include curves for responses that never peaked higher than 10% of responses. This decision was made to declutter the plots, thereby allowing readers to focus on the relevant responses.

TABLE II. Statistics of the two samples.

	American	Japanese
Sample size	2348	1531
Mean score	9.05	11.22
Standard deviation	7.39	7.56
Median score	7	9

III. RESULTS

Descriptive statistics of FMCE pretest scores for both samples, presented in Table II, show that the mean score for Japanese students was two points higher than the mean American score. A two-tailed t test indicates that these averages were significantly different ($t = 8.8$, $p < 0.005$). The score distributions of the two samples are similarly shaped (Fig. 4); however, very few Japanese students had extremely low scores. Although the American sample size was larger than the Japanese sample size, both were sufficient for generating IRCs.

In this section, we present and interpret the IRCs. First, we examine the overall results of the analysis. Next, we compare the IRCs generated for each item cluster. Then, we consider the symmetrical items, those that ask identical questions except with opposite directions of motion. Finally, we look at some specific items, such as the items excluded from the SNS and an item that is common to both the FMCE and the FCI.

A. Overview

The IRCs in Fig. 3(a)–3(h) are remarkably similar across the two samples. Students with low SNSs tended to select

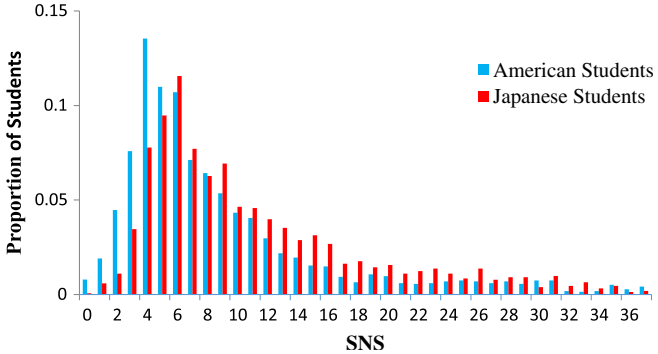

FIG. 4. Proportions of American students (blue) and Japanese students (red).

the same distractors across cultures, which is consistent with previous research [4]. There were many items that showed the same overall structure but were noticeably more difficult for one sample than the other, as indicated by the horizontal position of the curves. We did identify a few items that are substantively different in structure, which will be described in subsequent sections. The students tended to interact with the items similarly, selecting the same distractor responses across ability levels for nearly all items. This similarity provides some evidence that Japanese and American students have similar views when they enter college physics courses and that the translation of the assessment was appropriate.

B. Item cluster comparisons

1. Force sled cluster (items 1–4, 7)

The force sled questions ask students to select the net force that would cause various motions of a frictionless sled. The IRCs of the two samples appear to be very similar [Fig. 3(a)]. In fact, all the force sled IRCs appear to be similar to each other, taking into account the approximate form of dichotomous items wherein students select only from two of the available options. Most preinstruction students answer as if they believe that the force on an object is proportional to its velocity ($F \propto v$), selecting the responses ABCGE. Some students, usually those with high SNSs, select the correct response set BDFFB. Our results, shown in items 1–4 and 7 in Fig. 3(a), are consistent with those of Smith and Wittmann [12].

The most common incorrect answers in items 1 and 4 say that increasing speed to the left or to the right would be caused by an increasing force to the left or to the right, respectively. On item 2, which asks about a sled moving at a constant velocity, most students selected a constant force, although some midlevel proficiency students selected a decreasing force. For the decreasing speed items 3 and 7, the common incorrect response was that the force would be in the direction of motion but decreasing over time. For these items, Japanese students were more likely to select distractors G and A, respectively, to indicate that the force would oppose the direction of motion and be increasing. The response pattern ABGGA, which appears to have been selected by Japanese students more than by American students, indicates that the students understand the directionality of forces but not the magnitude.

