Validating the Japanese translation of the Force and Motion Conceptual Evaluation and comparing performance levels of American and Japanese students

Michi Ishimoto,[1] Ronald K. Thornton,[2] and David R. Sokoloff[3]

[1]*School of Environmental Science and Engineering, Kochi University of Technology, Tosayamada-cho,
Kami-shi, Kochi 782-8502, Japan*
[2]*Departments of Physics and Education, Center for Science and Mathematics Teaching,
Tufts University, Medford, Massachusetts 02155, USA*
[3]*Department of Physics, University of Oregon, Eugene, Oregon 97403, USA*

This study assesses the Japanese translation of the Force and Motion Conceptual Evaluation (FMCE). Researchers are often interested in comparing the conceptual ideas of students with different cultural backgrounds. The FMCE has been useful in identifying the concepts of English-speaking students from different backgrounds. To identify effectively the conceptual ideas of Japanese students and to compare them to those of their English-speaking counterparts, more work is required. Because of differences between the Japanese and English languages, and between the Japanese and American educational systems, it is important to assess the Japanese translation of the FMCE, a conceptual evaluation originally developed in English for American students. To assess its appropriateness, we examined the performance of a large sample of students on the translated version of the FMCE and then compared the results to those of English-speaking students. The data comprise the pretest results of 1095 students, most of whom were first-year students at a midlevel engineering school between 2003 and 2012. Basic statistics and the classical test theory indices of the translated FMCE indicate that its reliability and discrimination are appropriate to assess Japanese students' concepts about force and motion. In general, the preconcepts of Japanese students assessed with the Japanese translation of the FMCE are quite similar to those of American students assessed with the FMCE, thereby supporting the validity of the translated version. However, our findings do show (1) that only a small percentage of Japanese students grasped Newtonian concepts and (2) that the percentage of Japanese students who used two different concept models together to answer some questions seems to be higher than that of American students.

I. INTRODUCTION

The use of standard examinations, such as the Force Concept Inventory (FCI) [1] and the Force and Motion Conceptual Evaluation (FMCE) [2], to test student understanding of physics concepts has led to the development of many effective instructional methods in the United States [3]. In Japan, many physics instructors have attributed student learning to students' individual characteristics rather than to the instructional methods employed. These educators do not have the means to measure the efficacy of their instructional methods. The dissemination of concept inventories, such as the FCI and the FMCE, is a first step in improving instructional methods scientifically. Thus far, no original concept inventories have been developed in Japan. For this reason, the Japanese translation of established concept inventories in the United States is the fastest, most convenient way for Japanese educators to obtain concept inventories. Because of differences between the Japanese and English languages, and between the Japanese and American educational systems, it is important to assess the Japanese translation of the concept inventories, tests originally developed in English for American students.

In Japan, the subject of science has been taught in the Japanese language for 150 years. All technical terms have been translated into Japanese so that all Japanese can teach and learn science without learning foreign languages. We surmise that most Japanese physics teachers have limited English proficiency and do not have a full understanding of the nature of the concept inventories, which differ from regular physics problems. From about the year 2000 onward, several Japanese translations of the FCI (written in plain English) have been conducted. In general, the translators managed to produce a well-written introductory physics aptitude test but paid little attention to the distractors. They used their translated versions to calculate the normalized gain so as to evaluate their instructions. Beginning in 2009, several Japanese researchers in physics education worked to amalgamate the translated versions of

the FCI in existence at that time and then uploaded the amalgamated FCI (we refer to it as the "FCIJ" in this study) to the Arizona State University Modeling Instruction Program website in 2011 [4]. However, a translation does not guarantee an accurate transfer of the meaning of the original work. To the best of our knowledge, validation of the FCIJ is still required.

In Japan, the FMCE is not as widely used as the FCI owing to the longer time required in administering the FMCE. The time factor is especially important in high schools where more instructors are interested in finding ways to improve their instructional delivery. In addition, physics instructors using the FCI are not aware of the merits of using the FMCE for class assessment purposes. One of the authors of the present study translated the FMCE and has been using the translated version to evaluate the effectiveness of class instruction since 2003. We refer to this translated version as the "FMCEJ" in this study. We chose to evaluate the test quality and validity of the FMCEJ rather than the FCIJ for the following four reasons:

(1) The FMCE provides a more detailed measure of student understanding by virtue of having a greater number of items covering a narrower range of topics to probe [5].

(2) We have a large set of pretest data to validate.

(3) American data are available for comparison purposes.

(4) The FMCE is in a clustering format, thus allowing detailed analysis.

We used indices of the classical test theory to validate the translation because they are simple to calculate and are considered a standard procedure for ensuring the quality of the concept inventory [6].

II. METHODOLOGY

To be considered useful tools for assessing student learning, concept inventories have to meet the conditions of a high-quality test and elicit students' concepts quantitatively. We used the classical test theory indices to examine the test quality of the FMCEJ because the indices distinguish the quality of multiple-choice tests [6]. We calculated the indices to determine if they satisfy the recommended criteria of a high-quality test [7]. To examine whether the FMCEJ measures Japanese students' concepts with reasonable accuracy, we assumed that Japanese students and American students share the same concepts of motion and force before they receive formal instruction in physics. Based on this assumption, American and Japanese data should show similar responses on the pretest. If the similarity is reasonably close, the test is likely to be valid. Although test validity is usually examined by interviewing students, we did not establish interviewing methods for validation of physics concepts in Japan. Therefore, we decided to compare statistically our data

from Japanese students with data from American students for validation. For the purpose of comparison with our data, we used Smith and Wittmann's [8] data to represent American data.

A. Students surveyed

For the purposes of the present study, the FMCEJ was administered to mostly first-year students attending a basic mechanics class at the Kochi University of Technology (KUT) between 2003 and 2012. This midsized public university has an engineering division that consists of three schools: the School of Systems Engineering (SE), the School of Information (I), and the School of Environmental Science and Engineering (ESE). We obtained data from 1095 students who answered all 47 questions on the FMCEJ. (The surveyed students were distributed among the schools as follows: 40% SE, 5% I, and 55% ESE, respectively.) Each year, more than 500 000 university candidates in Japan write the National Center Test (NCT) for University Admissions, a standardized academic aptitude test. On average, the KUT students surveyed for the present study earned a score of 63 points (of 100 points) on the NCT in physics, which is comparable to the national average of 62.68 points reported in 2013 [9]. Thus, we infer that the data must be suitable for assessing the physics preconcepts of typical Japanese students.

