Analyzing false positives of four questions in the Force Concept Inventory

Jun-ichiro Yasuda
Institute of Arts and Sciences, Yamagata University, Yamagata, Yamagata 990-8560, Japan

Naohiro Mae
Faculty of Engineering Science, Kansai University, Suita, Osaka 564-8680, Japan

Michael M. Hull
Austrian Educational Competence Centre Physics, University of Vienna, Vienna 1090, Austria

Masa-aki Taniguchi
Center for Teacher Education, Meijo University, Nagoya, Aichi 468-8502, Japan

(Received 13 December 2017; published 8 March 2018)

In this study, we analyze the systematic error from false positives of the Force Concept Inventory (FCI). We compare the systematic errors of question 6 (Q.6), Q.7, and Q.16, for which clearly erroneous reasoning has been found, with Q.5, for which clearly erroneous reasoning has *not* been found. We determine whether or not a correct response to a given FCI question is a false positive using subquestions. In addition to the 30 original questions, subquestions were introduced for Q.5, Q.6, Q.7, and Q.16. This modified version of the FCI was administered to 1145 university students in Japan from 2015 to 2017. In this paper, we discuss our finding that the systematic errors of Q.6, Q.7, and Q.16 are much larger than that of Q.5 for students with mid-level FCI scores. Furthermore, we find that, averaged over the data sample, the sum of the false positives from Q.5, Q.6, Q.7, and Q.16 is about 10% of the FCI score of a midlevel student.

DOI: 10.1103/PhysRevPhysEducRes.14.010112

I. INTRODUCTION

The Force Concept Inventory (FCI) is one of the most widely used tests in physics education [1,2]. The FCI is a 30-item, five-choice survey to probe student conceptual understanding of Newtonian mechanics. One of its features is that its distractors are designed based upon knowledge of students' naïve conceptions. It uses everyday speech in order to better elicit what the student personally considers to be correct as opposed to an answer memorized by rote from physics class. The FCI has played an important role in analyzing the effects of newly developed pedagogy including interactive-engagement methods [3,4].

When developing or surveying with an assessment test, it is necessary to analyze its validity [5]. In the specific case of the FCI, we need to consider to what extent the test actually measures whether a student is thinking of the concept of force as a Newtonian thinker does. Prior to the study presented in this paper, the validity of the FCI had been evaluated in various ways. For example, by interviewing students and professors, Hestenes *et al.* confirmed that respondents correctly understood the wording and diagrams of the questions [1,6]. In addition, Stewart *et al.* used a context-modified test to investigate context sensitivity of the FCI (for example, whether students answer a question differently if the truck and car in an original FCI question are replaced with a bowling ball and marble), and they confirmed that the average test score is not particularly context dependent [7]. Recently, DeVore *et al.* examined the effects of testwiseness in the FCI [8]. Testwiseness is defined as the set of cognitive strategies used by a student that is intended to improve his or her score on a test, for example, by avoiding "none of the above" or "zero" distractors. They found that overall scores are not substantially affected by testwiseness; however, the effect on individual items could be substantial.

Although the FCI has been validated from various viewpoints, there remains room for discussion of the false positives, which is when a respondent answers a question correctly without understanding the physics concept being tested in the question. For instance, false positives will appear 20% of the time when a respondent chooses his or her answer randomly, since the FCI is a five-choice survey. Although they introduced powerful distractors to reduce false positives [6], Hestenes *et al.* found that false positives were nevertheless "fairly common" in interviews of students [1]. Several

validation studies identified that question 16 (Q.16) is particularly prone to false positives [9–17]. Specifically, a number of students get the correct answer to Q.16 by incorrectly applying Newton's first law (e.g., in the case of Ref. [11], 8 in 10 correct responses), although Q.16 should be solved with Newton's third law (see below for a description of the problem). Although they did not explicitly specify Q.16, Hestenes *et al.* also mentioned that some students in their study confused the balance of forces on a single object with the equal and opposite forces on different objects in an interacting pair [1].

False positives are a source of systematic error on the results of the FCI. Systematic errors are errors associated with observation or instrument inaccuracy that are not random and, as such, push experimental results in a consistent direction [18]. The systematic error from false positives tends towards scores that are higher than what would be measured without this error. Hake described false positives as one of five sources of systematic errors of the FCI, but he did not examine them in detail [3]. Similarly, although Hestenes *et al.* addressed false positives with the statement "except possibly for high scores (say, above 80), the Inventory score should be regarded as an upper bound on a student's Newtonian understanding" [1], they did not examine *how much* the "Newtonian understanding" is below the upper bound. Wang and Bao analyzed the FCI with item response theory (IRT) and calculated the guessing parameters for each of the FCI questions [10]. However, they did not use these parameters to calculate systematic error.

We build upon this body of research and present here a method to estimate the systematic error due to false positives. If the systematic error from false positives is sufficiently small relative to, for example, the total score of the FCI, then the effect of false positives can be regarded as negligible and, consequently, ignored. In this paper, however, we show that, for students with middle-range total scores, the size of the systematic error from false positives is larger than 10% of that total score. Although we do not think that this is small enough to ignore, we also do not recommend modification to Q.16 or any other item on the FCI, in part because of the large amount of data that has already been accumulated with it. On the contrary, we think that the benefit of *quantifying* the systematic error as we do in this paper is that it allows for one to correct for it by appropriately reducing student scores *ex post facto*.