2. Reverse direction cluster (items 8–10, 11–13, and 27–29)

The reverse direction cluster has three sets of three items, each asking about the net force on (or acceleration of) an object as it moves up, reaches the top of its trajectory, and falls back down. The most common model was again the $F \propto v$ model, where students give answer GDB to express forces that are proportional to the motion rather than a change in motion. Smith and Wittmann [12] also identified a group of students who answered questions about the upward portion of motion as "the force is upward and increasing." We can see this response F in the IRCs for items 8, 11, and 27 [Fig. 3(b)]. Items 9, 12, and 28 ask about the force on the object at the moment that it changes direction. The items appear to be nearly dichotomous, where students answered either that the net force was zero (incorrect) or that it was downward and constant (correct). Item 29 asks about the acceleration of the coin as it falls, and the IRC shows that students selected from a number of different distractors, even more than for the analogous items 10 and 13. The IRCs appear similar overall, although some of the items appear noticeably easier for the American students. The first three items, which use the context of the toy car on the ramp, are more similar across groups than the second set of items, which uses the context of the coin toss. The difference between these two sets of items may indicate that it is the coin toss, indicating that the context itself might be the reason that the items were more difficult for the Japanese students.

3. Force graphs cluster (items 14, 16–21)

The force graphs questions ask students to select the graph of force vs time that matches the described motion of a toy car. The most common incorrect student view is the same as with the force sled cluster $F \propto v$. The resulting incorrect response pattern is ACBHDF*, where the seventh symbol, the asterisk, indicates a variety of responses, which is reflected in the IRCs of Fig. 3(c). The IRCs of the two groups are essentially the same, although items 16 and 18 appear to be slightly easier for the American students.

4. Acceleration graphs cluster (items 22–26)

Items 22–26 ask students to select the acceleration vs time plot that matches the description of a toy car in motion. The most common view is confusion between acceleration and velocity, $a \propto v$. The IRCs of the two groups are very similar, although item 22 appears to be easier for the American students [Fig. 3(d)]. The correct response curves of the acceleration graphs are shifted far to the left, where even students in the middle of the SNS range have a high probability of answering correctly. This indicates that these items are easier than others on the FMCE.

5. Newton III cluster (items 30–32, 34, 36, 38)

The Newton III items ask about the forces on two objects in situations where they collide or push each other. The two incorrect views that students tend to display are mass dependence and action dependence [4]. The former group thinks that a *larger* object exerts more force, whereas the latter group thinks that a *faster* object exerts more force. Students might have both incorrect ideas (mass dependence

and action dependence) and may or may not see the two as compensatory. The correct view is that colliding, pushing, and pulling objects always exert the same force on one another. The first four questions of the Newton III cluster ask students to compare the forces in the collision of a car and a heavier truck (items 30–32) and those in the collision of a car and a truck with equal mass (item 34) [Fig. 3(e)].

These items produced very different IRCs across the two samples. The majority of American students appeared to use a mass-dependent model on item 30 and an action-dependent model on items 32 and 34. The correct response curves of these items were shifted far to the right, such that only the most proficient American students answered these four items correctly. By contrast, the correct answer curves for the Japanese students hover near 50% for a wide range of SNSs. These curves have shallow slopes, indicating poor discrimination, and many more Japanese students were likely to answer these four items correctly. Item 31 appears to be an exception, however. It asks about the forces between a fast-moving car and a slow-moving car, but heavier, truck. Smith and Wittmann [12] identified item 31 as a possible false positive for American students. Some students used both mass-dependent and action-dependent thinking, and they assumed that the two compensate and cancel out each other. As a result, they reported that the two objects would exert the same force. It appears that the Japanese students did not make the same assumption because a smaller proportion answered the item correctly. The majority of Japanese students chose answer F ("Not enough information is given to pick one of the answers"). It seems possible that Japanese students used ideas that had mass dependence and action dependence but did not make the assumption that the two would be equally influential. The differences in this cluster may account for much of the mean score difference (2 points) between the Japanese and American scores (see Table II).

Items 36 and 38 ask students to consider forces between a car and a broken-down truck while the car is pushing the broken-down truck. The IRCs of the two groups agree and indicate that most students viewed them as action-dependent phenomena. These two items are so difficult that they provide almost no information about the population of preinstruction students, providing information only about the highest-scoring individuals. We speculate about possible causes of the differences in Sec. V.

6. Velocity graphs cluster (items 40–43)

The velocity graphs cluster asks students to select the velocity vs time plot that represents the motion of a toy car. The IRCs of the two groups are very similar, where all students, except for very low-scoring students, give correct responses [Fig. 3(f)]. The incorrect responses among low-scoring students indicate that the velocity is proportional to the position.