B. Basic statistics of the FMCEJ for two kinds of scoring systems

The FMCE is designed to probe students' views of force and motion concepts using clusters with respect to subjects (kinematics, Newton's second law of motion, Newton's third law of motion, and energy conservation) and representation (natural language or graphical questions). According to Thornton and Sokoloff [2], an accurate evaluation could be obtained with 40 of the 47 questions.
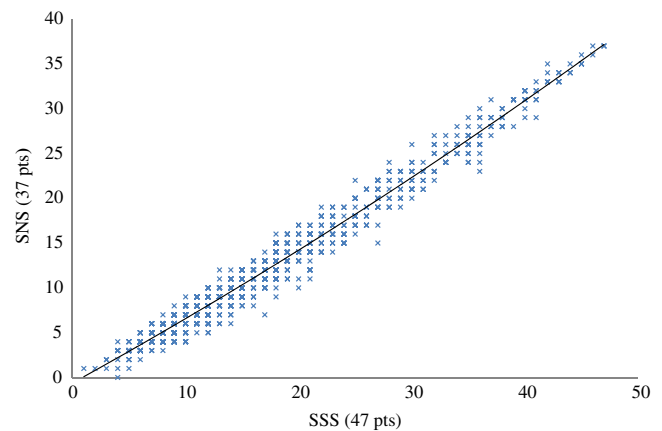

FIG. 1. The correlation between the SSSs and the SNSs of the FMCEJ. The correlation coefficient is $r = 0.99$ ($N = 1095$ students). The solid line shows a best fit (regression) to the data.

TABLE I.   Basic statistics for two scoring systems: simple sum score and single number score.

Basic statistics	SSS	SNS
Mean	16.05	11.35
Standard error	0.29	0.22
Median	13	9
Standard deviation	9.51	7.32
Range	46	37
Confidence interval (95.0%)	0.56	0.43

They omitted questions 5, 6, 15, 33, 35, 37, and 39. Thornton *et al.* [5] said that single-number scores (SNSs) could be useful in some circumstances and designed a SNS with a total of 33 points. (Their approach entailed excluding the portion on energy conservation, which consists of 4 questions.) They suggested that 9 of the 40 questions be divided into three sets of 3 questions, with a score of 2 points being obtained if all 3 questions in a set are answered correctly [$(2 \text{ points} \times 3) + (40 - 4 - 9 \text{ points}) = 33 \text{ points}$].

In the present study, we use their SNSs and 4 points from the 4 questions on energy conservation and refer to them as our SNSs. The inclusion of the 4 questions on energy conservation thus changes the possible total score on the

FIG. 2.   Total score distributions of the pretest results of the FMCEJ. (a) The SSS distribution (47 pts) of the pretest results of the FMCEJ (mean = 16 pts, median = 13 pts, N = 1095 students). (b) The SNS distribution (37 pts) of the pretest results of the FMCEJ (mean = 11 pts, median = 9 pts, N = 1095 students).

FMCEJ to 37 points. The SNSs are well correlated to the scores of the simple sum scores (SSSs) for the 47 questions (Fig. 1). The SSS values overestimate the low-scoring students and underestimate the high-scoring students. Table I shows the statistical values of the SSSs and the SNSs questions. The score distributions of the SSS and the SNS are shown in Fig. 2.

III. RESULTS

Evaluation of the FMCEJ

Table II shows the evaluation of the analysis of the FMCEJ, based on the classical test theory [6]. The difficulty index is the ratio of correct responses to the total number of responses for each item question. The average difficulty index over 47 items of the FMCEJ was 0.34 for the pretest, and the value was expected to rise for the posttest. The discrimination index determines the discrimination power of an individual test item. It measures the extent to which a single test item distinguishes students who know the material well from those who do not. The point biserial coefficient is defined as "a measure of consistency of a single item with the whole test. It reflects the correlation between students' scores on an individual item and their scores on the entire test" [7]. As shown in Table II, the average of these indices of the FMCEJ is above the desired values [7]. The reliability index KR-20 is a measure of the self-consistency of a whole test. Ferguson's δ is a measure of the discriminate power of an entire test, meaning that it reflects how broadly the total scores of a sample are distributed over a possible range. The reliability index and the discriminatory power of the pretest were 0.91 and 0.98, respectively. The index values indicate that the FMCEJ is a high-quality concept inventory test and is appropriate for evaluating individuals' understanding of concepts of force and motion.

Table III shows each item's difficulty index, point biserial coefficient, and discrimination index based on data from the top 25% of scorers and the bottom 25% of scorers. The columns labeled "MC" and "NA" in Table III show the FMCEJ distractor symbols for the most common sense
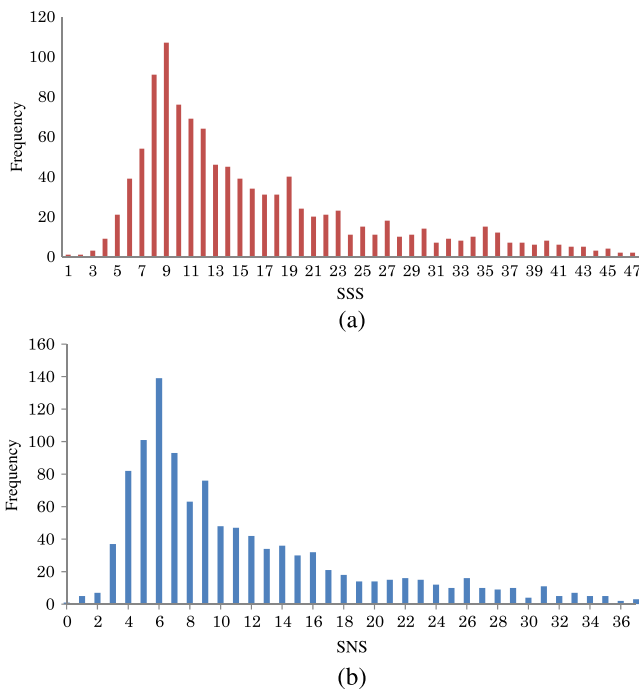
TABLE II.   Evaluation of the FMCEJ written by Japanese students.