The remainder of this paper is organized as follows. In Sec. II, we describe our methods for collecting and analyzing the data. In Sec. III, we describe the results of our data. Finally, in Sec. IV, we summarize this study and we discuss the limitations and validity of our research.

II. METHODOLOGY

A. What are false positives?

Our research project centered upon analysis of a modified version of the FCI that we designed in order to calculate the systematic error from false positives. We begin this section of the paper by elaborating on what is meant by a "false positive." We then describe the design of our survey and our analysis methods.

TABLE I. Contingency table of answers.

		True attribute of a respondent	
		Understanding	Not understanding
FCI Question	Correct	True positive	False positive
	Incorrect	False negative	True negative

We classify a student's response to a particular FCI question as being one of the four options shown in Table I. In classical test theory, it is assumed that each examinee has a "true attribute" [19], which in this case would be Newtonian understanding [1]. When a respondent answers a question correctly while also understanding the physics content being tested in the question, the response is classified as a "true positive." However, if he or she does not understand the content, the answer is considered a "false positive." "True negative" and "false negative" are defined in a similar manner. False positives will appear when a respondent happens to answer correctly by random guessing, when a respondent has seen a similar situation in class and has simply memorized the correct answer without understanding, etc. On the other hand, false negatives will appear, for example, when the wording or diagram of a question is not clear, or when a respondent is careless or inattentive.

Hestenes *et al.* considered false positives and false negatives on the FCI and wrote that, according to interview-based studies, false positives "were fairly common" [1] but that "the probability of a false negative [is] certainly less than ten percent" [6]. Since, in accordance with the findings of Hestenes *et al.*, we expect systematic errors due to false positives to be greater than systematic errors due to false negatives, we focus on false positives in our research. Of course, for greatest accuracy, the systematic error due to false negatives should also be investigated, and future research could attend to this. We presume it would be a smaller effect that would tend to somewhat counter the effect from false positives.

B. Survey design

In order to judge whether the correct answer of a certain respondent is a true positive or a false positive, it is necessary to judge whether the respondent understands the content tested in the question. In this study, we judge that a correct answer is a true positive if the respondent correctly answers a corresponding set of subquestions which we designed and inserted into the survey [13,14]. Many individual questions of the FCI test understanding of several concepts simultaneously. Each individual subquestion is designed to test student understanding on just one of

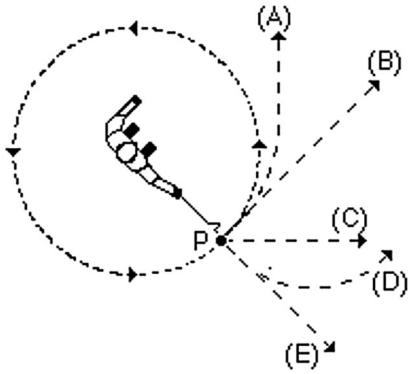

FIG. 1. Question 7 (Q.7) of the FCI [1]. [Question statement] A steel ball is attached to a string and is swung in a circular path in a horizontal plane as illustrated in the accompanying figure. At the point *P* indicated in the figure, the string suddenly breaks near the ball. If these events are observed from directly above as in the figure, which path would the ball most closely follow after the string breaks?

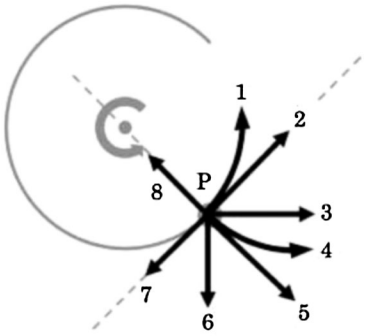

FIG. 2. Subquestions of the FCI Q.7. [Question statement] Which of the eight choices best represents the direction of the following variables, just after the string breaks? If you think a variable is zero or not at all horizontal, write 9; 1. force acting on the ball; 2. velocity of the ball.

the concepts required to answer the corresponding FCI question. As described below, the design of these subquestions is based upon previous student interviews [11], where incorrect reasoning was observed.

As an example, we show Q.7 in Fig. 1, and the corresponding subquestions in Fig. 2. Question 7 asks for the trajectory of a ball after the string breaks, which requires understanding of the force acting on the ball after the string breaks as well as understanding of what the subsequent velocity of the ball would be. The subquestions of Q.7 test for understanding of these concepts individually.

If a respondent answers a certain FCI question correctly and answers all corresponding subquestions correctly, we treat that student's response to be a true positive. On the other hand, if a respondent answers a certain FCI question correctly but answers even one subquestion incorrectly, we treat that student's response to be a false positive. This judgement method is summarized in Table II.

TABLE II. Contingency table of answers with subquestions.

		Subquestions	
		All correct	Not all correct
FCI Question	Correct	True positive	False positive
	Incorrect	False negative	True negative

In order to prevent the survey from becoming too much of a burden for respondents, we created subquestions for only FCI questions 5, 6, 7, and 16. In what follows, we justify our selection of these four questions and describe how we designed the subquestions.