7. Energy cluster (items 44–47)

These questions ask students to predict the speed and kinetic energy of a sled after sliding down a hill. The IRCs in Fig. 3(g) indicate that the items are relatively easy and are easier for the Japanese students than for the American students. All four items are poorly discriminating, as indicated by the shallow slopes of the correct answer curves. Incorrect responses on items 44 and 45 reflect the common incorrect view that steeper hills cause greater speed. Items 46 and 47 ask students to compare the speed and kinetic energy, respectively, of the sled after sliding down a higher but less steep hill. The IRCs of item 46 show that students selected a variety of responses, possibly because students were torn between height dependence and steepness dependence.

The correct answer curves for the Japanese students locate higher than those for the American students in the midlevel SNS regions on items 44, 45, and 47. It appears that the Energy items partly account for the higher SNS averages of the Japanese students. We will speculate about possible causes of the differences in Sec. IV.

C. Symmetrical item pairs

Six pairs of FMCE items ask the same question but with motion in opposite directions (Table III). Items 1 and 4 in the force sled cluster ask students to identify the force on a constantly accelerating sled to the right and to the left, respectively. The IRCs in Fig. 3(a)–3(h) are almost identical. In the force graphs cluster, items 14 and 17 and items 16 and 19 have the same kind of symmetry. In this case, the IRCs reveal that the distractor was much more attractive to

TABLE III. Description of items of symmetrical pairs.

Item pair	Cluster	Question	IRC comparison
1 and 4	Force sled	Increasing speed to the right and left	Very similar
3 and 7	Force sled	Decreasing speed to the right and left	Very similar
14 and 17	Force graphs	Constant speed to the right and left	Distractor less effective for item 17
16 and 19	Force graphs	Increasing speed to the right and left	Distractor less effective for item 19
22 and 25	Acceleration graphs	Increasing speed to the right and left	Distractor less effective for item 25
24 and 26	Acceleration graphs	Constant speed to the right and left	Very similar

low-scoring students when motion was to the right. The same is true for items 22 and 25 in the acceleration graphs cluster. It is possible that the left-moving object in graphs confuses the low-scoring students by adding to the cognitive load. It is also possible that some students have specific assumptions about left motion being different from right motion. The trends are the same in the Japanese and American samples.

D. Unscored items (items 5, 15, 33, 35, 37, 39)

Thornton *et al.* [15] said that "items 5, 15, 33, 35, 37, and 39 [whose IRCs are shown in Fig. 3(h)] are frequently answered 'expertly' by students even before they are consistently Newtonian thinkers" and that item 6 in Fig. 3(h) "is sometimes answered incorrectly, even by experts." In this section, we focus on the items that are excluded from the SNS.

Item 5 is considered to be a confusing item that receives many false-positive results [15]. The question is worded as follows: "The sled was started from rest and pushed until it reached a steady velocity toward the right. Which force would keep the sled moving at this velocity?" A physicist would consider this question to be identical to the "constant velocity" question in item 2; however, the "getting up to speed" aspect of the item makes it difficult for new physics students. It can entangle the item with the naïve "impetus" reasoning that motion is given to an item and dies away over time. The IRCs show that item 5 was easier than item 2 for both the Japanese and American students, indicating that this contextual effect alters student responses in both Japanese and English.

Item 6 was intended to probe students' belief that the net force is in the direction of the acceleration. It asks students to identify the force on the sled as it slows down and accelerates to the right. The wording of the item confused American students; therefore, Thornton and Sokoloff [2] considered the item to be a false negative. The IRCs in Fig. 3(a)–3(h) confirm that even students with high SNSs in both samples do not always answer correctly. We speculate that the confusion is likely from the context rather than from the translation because the original wording of the question and answer choices is simple and straightforward to translate. It is interesting to note that the low-scoring American students were more likely to choose a specific distractor (C), whereas the Japanese students selected a variety of distractors.

Item 15 asks about the net force acting on an object standing still. Item 33 asks about the forces involved in a collision of two identical objects with the same speed. The two items can be answered correctly without a Newtonian understanding of forces. The two items were included to probe students' reading comprehension. As expected, the IRCs of items 15 and 33 show that almost all students answered these items correctly.