Evaluation measure	FMCEJ (47 points)	Desired values
Difficulty index	Average of 0.34	0.3–0.9
Discrimination index	Average of 0.46	≥ 0.30
Point biserial coefficient	Average of 0.49	≥ 0.2
KR-20	Average of 0.91	≥ 0.7
Ferguson's δ	0.98	≥ 0.9
Number of students	1095	N/A
Average	16.05 ± 0.29	N/A
Standard deviation	9.51	N/A
Confidence interval (95.0%)	0.564	N/A

TABLE III.   Difficulty index (p), point biserial coefficient (PBC), discrimination index (D), and MC of the FMCEJ.

Item No.	p	PBC	D[a] (25%)	MC[b]	NA	Item No.	p	PBC	D (25%)	MC	NA
1	0.24	0.61	0.58	a	0.01	25	0.31	0.68	0.74	f	0.03
2	0.16	0.56	0.45	b	0.04	26	0.38	0.66	0.83	a	0.02
3	0.35	0.57	0.66	c	0.02	27	0.22	0.68	0.66	g	0.00
4	0.26	0.54	0.52	g	0.01	28	0.16	0.63	0.49	d	0.01
5	0.48	0.47	0.62		0.04	29	0.29	0.57	0.62	b	0.00
6	0.14	0.54	0.39		0.10	30	0.17	0.29	0.24	a	0.01
7	0.41	0.53	0.65	e, a	0.02	31	0.18	0.27	0.25	f	0.03
8	0.11	0.53	0.32	g	0.01	32	0.35	0.41	0.52	b, f	0.13
9	0.09	0.54	0.27	d	0.02	33	0.88	0.21	0.19		0.04
10	0.19	0.48	0.41	b	0.01	34	0.34	0.44	0.54	b	0.11
11	0.16	0.67	0.51	g	0.00	35	0.32	0.37	0.39		0.02
12	0.14	0.62	0.41	d	0.01	36	0.06	0.35	0.15	c	0.05
13	0.25	0.57	0.54	b	0.00	37	0.44	0.33	0.37		0.17
14	0.13	0.67	0.42	a	0.05	38	0.07	0.30	0.14	b	0.20
15	0.96	0.10	0.05		0.02	39	0.47	0.45	0.59		0.01
16	0.18	0.71	0.53	c	0.01	40	0.93	0.10	0.08		0.01
17	0.11	0.67	0.39	b	0.07	41	0.80	0.23	0.27	g	0.06
18	0.12	0.68	0.40	h	0.06	42	0.87	0.18	0.17		0.02
19	0.15	0.69	0.48	d	0.06	43	0.89	0.16	0.15		0.07
20	0.23	0.53	0.48	f	0.14	44	0.37	0.53	0.66	a	0.01
21	0.19	0.67	0.57	f, h, a	0.12	45	0.55	0.45	0.61	a	0.01
22	0.39	0.68	0.83	e, a	0.01	**46**	0.42	0.46	0.61	**d**, c	0.02
23	0.26	0.70	0.72	**g**	0.03	**47**	0.52	0.45	0.61	**d**, c	0.01
24	0.34	0.67	0.79	b	0.07						

[a]The D value is based on data from the top 25% of scorers and the bottom 25% of scorers.
[b]The 10 blank spaces in the MC columns of Table III indicate that no MCs were probed. The first 7 blank spaces correspond to the 7 items (5, 6, 15, 33, 35, 37, and 39) regarded by Ref. [2] as inaccurate for the assessment. The remaining 3 blank spaces correspond to velocity graph questions (items 40, 42, and 43); the correct response rate was nearly 90% for these items, and, therefore, the other responses were too minute to be listed.

concepts (MCs) and the proportion of "no correct answers" (NAs). The results showing that KUT students' MC symbols agree with those of American students support our assumption that Japanese students and American students share the same concepts of motion and force before they receive formal instruction in physics. This similarity supports the FMCEJ as a valid form of assessment. The items with different MCs from those of the U.S. counterparts are items 23, 46, and 47 (presented in bold type). We discuss these three items in greater depth in the cluster section. The 10 blank spaces in the MC columns of Table III indicate that no MCs were probed. The first 7 blank spaces correspond to the 7 items considered to be irrelevant for listing MCs. According to Thornton and Sokoloff [2], the students' responses for these 7 items reflect something other than the intent of these questions, so we left them blank. The remaining 3 blank spaces correspond to velocity graph questions. Because the correct response rate was near 90% for these questions, there is no point to elicit MCs on these items.

The classical test theory indices of the FMCEJ are above the desired values, and the MCs of the Japanese students, with the exception of the three items (items 23, 46, and 47), agree with those of American students. Thus, we regard the FMCEJ as a high-quality test that can be used to compare

Japanese students' and American students' understanding of Newtonian concepts.

IV. ANALYSIS

A. Difficulty level of clusters

Smith and Wittmann [8] used the following clusters to analyze FMCE data: force sled (FS), reverse direction (RD), force graph (FG), acceleration graph (aG), Newton's third law (N3), velocity graph (vG), and energy (KE). We used these same clusters to analyze our data. In the KE cluster, the same questions are posed about velocity and kinetic energy. Because the difficulty indices of the velocity questions are intrinsically lower (by about 10% and higher) than those of the kinetic energy questions, we suspected that some students answered the kinetic energy questions correctly without understanding the concept. To examine this hypothesis, we included the velocity energy (vE) cluster, a subset of the KE cluster, in our analysis in Table IV.