Although false positives exist whenever a student randomly guesses correctly on a given question, we have found that additional false positives arise on Q.6, Q.7, and Q.16 as a result of clearly incorrect reasoning [11]. (The protocol and conditions of the interviews of Ref. [11] are shown in the Appendix.) We refer to these three questions as "FP-reasoning" questions. Since the FP-reasoning questions tend to induce false positives more than other questions, they accordingly introduce relatively large systematic errors. Other than Q.6, Q.7, and Q.16, questions that tend to induce false positives have not been found.

For Q.7 and Q.6 (which is analogous to Q.7), some students revealed in interviews that they had chosen the correct answers for erroneous reasons such as "because the direction of the velocity is the same as the direction of the force acting on the ball after the string breaks." Since Q.6 and Q.7 are designed to evaluate a student's understanding of Newton's first law in situations where no force acts [1,2], we consider a response based upon this reasoning to be a false positive. In accordance with these findings, we created two subquestions for each of these FCI questions, one asking about the force on the ball and the other asking about the velocity of the ball, as shown in Fig. 2. For Q.16, in which a car pushes a truck at constant speed, some interviewees chose the correct answer that the forces are equal for erroneous reasons such as "the forces are balanced because both vehicles are moving at a constant speed." Based upon this, we created two subquestions for Q.16, one asking about the *reaction force* to the car's force on the truck and the other asking about the force that *balances* that pushing force (see Supplemental Material in Ref. [20] for all subquestions).

Again, other than Q.6, Q.7, and Q.16, there have been no cases found of students answering FCI questions correctly with clearly erroneous reasoning (that is, FP reasoning consists of *only* those three questions) [11]. Nevertheless, we considered that students might still obtain false positives on the other 27 FCI questions by, for example, some form of guessing. To contrast these questions with the FP-reasoning questions, we use the term "FP guessing" in this paper. We want to compare the systematic error from the FP-reasoning questions with that from the FP-guessing questions, and for this we created subquestions also for Q.5.

We specifically chose Q.5 (instead of one of the other 26 FP-guessing questions) because we could create subquestions for it in a straightforward manner. Question 5 has students select one out of five combinations of four forces acting on a ball traveling in a passage. Therefore, we made one subquestion for each possible force listed in the original FCI question, with each subquestion independently asking whether or not that force was acting on the ball.

In summary, the survey instrument used in this study consisted of a total of 40 questions, the 30 original Japanese-language FCI questions [21], 4 subquestions for Q.5, 2 subquestions for Q.6, 2 subquestions for Q.7, and 2 subquestions for Q.16. So as to minimize hints given by the subquestions for the FCI questions, we placed the subquestions of Q.6, Q.7, and Q.16 after the 30 FCI questions. Conversely, to avoid having Q.5 give hints for its subquestions, the Q.5 subquestions were placed in front of the 30 FCI questions. In addition, students were instructed not to return to previous questions. We checked the clarity of the wording and diagrams of the subquestions by interviewing a few students to confirm that they understood the intent of the questions.

C. Data collection

We surveyed students at the beginning of introductory physics courses at one public university and three private universities in Japan in April 2015 and April 2017. These four universities are middle-rank universities in Japan. The total number of survey responses was 1145. From this, we excluded the responses of students who did not answer some of the questions, who wrote a letter which was not one of the choices available for a given question, or who wrote the same or serial letters continuously. In total, the number of valid responses was 1110. Most of the respondents were first-year students. Students were from different departments, mainly the departments of science, technology, and agriculture. The survey was conducted during class, so as to help ensure that students would concentrate on the survey. The respondents were not given any incentive to participate (in the form of money or extra credit).

D. Analysis method

In this paper, we calculate the systematic error of the FCI by subtracting the "true score" from the "raw score." The raw score is what the FCI measures directly it is the number of correct answers ("raw positives"). The true score in this study is taken to be the number of true positives (see Table II). According to classical test theory [19], if we represent the raw score as a random variable S_{raw} and the true score as a random variable S_{true}, the relationship between S_{raw} and S_{true} is written as

$$S_{\mathrm{raw}} = S_{\mathrm{true}} + E_{\mathrm{sys}} + E_{\mathrm{stat}}, \tag{1}$$

where E_{sys} is a random variable which represents systematic error and E_{stat} is a random variable which represents statistical error. The systematic error E_{sys} includes false positives and false negatives and can generally be a function of the raw score S_{raw}. For example, since high-performing students are less likely to get false positives (particularly on Q.6, Q.7, and Q.16), their systematic error is smaller. The statistical error E_{stat} includes errors from the motivation and attention of respondents, and from the environment of the classroom.

Classical test theory assumes the behavior of E_{stat} to be random, with the expected value $\langle E_{\mathrm{stat}} \rangle$ equal to zero [19]. Therefore, the expected value of Eq. (1) can be written as

$$\langle S_{\mathrm{raw}} \rangle = \langle S_{\mathrm{true}} \rangle + \langle E_{\mathrm{sys}} \rangle. \tag{2}$$

We assume that the systematic error from false positives is dominant and neglect false negatives. Therefore, we rewrite Eq. (2) as

$$\langle S_{\mathrm{raw}} \rangle = \langle S_{\mathrm{true}} \rangle + \langle E_{\mathrm{fp}} \rangle, \tag{3}$$

where E_{fp} is the systematic error from false positives.