Thornton *et al.* [15] excluded items 35 and 37 from the SNS items because they are somewhat misleading. They ask students about the forces between two vehicles as one pushes the other but not enough to cause any acceleration. The items tend to be false positives for American students. The IRCs of the two items show a larger proportion of correct responses than for similar items (items 36 and 38), thereby confirming the observation by Thornton *et al.* [15]. Upon closer examination of the IRCs, we find that they function more like normal items for the Japanese students, with lower-scoring students more likely to select a specific distractor. We suspect the difference may be caused by cultural or translation effects, which we describe in Sec. IV.

The pushing items of the Newton III cluster (items 35–38) ask about the forces when the vehicles are not moving, are increasing speed, are at a cruising speed, and are decreasing speed. The context of these items is unusual, and the situation was particularly difficult to translate into Japanese. The IRCs of items 35 and 37 are very different across the Japanese and American samples. The IRC of item 35 for the Japanese sample seems to act as a typical item, with one strong distractor that decreases as the proportion of correct answers increases. The IRC of item 37 for the Japanese sample is messier, with low-scoring students selecting from a variety of responses.

Item 39 asks about forces between two people pushing off each other in rolling chairs. It is identical to item 28 of the FCI, and the translations of the item in the FMCE [4] and the FCI [18] are the same. The IRC plot for FCI item 28 is given in an article by Morris *et al.* [6], a study that used an American student sample (see Fig. 1 of Ref. [6]). The IRCs of these two identical items, on different tests, appear nearly identical for the American samples. The IRC for Japanese students in Fig. 3(h) is nearly identical to the IRC for a sample of more than 5000 Japanese high school and college students from a previous study using the FCI [19]. The similarity of the IRCs across instruments is an indicator of sample independence within the nation but, at the same time, explores differences across national samples. The IRCs presented in Fig. 3(h) confirm that many low-scoring students answer the item correctly. In terms of IRT, the shallow slope of the correct response curve indicates that the item discriminates very poorly.

IV. SUMMARY

The IRCs of two different samples taking the FMCE in two different languages were surprisingly similar. Many items showed noticeable differences in item functioning, specifically in the difficulty (horizontal positioning of the correct answer curve). When the items are visibly different, tests of DIF will be statistically significant, as they are very sensitive. However, the overall structure of the items was similar—students in both samples selected the same responses. The small differences in functioning would cause problems if the FMCE were to be used as a single, large-scale assessment for both American and Japanese college students. Nonetheless, the curves are remarkably

similar given that the two versions of the test are not in the same language. A few of the items were markedly different and need to be examined more closely to establish the validity of the Japanese translation.

There are a set of notable, specific findings from the IRC comparisons of the American FMCE and the Japanese translation of the FMCE:

(1) Four common characteristics
   (a) The majority of items operated as if they were dichotomous, with most students selecting either the common incorrect response or the correct response.
   (b) The items that Thornton *et al.* [15] omitted from their scoring rubric appeared to be problematic in terms of IRCs.
   (c) Right-direction items were easier for both groups of students than left-direction items in the graph representation but not in the verbal representation.
   (d) The velocity graphs cluster was not discriminating and did very little to distinguish between students of different proficiency levels.
(2) Two main differences
   (a) The Newton III and energy items were easier for Japanese students.
   (b) Coin toss items were easier for American students.

V. DISCUSSION

The use of IRCs with FMCE data allows us to compare how students in different populations interact with individual items. These behavioral results give us some evidence about how students think about force and motion and how they interpret each item. The overall similarity of the IRCs indicates that the students tended to have the same naïve ideas about force and motion and that they interpreted the items similarly. This similarity is surprising, given the number of reasons that the populations might approach items differently. The translation appears to have little impact on how students interact with the assessment, thus providing some evidence that the translation is valid.

In this section, we present our speculations about the causes of the differences observed across the IRCs. We believe these speculations provide a starting point for future studies.