Table IV shows the average difficulty indices for the clusters in the group of students, with the corresponding SNS in the first column. The second column shows the number of students with the corresponding SNS, and the third column shows the population percentage

TABLE IV.   Average difficulty index in clusters in terms of SNS.

SNS	No. of students	Cum (%)	FS	RD	FG	aG	N3	vG	KE	vE
37	3	0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
36	2	0.5	1.00	1.00	1.00	1.00	1.00	0.88	0.88	0.75
35	5	0.9	0.88	1.00	1.00	1.00	0.80	1.00	0.95	0.90
34	5	1.4	0.88	1.00	0.86	1.00	0.77	1.00	1.00	1.00
33	7	2.0	1.00	0.86	0.96	0.97	0.62	0.96	0.93	0.86
32	5	2.5	0.88	1.00	0.97	1.00	0.37	1.00	0.90	0.90
31	11	3.5	0.91	0.85	0.90	0.95	0.42	0.98	0.98	0.95
30	4	3.8	0.75	0.83	0.64	0.95	0.71	1.00	0.94	0.88
29	10	4.7	0.94	0.53	0.94	1.00	0.48	1.00	0.65	0.65
28	9	5.6	0.93	0.70	0.87	0.87	0.43	0.78	0.75	0.67
27	10	6.5	0.82	0.60	0.71	0.98	0.33	0.98	0.88	0.90
26	16	7.9	0.81	0.42	0.75	0.95	0.41	0.86	0.89	0.91
25	10	8.9	0.76	0.47	0.77	0.96	0.27	0.90	0.75	0.75
24	12	10.0	0.75	0.36	0.57	0.93	0.33	0.94	0.92	0.83
23	15	11.3	0.81	0.24	0.64	0.93	0.29	0.95	0.70	0.60
22	16	12.8	0.74	0.29	0.55	0.73	0.31	0.98	0.81	0.72
21	15	14.2	0.55	0.27	0.46	0.87	0.41	0.98	0.68	0.63
20	14	15.4	0.50	0.29	0.31	0.87	0.30	0.98	0.89	0.86
19	14	16.7	0.56	0.24	0.32	0.91	0.30	0.84	0.71	0.61
18	18	18.4	0.59	0.09	0.28	0.74	0.33	0.92	0.79	0.69
17	21	20.3	0.55	0.13	0.22	0.71	0.24	0.93	0.81	0.86
16	32	23.2	0.47	0.04	0.17	0.69	0.31	0.95	0.77	0.70
15	30	25.9	0.39	0.09	0.17	0.65	0.27	0.93	0.68	0.65
14	36	29.2	0.28	0.08	0.09	0.55	0.30	0.96	0.78	0.69
13	34	32.3	0.23	0.06	0.11	0.58	0.25	0.93	0.64	0.53
12	42	36.2	0.23	0.06	0.08	0.51	0.23	0.92	0.57	0.48
11	47	40.5	0.27	0.01	0.05	0.31	0.22	0.93	0.65	0.57
10	48	44.8	0.20	0.00	0.07	0.32	0.22	0.94	0.46	0.40
9	76	51.8	0.23	0.01	0.04	0.20	0.20	0.90	0.42	0.34
8	63	57.5	0.17	0.02	0.03	0.13	0.16	0.90	0.41	0.27
7	93	66.0	0.18	0.00	0.02	0.08	0.13	0.86	0.33	0.25
6	139	78.7	0.11	0.00	0.02	0.04	0.09	0.87	0.28	0.18
5	101	87.9	0.08	0.00	0.02	0.04	0.05	0.84	0.16	0.10
4	82	95.4	0.03	0.00	0.01	0.01	0.03	0.79	0.10	0.07
3	37	98.8	0.01	0.00	0.02	0.01	0.04	0.56	0.08	0.04
2	7	99.5	0.06	0.00	0.02	0.06	0.10	0.14	0.04	0.07
1	5	99.9	0.08	0.00	0.00	0.00	0.03	0.00	0.10	0.10
0	1	100.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average			0.28	0.10	0.16	0.34	0.20	0.87	0.46	0.40

of the students whose scores are equal to or higher than all the students. The values for the seven clusters represent the average difficulty index of the group. The average difficulty indices as a function of the SNS by cluster are presented in Fig. 3.

The values in the FS cluster correlate linearly to the SNS. The set of five items in the FS cluster represents students' SNS values. The RD values are very selective; that is, only a low percentage of the population provided correct answers. The values drop sharply below the SNS of 30 points (81% score). The FG questions were difficult for students to answer correctly, with values below the SNS of 20 points (54% score). The aG questions appeared to be less difficult for students to answer correctly than the FG

questions. The N3 questions were extremely difficult for students to answer correctly, and even students with high SNSs could not answer these questions well. This is consistent with results in English-speaking populations. The vG questions are of little use for the purposes of evaluation and assessment in this population because more than 90% of the students answered them correctly. Our finding that students with an SNS of 17 points answered more than 80% of the KE questions correctly indicates the ease with which the students answered these particular questions. However, the values for the vE questions are lower (not shown in Fig. 3), indicating that the difficulty indices for the KE questions included more false positives.
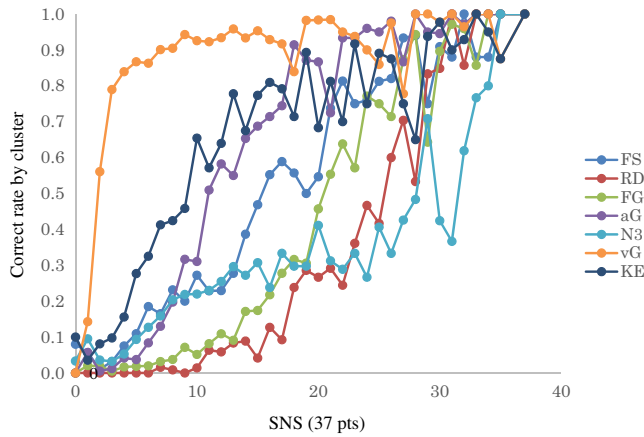
FIG. 3.   The average difficulty index in each cluster with respect to the SNS. The values in the FS cluster correlate linearly to the SNS.