Each question on the FCI has its own set of Eq. (1)–(3). We define the random variable of raw score, true score, and systematic error of the ith question on the FCI as S^i_{raw}, S^i_{true}, and E^i_{fp}, respectively. Each random variable can be either 0 or 1. For example, $S^i_{\mathrm{raw}} = 1$ if a student's response to the ith question is a raw positive and $S^i_{\mathrm{raw}} = 0$ if not. If a student's response to the ith question is a false positive, $S^i_{\mathrm{raw}} = 1$ but $S^i_{\mathrm{true}} = 0$ and so $E^i_{\mathrm{fp}} = 1$. With these notations and from Eq. (3), it follows

$$\langle S^i_{\mathrm{raw}} \rangle = \langle S^i_{\mathrm{true}} \rangle + \langle E^i_{\mathrm{fp}} \rangle. \tag{4}$$

We define the probability that each random variable takes 1 as P^i_{raw}, P^i_{true}, and P^i_{fp}. For example, P^i_{raw} is the probability that a student answers the ith question of the FCI correctly; in other words, the probability that a student's ith response is a raw positive. Since each random variable S^i_{raw}, S^i_{true}, and E^i_{fp} follows a Bernoulli distribution, it follows that $\langle S^i_{\mathrm{raw}} \rangle = P^i_{\mathrm{raw}}$, $\langle S^i_{\mathrm{true}} \rangle = P^i_{\mathrm{true}}$, and $\langle E^i_{\mathrm{fp}} \rangle = P^i_{\mathrm{fp}}$. Therefore, from Eq. (4),

$$P^i_{\mathrm{raw}} = P^i_{\mathrm{true}} + P^i_{\mathrm{fp}}. \tag{5}$$

Note that since $\sum_{i=1}^{30} P^i_{\mathrm{raw}} = \langle S_{\mathrm{raw}} \rangle$, $\sum_{i=1}^{30} P^i_{\mathrm{true}} = \langle S_{\mathrm{true}} \rangle$, and $\sum_{i=1}^{30} P^i_{\mathrm{fp}} = \langle E_{\mathrm{fp}} \rangle$, if we sum up Eq. (5) for all of the 30 questions of the FCI, we obtain the same equation as Eq. (3). Like with the total systematic error due to false positives, E_{fp}, we expect the probabilities P^i_{raw}, P^i_{true}, and P^i_{fp} to depend upon the total raw score, S_{raw}, and we will explicitly show the dependences as, for example, $P^i_{\mathrm{fp}}(S_{\mathrm{raw}})$, as needed for clarity. Note that this approach is subtly

different from that of our prior work [22], in that we did not calculate the systematic error from individual questions.

With our new approach, we turn our attention to the following tasks. First, we will analyze the behavior of the systematic errors from false positives on Q.5, Q.6, Q.7, and Q.16 using Eq. (5) and check for consistency between the results of this study and those of a prior study [11]. In this analysis, $P^i_{\text{raw}}(S_{\text{raw}})$ and $P^i_{\text{true}}(S_{\text{raw}})$ for each of the 31 possible values of S_{raw} $(0, 1, 2, \ldots, 30)$ are estimated using the following equations:

$$P^i_{\text{raw}}(S_{\text{raw}}) = \frac{N^i_{\text{raw}}(S_{\text{raw}})}{N(S_{\text{raw}})}, \tag{6}$$

$$P^i_{\text{true}}(S_{\text{raw}}) = \frac{N^i_{\text{true}}(S_{\text{raw}})}{N(S_{\text{raw}})}, \tag{7}$$

where $N(S_{\text{raw}})$ is the number of students with a given S_{raw}, $N^i_{\text{raw}}(S_{\text{raw}})$ is the number of those students who answered the ith question correctly (number of raw positives), and $N^i_{\text{true}}(S_{\text{raw}})$ is the number of *those* students who also answered the subquestions of that ith question correctly (number of true positives). Although the notation is different, the graphs of P^i_{raw} vs S_{raw} are equivalent to item response curves of the correct answers to the FCI questions [23,24], and so we will compare our results with the findings of this prior research. After calculating $P^i_{\text{raw}}(S_{\text{raw}})$ and $P^i_{\text{true}}(S_{\text{raw}})$, $P^i_{\text{fp}}(S_{\text{raw}})$ is calculated using Eq. (5). We will analyze the dependence of P^i_{raw}, P^i_{true}, and P^i_{fp} on S_{raw} and compare the size of P^i_{fp}—that is, the systematic errors—between the FP-guessing question (Q.5) and FP-reasoning questions (Q.6, Q.7, and Q.16).

Second, we will estimate a minimum value for the systematic error due to false positives on the entire FCI. We do this by first calculating the sum of the systematic errors $P^i_{\text{fp}}(S_{\text{raw}})$ for $i = 5$, 6, 7, and 16 and then examining its size relative to the raw score. As an equation, this ratio $R_{\text{fp}}(S_{\text{raw}})$ is calculated by

$$R_{\text{fp}}(S_{\text{raw}}) = \frac{\sum_{i=5,6,7,16} P^i_{\text{fp}}(S_{\text{raw}})}{S_{\text{raw}}}. \tag{8}$$

As already discussed, the total systematic error from false positives is the sum of the systematic errors P^i_{fp} from all 30 questions on the FCI. We will show that the sum of a subset of systematic errors, namely, from Q.5, Q.6, Q.7, and Q.16, is about 10% of the raw score in the middle score range. This automatically implies that the total systematic error must be even larger.