A. Educational background ("action-reaction law" and "energy conservation")

The two greatest differences in the IRCs between the two samples appeared in the collision situation of the Newton III cluster [Fig. 3(e)] and in the energy cluster [Fig. 3(g)]. We suspect that the higher proportions of correct responses on items 30, 32, 34, 44, 45, and 47 by the Japanese students may represent their rote memory from previous physics education. If the speculation is correct, the average SNSs of the two nations shown in Table II become about the same.

The collision items in the Newton III cluster, items 30–34, ask about forces on two objects in collision. On items 30, 32, and 34, more Japanese students (in the wide range of scores) chose correct responses, but they failed to answer item 31 correctly, which requires a deeper understanding of Newton's third law [Fig. 3(e)]. Because Newton's third law is introduced qualitatively as the action-reaction law in Japanese middle school, we suspect that they answered items 30, 32, and 34 correctly by rote memory; however, some students struggled with item 31—perhaps because their rote memory conflicted with their intuition, resulting in more low- and middle-scoring Japanese students choosing answer F ("Not enough information is given to pick one of the answers").

Energy cluster items, items 44–47 [Fig. 3(g)], ask students to predict the speed and kinetic energy of a sled after sliding down a hill. Incorrect responses on items 44 and 45 reflect the common incorrect view that steeper hills cause greater speed. Items 46 and 47 ask students to compare the speed and kinetic energy of the sled after sliding down a higher but less steep hill. The Japanese students with midlevel scores answered correctly on items 44, 45, and 47 but failed on item 46. The IRCs of item 46 show that the students selected a variety of responses, possibly because they were torn between height dependence and steepness dependence. Because mechanical energy conservation is also introduced qualitatively in Japanese middle school, we infer they answered items 44, 45, and 47 correctly by rote memory. However, they were less successful in answering item 46 correctly, perhaps because it portrayed a more sophisticated situation.

B. Personal experiences ("coin toss")

The proportions of correct responses by the American students were higher on items 11 and 13 [Fig. 3(b)], which ask about the force involved in coin flipping, than those on analogous items 8 and 10, which ask about the force on a toy car on an inclined plane. By contrast, items 11 and 13, which ask about the net force on a coin during a coin toss, were more difficult for the Japanese students compared with the analogous items 8 and 10, which ask about the net force on a toy car on an inclined plane [Fig. 3(b)]. Although Japanese students seldom practice coin tossing, they know what it is from the media. The American students' familiarity with this practice may have influenced their responses.

C. Translation (Newton's third law in a pushing situation)

The pushing items of the Newton III cluster, items 35–38 [Fig. 3(e)], ask about the forces involved when the vehicles are not moving, are increasing speed, are at a cruising speed, and are decreasing speed. Thornton *et al.* [15] excluded items 35 and 37 from the SNS items because

they were somewhat misleading. These items ask students about the forces between two vehicles as one pushes the other, but the forces are not great enough to cause any acceleration. The items tend to be false positives for American students. The IRCs of the two items show a larger proportion of correct responses than for two similar items, items 36 and 38, thereby confirming the observation made by Thornton *et al.* [15]. The IRCs of items 35 and 37 are very different across the Japanese and American samples. The IRCs of item 35 for the Japanese students seem to act as a typical item, with one strong distractor that decreases as the proportion of correct answers increases. The IRCs of item 37 for the Japanese students are messier, with low-scoring students selecting from a variety of responses.

It is possible that the difference in performance on these items is because of translation issues, particularly with the phrase "not hard enough" (item 35) and the term "cruising speed" (items 36–38). On the original FMCE (designed for American students), item 35 is worded as follows: "The car is pushing on the truck, but not hard enough to make the truck move." When the Japanese translation of this sentence is translated back into English, it is worded as follows: "The car is pushing on the truck, but the pushing force is insufficient to make the truck move." The use of the words "force" and "insufficient" in the Japanese translation might have caused differences in student responses. However, a literal translation of this sentence would have read too awkwardly to probe "natural" views. The next item asks about the force needed to increase speed, which implies a stronger pushing force than the force in item 35. The phrase "not hard enough to make the truck move" indicates that a stronger force could have made the truck move. The translation could prompt some Japanese students to think that the force of the pushing car is smaller than the force of the resisting truck. The corresponding illustration in which the truck is drawn so that it appears heavier than the car may further support this thinking.