B. FS cluster

The FS questions, written in natural language, probe the relationship between force and motion. The MC is that an applied force is proportional to the velocity of the sled, the $F \propto v$ model. The corresponding proportions are shown in Table V in bold type. A very low value was determined for item 5 of the FMCEJ, a finding that is consistent with a "false-positive response" observed by Thornton and Sokoloff [2]. Table V shows that in the case of an increase in the sled's speed ($\pm v \ up$), students' responses reflecting their use of the $F \propto v$ model were as high as 70%. However, in the case of a decrease in the sled's speed ($\pm v \ down$), students' responses reflecting their use of the $F \propto v$ model dropped to around 30%. Instead, students' responses reflecting their use of the opposite force against the velocity direction model (the brake model) increased to 70%. A comparison of students' responses in which the $F \propto a$ model was used in the case of a decrease in the sled's speed and in the case of an increase in the sled's speed revealed that the former was 10% higher (not the $F \propto a$ model). Therefore, in the case of a decrease in the sled's speed, we estimate that 20% of students used the $F \propto a$ model and another 20% used a variation of the brake model

to make up the 40% response rate of the $F \propto a$ model. Because 20% of the response rate of the brake model is actually the $F \propto a$ model, the rate of the common sense brake model would be 40% (60%−20% = 40%). Thus, we estimate that 20% used the $F \propto a$ model, 30% consistently used the $F \propto v$ model, and 40% used the brake model. Less than 1% of students answered all five items (items 1, 2, 3, 4, and 6) correctly, and 9% of the top scorers' average had a difficulty index above 80%. Thus, we infer that less than 10% of the students used the $F \propto a$ model consistently.

C. RD cluster

The three subclusters include a set of three questions asking about the net force of an object on an inclined surface and the force of an object with free fall motion and its acceleration. The motion of the object includes the point where the direction of the motion reverses. According to Thornton and Sokoloff [2], students are considered to provide correct answers only when all three questions within a given subcluster are answered correctly. Smith and Wittmann [8] suggested the direction-of-motion model ($F\|v$), which includes the $F \propto v$ model. Table VI shows that 10% of the students used the $F \propto a$ model, 60% used the $F \propto v$ model, and 80% used the $F\|v$ model to describe the net force of an object on an inclined surface (items 8, 9, and 10). A comparison of students' responses revealed that the percentage of correct responses was 5% lower for questions on the net force acting on a toy car on a ramp (items 8, 9, and 10) than for those on the force acting on a tossed coin (items 11, 12, and 13). The percentage of correct responses to questions on the acceleration of a tossed coin was higher than that on force. Because the concept of acceleration is more abstract than the concept of force, students' prior mathematics knowledge (i.e., acceleration as the derivative of velocity) might have contributed to this false positive. Only 52 of 1095 students answered the questions posed in all nine items correctly. The percentages of perfect correct responses for the sets of three questions were 6.7%, 6.5%, and 12.1% for the ramp, coin toss force, and coin toss acceleration subclusters, respectively. The highest percentage of correct responses in

TABLE V.   Response rates for force sled question (FS).

$\vec{v}$	F increase	F constant	F decrease	$F = 0$	$-F$ increase	$-F$ constant	$-F$ decrease	Item No.	$F \propto a$	$F\|v$	Brake
v increase	**0.74**[a]	0.24[b]	0.00	0.00	0.01	0.00	0.00	1	0.24	0.98	
v constant	0.02	**0.70**[a]	0.05	0.16[b]	0.00	0.02	0.00	2	0.16	0.78	
v decrease	0.01	0.01	**0.35**[a]	0.03	0.16	0.35[b]	0.08	3	0.35	0.35	0.59
$-v$ increase	0.01	0.01	0.00	0.01	**0.68**[a]	0.26[b]	0.03	4	0.26	0.96	
$-v$ constant	0.03	0.35	0.04	0.48[b]	0.00	**0.04**[a]	0.01	(5)[c]	0.48		
$-v$ decrease	0.20	0.41[b]	0.07	0.03	0.01	0.03	**0.23**[a]	7	0.41	0.23	0.67

[a]The corresponding proportions of the $F \propto v$ model are shown in Table V in bold type.
[b]The corresponding proportions of the $F \propto a$ model are shown in Table V in underlined type.
[c]Item 5 is inaccurate for assessment, according to Ref. [2].

TABLE VI. Response rates for reversing direction force and acceleration.

Subcluster	Direction	$F \propto a$	$F \propto v$	$F\|v$[a]	$F\|v - F \propto a$	Item No.
Net force (ramp)	Up	0.11	0.59	**_0.78_**[a]		8
	Top	0.09	0.82	**_0.82_**[a]		9
	Down	0.19	0.72	**_0.93_**[a]	0.74	10
Toss force	Up	0.16	0.64	**_0.79_**[a]		11
	Top	0.14	0.83	**_0.83_**[a]		12
	Down	0.25	0.71	**_0.79_**[a]	0.72	13
Toss acceleration	Up	0.22	0.52	**_0.64_**[a]		27
	Top	0.16	0.80	**_0.80_**[a]		28
	Down	0.29	0.61	**_0.92_**[a]	0.6	29

[a]$F\|v$ is the direction-of-motion model suggested by Ref. [8], which includes the $F \propto v$ model.

the acceleration subcluster might be attributed to students' prior mathematics knowledge.

Students' responses reflecting their use of the $F \propto a$ model to describe the force acting on a tossed coin that is at its highest point may be considered true Newtonian responses. If so, students' responses reflecting their use of the same model to describe the force on a tossed coin moving upward include false positives; a small percentage of these false positives may be attributed to students' use of the brake model. Students' responses reflecting their use of the $F \propto a$ model to describe the force acting on a tossed coin moving downward include false positives. Ten percent of these false-positive responses are likely attributed to students' use of the $F\|v$ model. Because the $F\|v$ model includes the $F \propto v$ model, the estimated percentage of the students who used the $F\|v$ model was near 80% in the case of slowing down with the upward motion (items 8 and 11), whereas 30% was observed in the FS cluster (items 3 and 7). The influence of context may explain this difference. The brake model is brought to mind in the case of applying force by human action in the FS cluster. Table VI shows that 80% of students used the $F\|v$ model and that 60% used the $F \propto v$ model, which is included in the former. The percentage of students who correctly answered the questions for all items in this cluster was only 5%. This is an expected result for this population.