III. RESULTS

A. Basic statistics

The summary of data from all valid responses is presented in Table III. Taken together, the distribution of the respondents is broad. Figure 3 represents the distribution of raw scores of the respondents. Note that since the number of respondents is low near a score of zero, the analysis is not so accurate in this range. Therefore, we will focus on the scores that are sufficiently far from zero in the following analysis.

TABLE III. Summary of basic statistics from each university (2015 + 2017 data). "N" is the number of respondents and "SD" is the standard deviation.

Samples	N	Mean	SD
U1	222	53.9%	22.6%
U2	209	36.8%	14.0%
U3	366	68.2%	18.8%
U4	313	46.6%	20.0%
Population mean		53.3%	
Population SD		22.5%	

B. Systematic errors of Q.5, Q.6, Q.7, and Q.16

In Fig. 4, P^i_{raw} and P^i_{true} of Q.5, Q.6, Q7, and Q.16 are plotted as a function of S_{raw}. P^i_{raw} and P^i_{true} are calculated using Eq. (6) and (7) for each of the 31 possible values of S_{raw}. Since in our survey there is no respondent with $S_{\text{raw}} = 0$ or 1, corresponding P^i_{raw} and P^i_{true} values are not shown in the graph. The error bars indicate Clopper-Pearson 95% confidence intervals [25,26] and are gray with caps for P^i_{raw} and black without caps for P^i_{true}. From the graphs, we can see the tendency that as S_{raw} increases, P^i_{raw} and P^i_{true} also increase as we expected. Note that the P^i_{raw} vs S_{raw} graphs of Q.5, Q.6, Q7, and Q.16 show no large discrepancies with the graphs of the correct responses of

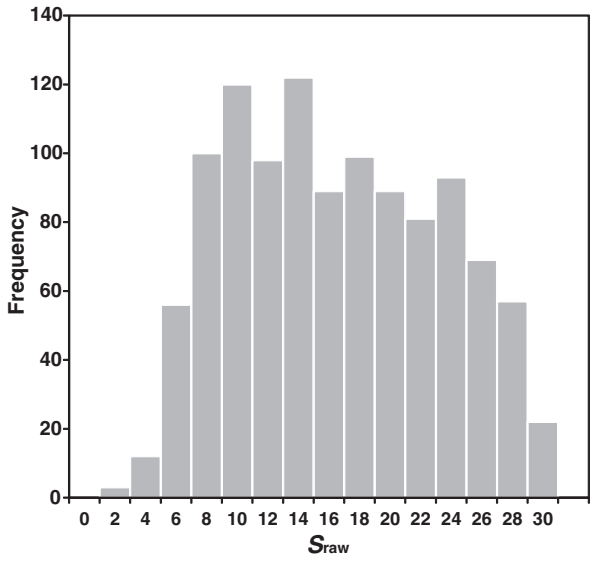

FIG. 3. Distribution of raw scores on the FCI ($N = 1110$). The values on the horizontal axis designate the highest score in a given bin.

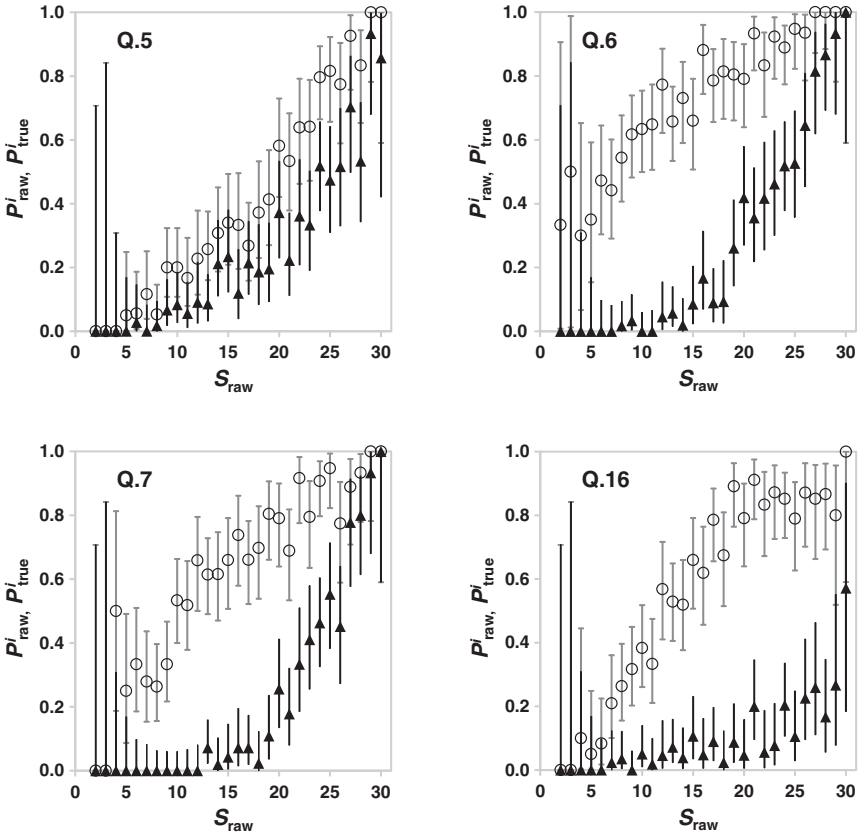

FIG. 4. P^i_{raw} (vertical axis, white circle) and P^i_{true} (vertical axis, black triangle) of Q.5, 6, 7, and 16 for each group of students with a given S_{raw} (horizontal axis). The error bars (P^i_{raw}, gray with caps; P^i_{true}, black without caps) indicate the Clopper-Pearson 95% confidence intervals of the dependent variable.

the item characteristic curves of Q.5, Q.6, Q7, and Q.16 presented in Refs. [23,24].