Items 37 and 38 use the term "cruising speed." The corresponding Japanese term of "cruising speed" is used in navigational systems of airplanes and ships. Japanese cars seldom have a cruise-control feature, and, more importantly, students do not apply this term to describe their driving. In the Japanese translation, the term "cruising speed" is accompanied by a footnote describing what it means. The term and its footnote may have taken an awkward series of questions and drawn extra attention to them.

Comparing IRCs enabled us to detect items with different responses caused by intrinsic differences in student views, contextual effects from cultural differences, and artifacts of translation. We discerned which items were artifacts of translation through inference. Adopting a qualitative approach in future studies (e.g., conducting interviews with students) should help to confirm which translated items require further revision.

Revising the translation is the only way to improve the accuracy of the data, and comparing IRCs is an objective method of detecting which items to work on. One way of determining the limit of translation improvement would be to compare the IRCs of English-speaking non-American students with very different backgrounds.

VI. CONCLUSIONS

The comparison of American and Japanese IRCs revealed similarities and differences for all possible responses across the range of ability levels. Overall, the IRCs were strikingly similar when background differences and possible contextual alterations caused by the translation were taken into account. The comparison elicited groups of items, leading us to attribute pattern differences to educational background differences, cultural differences, and artifacts of translation. Identifying items that contain artifacts of translation is an effective way of improving the quality of translation, which in turn would increase the accuracy of determining other causes of differences in student responses.

IRCs, as described by Morris *et al.* [5,6], are straightforward to create and can be analyzed qualitatively and intuitively. They provide information about all possible responses and about students across the range of ability levels. One drawback is the large sample size that is required to have sufficient data across the full range of ability levels. Comparing different populations can provide insights into students' backgrounds. In addition, comparing preinstruction and postinstruction samples allows the assessment of different approaches to instruction. IRCs can be used to compare data from very different populations, particularly those involving translation of the survey from which the data are gathered. Taking a closer look at item functioning can help the test validation process by considering whether the translated assessment produces comparable data. Comparable functioning allows the sharing of results and resources across pretests and posttests, across curricula, and across nations.

[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] L. Bao, T. Cai, K. Koenig, K. Fang, J. Han, J. Wang, Q. Liu, L. Ding, L. Cui, Y. Luo, Y. Wang, L. Li, and N. Wu, Learning and scientific reasoning, Science **323**, 586 (2009).
[4] 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).
[5] 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).
[6] 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).
[7] P. M. Sadler, Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments, J. Res. Sci. Teach. **35**, 265 (1998).
[8] 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).
[9] PhysPort, https://www.physport.org/assessments/assessment.cfm?I=5&A=FCI.
[10] P. M. Sadler, H. Coyle, J. L. Miller, N. Cook-Smith, M. Dussault, and R. R. Gould, The Astronomy and Space Science Concept Inventory: Development and validation of assessment instruments aligned with the K–12 National Science Standards, Astron. Educ. Rev. **8**, 010111 (2010).
[11] D. R. Sokoloff, R. K. Thornton, and P. W. Laws, *RealTime Physics Active Learning Laboratories, Module 1: Mechanics*, 3rd ed. (Wiley, Hoboken, NJ, 2011).
[12] 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).
[13] Japanese Ministry of Education, Culture, Sports, Science and Technology, The New Course of Study in Science in Middle School, http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2011/04/11/1298356_5.pdf.
[14] American Institute of Physics, High School Physics Enrollments, 1987–2013, https://www.aip.org/statistics/physics-trends/high-school-physics-enrollments-1987-2013.
[15] 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).
[16] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).
[17] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory* (Sage, Newbury Park, CA, 1991).
[18] G. A. Davenport and H. J. Rogers, Strategic measurement model selection for conceptual diagnostic assessments, *Paper presented at the 2013 annual meeting of the American Educational Research Association. Retrieved 2017, from the AERA Online Paper Repository* (San Francisco, CA, 2013).
[19] M. Ishimoto, H. Nitta, and R. Lang, 2014 Survey on Physics Education in Japan (2): Item Response Curves of the FCI Data, *Proceedings of the Annual Meeting of the Physical Society of Japan* (Tokyo, 2015), https://www.jstage.jst.go.jp/article/jpsgaiyo/70.1/0/70.1_3482/_article/-char/ja/.