D. Graphical representation (*F*G, *a*G, and *v*G)

For the FG questions, students were asked to refer to the force-versus-time graph corresponding to the horizontal motion of a toy car. The concepts probed here are identical to those of the FS cluster, the only difference being the graphical representation.

The Newtonian response is that an applied force is proportional to the rate of its velocity change, the $F \propto a$ model. Students' responses reflecting their use of the $F \propto a$ model were consistently around 15% for five items in the FG cluster (Table VII). This finding is in contrast to students' responses reflecting their use of the $F \propto a$ model in the FS cluster, which varied from 16% to 41% (Table V). The corresponding proportions are shown in Table VII in underlined type.

The MC is that an applied force is proportional to the velocity of the toy car, the $F \propto v$ model, which is similar to those of the FS cluster. The corresponding proportions are shown in Table VII in bold type. In the case of constant speed, students' responses reflecting their use of the $F \propto v$ model were as high as 80% and their use of the $F \propto a$ model were 13%; these students' responses were very

TABLE VII. Response rates for force graph.

Direction magnitude[c]	Force-versus-time graph						Item No.	$F \propto a$	$F \propto v$	$F\|v$	Brake
	F increase	F constant	F decrease	$F = 0$	$-F$ increase	$-F$ constant	$-F$ decrease					
v increase	**0.80**[a]	_0.18_[b]	0.00	0.00	0.00	0.00		16	0.18	0.8	0.98	
v constant	0.03	**0.77**[a]	0.01	_0.13_[b]	0.00	0.01	0.00	14	0.13	0.77	0.94	
v decrease	0.01	0.00	**0.56**[a]	0.00	0.09	_0.12_[b]		18	0.12	0.56	0.57	0.21
$-v$ increase	0.04	0.01	0.01	0.00	**0.70**[a]	_0.15_[b]		19	0.15	0.7	0.85	0.06
$-v$ constant	0.01	0.02	0.03	_0.11_[b]	0.06	**0.68**[a]		17	0.11	0.68	0.86	0.06
$-v$ decrease												

[a]The corresponding proportions of the $F \propto v$ model are shown in Table VII in bold type.
[b]The corresponding proportions of the $F \propto a$ model are shown in Table VII in underlined type.
[c]The direction and magnitude column, "v" represents the direction of motion to the right, and "$-v$" to the left.

TABLE VIII. Response rates for acceleration graph.

| Direction magnitude[c] | Acceleration-versus-time graph | | | | | | | Item No. | $F \propto a$ | $F \propto v$ | $F\|v$ | Brake |
	F increase	F constant	F decrease	$F = 0$	$-F$ increase	$-F$ constant	$-F$ decrease					
v increase	**0.58**[a]	0.39[b]	0.00	0.03	0.00	0.00		22	0.39	0.58	0.97	
v constant	0.01	**0.53**[a]	0.02	0.38[b]	0.00	0.01	0.00	26	0.38	0.53	0.94	
v decrease	0.02	0.01	**0.52**[a]	0.01	0.10	0.26[b]		23	0.26	0.52	0.56	0.36
$-v$ increase	0.03	0.02	0.02	0.01	**0.57**[a]	0.31[b]		25	0.31	0.57	0.85	0.07
$-v$ constant	0.01	0.03	0.05	0.34[b]	0.02	**0.48**[a]		24	0.34	0.48	0.84	0.09
$-v$ decrease												

[a]The corresponding proportions of the $F \propto v$ model are shown in Table VIII in bold type.
[b]The corresponding proportions of the $F \propto a$ model are shown in Table VIII in underlined type.
[c]In the direction and magnitude column, "v" represents the direction of motion to the right, and "$-v$" to the left.

similar to those in the FS cluster. In the case of an increase in the toy car's speed ($\pm v$ up), students' responses reflecting their use of the $F \propto v$ model were as high as 80%, similar to those in the FS cluster.

In the case of a steadily decreasing positive velocity, 14% of students (not shown in Table VII) chose the force presented by a straight declining line with a positive initial value (right direction) and a negative final value (left direction). We surmise that the students who chose this answer had decided to use a mixed model comprising the $F \propto v$ model and the brake model. Students' responses reflecting their use of the $F \propto v$ model dropped to around 56% and more (depending on how the 14% of students described earlier are classified), which is in contrast to 35% in the FS cluster (Table V). Students' responses reflecting their use of the opposite force against the velocity direction model (the brake model) decreased to 9% and more (again, depending on how the 14% students described earlier are classified), which is in contrast to 40% in the FS cluster. In either case, students used the brake model to a lesser extent in the FG cluster. We speculate that the abstract representation of graphics might encourage students to respond with more consistency and less intuition.

For the aG questions, students were asked to refer to the acceleration-versus-time graph corresponding to the horizontal motion of a toy car. Students' responses reflecting their use of the $a = dv/dt$ model (the correct model) was consistently around 30% for five items (Table VIII). Compared with the correct response rate of the FG cluster, the rate was generally up by 15%. Students' responses reflecting their use of the $a \propto v$ model (the MC response) was consistently around 50%. On item 23 (a decrease in the toy car's speed), the Japanese students surveyed for the present study most commonly used the $a \propto v$ model (52%), whereas Smith and Wittmann [8] found that American students most commonly used the brake model. The brake model was used in only 10% of responses by students at KUT. The students at KUT consistently favored the $a \propto v(F \propto v)$ model. On item 3 in the FS cluster, Japanese students favored the brake model as much as did

American students. We surmise that the students' use of math knowledge (the $a = dv/dt$ model) explains the consistency.