The systematic error of each question corresponds to the difference between $P^i_{raw}(S_{raw})$ and $P^i_{true}(S_{raw})$. In Fig. 5, P^i_{fp} of Q.5, Q.6, Q7, and Q.16 are plotted as a function of S_{raw}. Note that the systematic error for all four questions in the low or high score range is smaller than that in the middle score range. This behavior makes sense because students with high scores are more likely to understand mechanics well and their P^i_{true} should correspondingly be high as well. On the other hand, for the low score range, P^i_{true} reaches P^i_{raw} because students who do not answer the original FCI question correctly are, by definition, unable to get a true positive on that question. In the limit of a student with an S_{raw} of zero, S_{true} on all questions is also zero. Note also that the systematic errors of FP-reasoning questions (Q.6, Q.7, and Q.16) are much larger than that of the FP-guessing question (Q.5) in the middle score range. For example, when $S_{raw} = 15$, $P^5_{fp} = 0.11$ but $P^6_{fp} = 0.57$, $P^7_{fp} = 0.62$ and $P^{16}_{fp} = 0.55$. This finding is consistent with the results of our previous interview study [11] where we found respondents answering correctly with clearly erroneous reasoning for Q.6, Q.7, and Q.16, but we did not for Q.5.

C. Combined effect of systematic errors

In the previous section, we calculated the systematic errors from false positives of Q.5, Q.6, Q.7, and Q.16. In this section, we consider their combined effect. We analyze R_{fp}, which is defined as, for a given raw score, the ratio of the sum of the systematic errors of Q.5, Q.6, Q.7, and Q.16 to that raw score [Eq. (8)]. Figure 6 shows the dependency of R_{fp} upon S_{raw}. The error bars indicate the square-and-add intervals of the dependent variable for each question [26–28], which are calculated by the following equations for each S_{raw}:

$$\delta R^L_{fp} = \sqrt{(R^{5L}_{fp})^2 + (R^{6L}_{fp})^2 + (R^{7L}_{fp})^2 + (R^{16L}_{fp})^2}, \quad (9)$$

$$\delta R^U_{fp} = \sqrt{(R^{5U}_{fp})^2 + (R^{6U}_{fp})^2 + (R^{7U}_{fp})^2 + (R^{16U}_{fp})^2}, \quad (10)$$

where δR^L_{fp} (δR^U_{fp}) is the lower (upper) error of R_{fp} and δR^{iL}_{fp} (δR^{iU}_{fp}) is the lower (upper) Clopper-Pearson 95% error of P^i_{fp}/S_{raw}.

In the graph, for $S_{raw} < 10$, the error bars are so large that any trend is obscured; however, for $10 \leq S_{raw} < 20$, we can see that R_{fp} is roughly constant as the raw score increases,

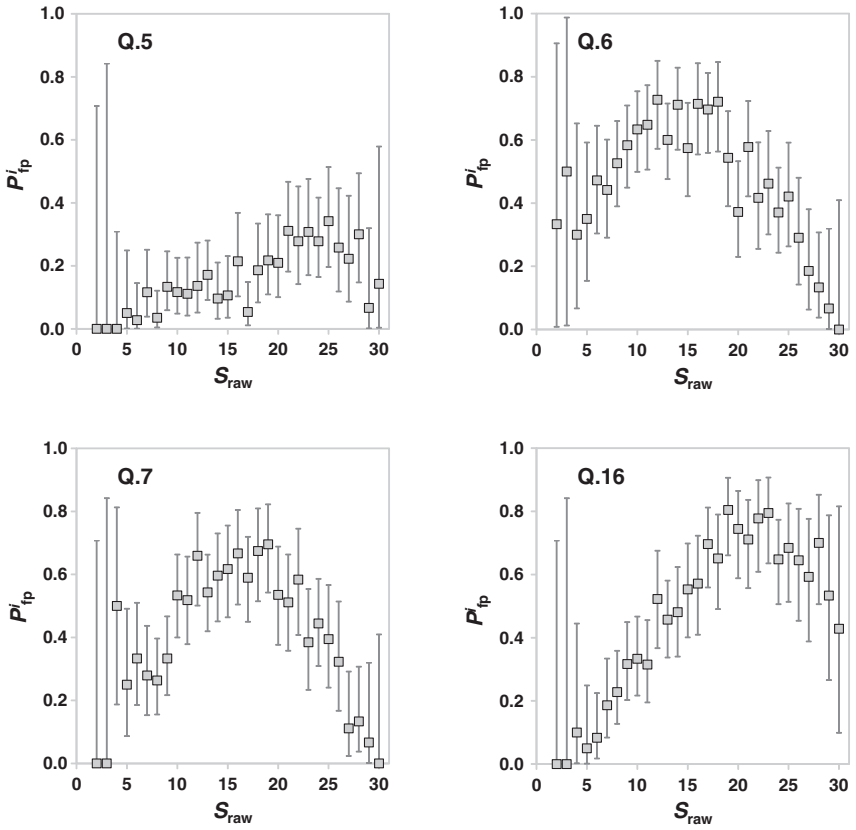

FIG. 5. P^i_{fp} (vertical axis) of Q.5, 6, 7, and 16 for each group of students with a given S_{raw} (horizontal axis). The error bars indicate the Clopper-Pearson 95% confidence intervals of the dependent variable.

beginning at $R_{fp} \sim 0.16$ at $S_{raw} = 10$ and dropping only marginally to $R_{fp} \sim 0.12$ at $S_{raw} = 19$. Beyond this range, R_{fp} decreases more rapidly as S_{raw} increases to a final value of $R_{fp} \sim 0.02$ when $S_{raw} = 30$. This result that the systematic error from false positives becomes smaller for high raw scores is consistent with the statement from Hestenes *et al.*

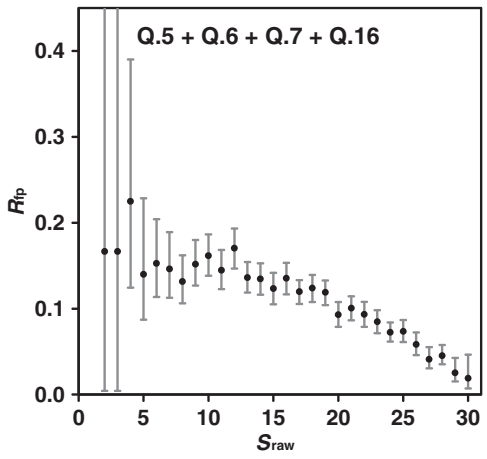

FIG. 6. The ratio of the sum of the systematic errors of Q.5, Q.6, Q.7, and Q.16 to a given raw score S_{raw} for each group of students with that S_{raw}. The error bars indicate the square-and-add intervals of the dependent variable for each question.

[1] that the gap between a student's Newtonian understanding and inventory score is smaller for students with a raw score more than 80%. Regarding students who do not have such high scores, we can say that at least in the middle score range, $10 \leq S_{raw} < 20$, the size of the systematic error from false positives on these four questions combined is roughly 10% of the raw score. It follows that the total systematic error from false positives on all 30 FCI questions (of which Q. 5, Q. 6, Q. 7, and Q. 16 is but a subset) must be even larger than this. As a concrete example of what this implies for true score, on average, a student with a raw score of 15 has more than 1.5 of those points coming from false positives and so has a true score of less than 13.5.

IV. DISCUSSION

A. Summary

In this study, we analyzed the systematic errors from false positives (which occur when a student answers a question correctly without understanding the content tested by the question) of Q.5, Q.6, Q.7, and Q.16 of the FCI. We determined whether or not a correct response to a given FCI question was a false positive by using subquestions. In our previous interview study, we had found respondents answering Q.6, Q.7, and Q.16 correctly with clearly erroneous reasoning. In this paper, we compared the

systematic error from these three questions with the systematic error from Q.5, for which no clearly erroneous reasoning had been found. From the graphs of P^i_{fp} (that is, systematic error due to false positives) vs S_{raw}, we found that the systematic errors of Q.6, Q.7, and Q.16 are much larger than that of Q.5 for most raw scores, as we expected. We also found that the sum of the systematic errors of Q.5, Q.6, Q.7, and Q.16 for a given raw score is about one-tenth of that raw score in the middle score range. It follows that the total systematic error from false positives on all 30 FCI questions must be even larger than this.

Nevertheless, we do not presently propose modification to these questions, because (i) a vast body of data has already been accumulated with the current version of the FCI, and (ii) as we describe below, it is possible that systematic errors have little influence on the test statistics used with the FCI. If that is the case, then the effect of false positives can be regarded as negligible and, consequently, ignored. Rather, our approach is to quantify the systematic error so as to correct for it by subtracting it from the raw score, bringing us to a closer estimate of a student's true score.

Our research is meaningful in that it is the first presentation of a method to estimate the systematic error due to false positives on the FCI. Although IRT, as described above, is similar in some aspects to our approach, it has not been used to estimate the size of the systematic error arising from false positives on the FCI. To do so requires teasing apart correct answers that arise from false positives from those that arise from true positives, as we have demonstrated in this paper. We feel that IRT in its current form is insufficient to make such a distinction.

B. Limitations and future work

It should be noted that we, like many test theorists, do not presume that we can know the value of an individual student's true attribute of understanding with certainty. Although our aim is to approximate and correct for the systematic error of the score as much as possible, we do not expect to reach a point where we can know the size of that error without uncertainty. Moreover, even if the systematic error were known perfectly, there would still be statistical error. We adopt from classical test theory the assumption that the statistical error behaves like a random variable which averages out to zero across a large group of subjects [19]. However, this assumption is axiomatic and may be unfounded. Depending upon this assumption is an inherent limitation to our research.