In the vG cluster, a correct response rate of 90% was determined for all questions except for item 41 (leftward motion), which had a correct response rate of 80%, and the $v \propto x$ model, which had a response rate of 8%. In general, the correct response rate was a few percentage points lower for questions involving leftward motion. We suspect that the unfamiliar situation (leftward motion) might have influenced students to answer incorrectly.

In general, we surmise that the graphical representation of questions encouraged the Japanese students to recall their mathematics knowledge and to use abstract ideas to a greater extent than did questions written in natural language.

E. N3 cluster

The N3 cluster comprised questions about two different situations (pushing and colliding) in which a truck and a car in contact apply a force on each other. The correct response rates for questions 36 and 38 about a small compact car pushing a large truck that has broken down were less than 7%, whereas the correct response rates for questions about a truck and car colliding were 17% and higher. The fact that Japanese students are not familiar with the situation of a vehicle pushing another vehicle that has broken down might have influenced the response rates.

Item 30 refers to a situation where a car and a heavy truck traveling at the same speed collide with one another. The inclusion of this item on the FMCE is to determine students' use of the mass dependent model. In our survey, 80% of the students used the mass dependent model and 17% used the N3 model to answer the question posed in item 30. Item 34 refers to a situation where a car collides with a stopped truck with the same mass as the car. This item is included on the FMCE as a way to determine students' use of the action dependent model. In our survey, 41% of the students used the action dependent model, 34% used the N3 model, and 18% provided uncertain answers

TABLE IX. Response rates for Newton's third law.[a]

Mass	Colliding Velocity	$F > f$	$F = f$	$F < f$	Unknown	Item No.
$M > m$	$V = v$	0.77	0.17	0.01	0.03	30
$M > m$	$V < v$	0.08	0.18	0.2	0.51	31
$M > m$	$V = 0$	0.07	0.35	0.24	0.29	32
$M = m$	$V = v$	0.02	0.88	0.01	0.07	33
$M = m$	$V = 0$	0.03	0.34	0.41	0.18	34
Mass	Pushing Velocity	$F > f$	$F = f$	$F < f$	Unknown	Item No.
$M > m$	$V = v = 0$	0.60	0.32	0.04	0.02	(35)[b]
$M > m$	$V = v$ increasing	0.05	0.06	0.81	0.05	36
$M > m$	$V = v$ constant	0.03	0.44	0.28	0.17	(37)[b]
$M > m$	$V = v$ decreasing	0.56	0.07	0.07	0.20	38

[a]M and m are the masses of the truck and the car, respectively, V and v are the speeds of the truck and the car, respectively, and F and f are the forces exerted by the truck and the car, respectively.
[b]Items 35 and 37 are inaccurate for assessment, according to Ref. [2].

for the question posed in item 34. Because the students used the mass dependent model and the action dependent model together, the percentage of uncertain answers was 50% in item 31, which refers to a collision between a slow-moving, heavy truck and a fast-moving, light car. According to Smith and Wittmann [8], American students were relatively consistent in their use of the action dependent model. We estimated that 17% of the students in our survey used the N3 model for items 30 through 34.

Items 35 through 38 refer to a situation in which a small compact car pushes a heavy truck that has broken down on the road. The following two scenarios are given: (1) the car, pushing the truck, speeds up and (2) the car, pushing the truck, slows down when the truck puts on its brakes. In our survey, 80% of the students attributed the car's increase in speed to the car exerting a stronger force on the truck, and 60% attributed the car's decrease in speed to the truck exerting a stronger force on the car. In either case, the students used the action dependent model, as Smith and Wittmann [8] previously pointed out in their analysis of the FMCE. The most common responses on items 36 and 38 in our survey were "c" and "b," respectively, which agree with the American responses. We estimated the Newtonian responses to be 6% (Table IX). The difficulty indices of items 30, 31, 36, and 38 are very low, as are the point biserial coefficient and discrimination index values (Table III). The level of difficulty of these questions is

shown in Fig. 3. After effective instruction, many more students answered these questions correctly. As seen in the students' responses for item 31, a large number of the students used both the mass and action dependent models. Only 3% of the surveyed students correctly answered all the questions in the N3 cluster.

F. KE cluster

Items 44 through 47 refer to a situation in which a sled at rest at the top of a hill is released. The questions deal with the speed and KE of the sled after sliding down the hill. For items 44 and 45, the two hills have the same height but different steepness. The MC models are the slope dependent model (answer option "a") for the speed questions and the correct response (answer option "b") for the KE questions. If the students who stated that there was insufficient information to answer the questions posed in items 44 and 45 (answer option "d") are regarded, it can be presumed that 57% of the students attributed the greater speed of the sled to the steepness of the hill and that 40% attributed the greater KE of the sled to the steepness of the hill.

Items 46 and 47 refer to a situation in which a sled is released from rest at the top of a hill that is higher and less steep than the original hill, thereby adding the factor of height to the steepness. As shown in Table X, the most common responses were the correct responses for both the

TABLE X. Response rates for energy conservation.

Description of new hill compared with original hill	Correct	a	b	c	d	Item No.
Same height, steeper v	0.37	0.53	**0.37**[a]	0.05	0.04	44
Same height, steeper KE	0.55	0.35	**0.55**[a]	0.05	0.05	45
Higher, less steep v	0.42	**0.42**[a]	0.11	0.20	0.25	46
Higher, less steep KE	0.52	**0.52**[a]	0.16	0.12	0.18	47

[a]The proportion of the correct responses are shown in bold type.

TABLE XI.   Robustness of the concept of energy conservation with respect to the SNS level.

SNS[a]	$\frac{(Steep+Height)b}{Steep}$	$\frac{(Steep+Height)b}{Height}$	$\frac{(v+KE)c}{v}$	$\frac{(v+KE)c}{KE}$	$\frac{Steep}{Height}$	$\frac{v}{KE}$
High scorer ($n = 168$)	0.96	0.83	0.86	0.91	0.86	1.05
Mid scorer ($n = 336$)	0.89	0.68	0.82	0.75	0.76	0.92
Low scorer ($n = 591$)	0.5	0.21	0.38	0.25	0.42	0.66

[a]SNS: high score, >54%; mid score, 27% ≈ 54%; low score, <27%.
[b]The number of students who answered correctly on both "steep" hill and "higher" hill questions.
[c]The number of students who answered correctly on both "velocity" and "kinetic" energy questions.

speed (42%) and the KE (52%) of the sled after sliding down the hill. Based on the sum of incorrect answers, it can be presumed that 58% of the students incorrectly attributed the greater speed of the sled to the steepness of the hill and that 48% of the students incorrectly attributed the greater KE of the sled to the steepness of the hill. The percentages are consistent for both cases.

In our survey, those students who answered the questions posed in items 46 and 47 favored the steepness model in item 46 (greater speed of sled) to a larger extent than in item 47 (greater KE). We suspect that speed is a less abstract idea than KE, which cannot be seen in everyday life. Thus, the true proportion is likely closer to the students' responses regarding the speed of the sled rather than the KE of the sled.

The most common incorrect response (around 20%) provided by the Japanese students for the questions posed in items 46 and 47 was that insufficient information was given to answer the questions (answer option "d"). By contrast, the most common incorrect response given by American students was slope dependent (answer option "c"); that is, they incorrectly attributed the higher speed of the sled to the steepness of the hill.

Table XI shows the relationship between the sled's speed or KE and the conditions of the hill. The ratio of the intersection correct response of "steep" hill and "higher" hill (items 44 through 47) to the correct response of steep hill (items 44 and 45) is 96%, whereas the ratio to the correct response of higher hill (items 46 and 47) is 83% among the SNS high scorers. The ratios decrease to 50% and 21%, respectively, among the low scorers. The difference sets for the higher hill minus steep hill questions for the students with middle and low SNS values are 32% and 79%, respectively. Thus, the correct response of steep hill contains a higher ratio of understanding of energy conservation, especially for students with middle to low SNSs.

The ratio of the intersection correct response of steep hill and higher hill to the correct response of velocity (items 44 and 46) is 86%, whereas the ratio to the correct response of KE (items 45 and 47) is 91% among the SNS high scorers. The ratios are 38% and 25%, respectively, among the low scorers. The difference sets for the "KE" minus "velocity"

questions for the students with middle and low SNS values are 25% and 75%, respectively, thus indicating that both middle and low scorers did not have a firm grasp of this concept. The ratio of the correct response to steep hill questions is lower than that to the higher hill questions, and the ratio of the correct response to speed questions is lower than that to KE questions for low scorers.

The high scorers had a firm grasp of the concepts of energy conservation, whereas the middle and low scorers used the correct concept conditionally. The questions about the sled's speed appear to be better indicators of students' understanding because there were fewer false-positive responses in general. The steep hill questions also included fewer false-positive responses among most students. Of the students surveyed, 25% correctly answered all four questions on energy conservation, indicating that they possessed a good understanding of its concepts. This value is substantially higher than the percentage of surveyed students considered to be Newtonian thinkers (<5%). Of the four items on energy conservation, item 44 is the most reliable estimate of correct energy concepts because this item had the lowest number of false positives. Item 44 also had the highest point biserial coefficient and discrimination index values in the cluster (Table III).

V. PRECONCEPTS: SUMMARY

In our survey, we found the apparent proportions of Japanese students who used the Newtonian $F \propto a$ model, the $F \| v$ including $F \propto v$ model, and the brake model are roughly 10% to 20%, 40% and higher, and 40%, respectively. Although the students used a particular model depending on the context of the question, the percentage of students who used the $F \propto a$ model consistently is estimated to be less than 5%. The graphic representation produces more consistent use of the concept model among the students.

The common sense concept model most frequently used by students to answer questions on Newton's third law was the action dependent model, followed by the mass dependent model. The proportion of students with a good understanding of Newton's third law model is estimated to be less than 5%.

The common sense concept model most frequently used by students to answer questions on energy conservation was the steepness dependency model (three-quarters of the students used this particular model). The percentage of students with a good understanding of the energy conservation model is estimated to be less than 25%.

We conclude that, much like the situation that has been described in American high schools, only a small percentage of students graduating from high school in Japan are Newtonian thinkers.

VI. DISCUSSION AND CONCLUSIONS

This extended and detailed analysis of the responses of 1095 "average" Japanese students (as indicated by national testing) on the FMCEJ indicates that these students responded to multiple-choice questions (items) with keys (correct answers) and distractors (incorrect answers) in a manner consistent to the views found among American students. Basic statistics and the classical test theory indices of the FMCEJ indicate that its reliability and discrimination

are appropriate to assess Japanese students' concepts about force and motion. The preconcepts of Japanese students assessed with the FMCEJ are quite similar to those of American students assessed with the FMCE, thereby further supporting the validity of the FMCEJ.

We, unfortunately, could not address the subtle differences between the responses of American students and those of Japanese students on a few items. We hope that a future collaborative study using a database containing item-level responses for each American student may provide insight into these differences.

Our next study will be a comparison of the data of two groups of Japanese students—students who have received instruction via the traditional lecture format and those who have received instruction via interactive engagement—in terms of pretest scores, NCT scores, and gender.

Finally, this simple way of evaluating a translated concept inventory may help in validating the quality and compatibility of translations, thereby making cross-cultural comparisons more reliable in the future.

[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] E. F. Redish, *Teaching Physics with the Physics Suite*, http://www2.physics.umd.edu/~redish/Book/.

[4] Force Concept Inventory (Japanese Translation), http://modeling.la.asu.edu/R&E/Research.html.

[5] 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).

[6] P. V. Engelhardt, An introduction to classical test theory as applied to conceptual multiple-choice tests, in *Getting*

Started in PER, edited by C. Henderson, and K. A. Harper (American Association of Physics Teachers, College Park, MD, 2009), http://www.per-central.org/items/detail.cfm?ID=8807.

[7] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. **2,** 010105 (2006).

[8] 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).

[9] National Center for University Entrance Examinations, *National Center Test for University Admission average scores* (2013), http://www.dnc.ac.jp/modules/center_exam/content0565.html.