Regarding the validity of the subquestions themselves, we ensured the clarity of the wording and diagrams of the subquestions by interviewing a few students to confirm that they understood the intent. Nevertheless, to increase the accuracy with which we measure systematic error, there is work to be done in further validating the subquestions. Consider, however, the case of a respondent answering a given FCI question and its corresponding subquestions correctly *but with erroneous reasoning*. Although this would be a false positive *on the subquestions* (and, consequently, on the original FCI question itself), our coding scheme would *misjudge* that positive to be a true positive. This will inflate P^i_{true}. In other words, if these false positives of subquestions in our data were to be accounted for, the correction would lower our estimation of P^i_{true} and increase our estimation of the systematic error P^i_{fp}. That is, *ignorance* of false positives on the subquestions makes our estimation of systematic error (10% of the raw score) all the more conservative. Of course, it is possible for a respondent to get a false negative on the subquestions as well, but we suspect this is much less frequent, just as false negatives are much less frequent than false positives on the original FCI questions.

Although we have determined by creating subquestions to four questions on the FCI that R_{fp} for the FCI as a whole is greater than 0.10 in the middle score range, we cannot say accurately *how much* larger it is without the creation and administration of subquestions for the remaining 26 FCI questions. Also, systematic errors from false negatives should be studied, as they would offset the effect of false positives, although, again, we expect that it would be a smaller effect.

The FCI is used primarily for assessments of physics courses. As such, correcting the raw score for systematic errors is important mainly as the first step towards considering if and to what degree these errors influence the statistics that are used with the test. Even if the systematic error has a large effect on the raw score, in the process of calculating the difference of the scores, the systematic errors may be canceled [3]. For example, Stewart and Stewart [29] showed that test score is substantially affected by guessing but, so long as the true score has a linear dependency on raw score, the normalized gain is unchanged. Our data (Fig. 4), however, suggest that there is no reason to assume such a linear dependency of true score to raw score, and so we feel that work remains to determine if and how normalized gain is affected by systematic errors from false positives. It is necessary to also consider if and how systematic error affects test statistics (e.g., t statistics) and statistics that measure effect size (e.g., Cohen's d) [30].

Using our model, one can study the relationships between the responses of the FCI questions as some other researchers have done [12,17,31]. To be clear, as we described below Eq. (5), the probability of obtaining a true positive on a given question, P^i_{true}, depends on the raw score, S_{raw} (for example, we expect students with higher S_{raw} to also have higher P^i_{true} values). This S_{raw} is the sum of the raw scores for each question S^j_{raw}, and in this sense our model does include interactions between the responses of the FCI questions, albeit implicitly. To analyze dependencies explicitly, S_{raw} can be decomposed into its constituent questions. From there, one can calculate P^i_{true} for students

who did and who did not answer the jth question correctly (S^j_{raw} equals 1 or 0, respectively).

Finally, there is the question of how generalizable our results are. Although the data in this paper comes exclusively from four universities in Japan, we hope to administer the survey in the United States and other countries. To that effect, we welcome educators interested in administering our modified FCI to contact the first author.

ACKNOWLEDGMENTS

The authors thank Tiffany-Rose Sikorski for her insightful comments on our work. This work was supported by JSPS KAKENHI Grant No. JP16K00948.

APPENDIX: PROTOCOL AND CONDITIONS OF OUR INTERVIEWS

The subquestions created and inserted into the modified FCI were based upon false positives discovered during validation interviews of the Japanese FCI. As discussed in Ref. [11], we designed the protocol of these interviews with the goal of investigating the reasoning of students when completing the FCI. We conducted semistructured interviews with a total of 16 students from 3 universities in Japan from the middle of July until the middle of September in 2010. Although we did the interviewing ourselves in most of the cases, on occasion we asked our graduate students to conduct the interview instead. We selectively interviewed only students who had studied mechanics previously, either in high school or at the university. Although some of the interviewees were our students, they were not given extra credit for participating in the interview. All interviewees were given a small amount of financial compensation. We interviewed the students one by one in the classroom or the seminar room in our respective universities for one or two hours. We explained to the examinees the purpose of the research, the range of disclosure, and the treatment of the personal information before the interview. We assured them that the interviews would have no effect on their course grades. Then, we confirmed consent from the examinees. Given that consent, we recorded the interviews with an IC recorder or a video camera to preserve the statements of the interviewees accurately.

The procedure of our interview was as follows. First, we asked the examinees to complete the Japanese version of the FCI within 30 minutes. We instructed them to read carefully the statements of the questions in order to decrease careless mistakes. We then asked the examinees the reason why they chose their answers to each question. If the interviewer did not feel that the interviewee's explanation was sufficient, he created and asked follow-up questions in real time. We instructed the examinees to think out loud in order for us to better understand their thinking process. We refrained from giving the examinees correct answers during the interview, as we considered that they could serve as hints to the subsequent questions.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).
[2] The latest version of the FCI revised in 1995 and the revised Tables of Ref. [1] are available at http://modeling.asu.edu/R&E/Research.html (Retrieved 11/11/2017).
[3] 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).
[4] E. F. Redish, J. M. Saul, and R. N. Steinberg, On the effectiveness of active-engagement microcomputer-based laboratories, Am. J. Phys. **65**, 45 (1997).
[5] E. F. Redish, *Teaching Physics with the Physics Suite* (John Wiley & Sons, Hoboken, NJ, 2003), p. 96.
[6] D. Hestenes and I. Halloun, Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).
[7] J. Stewart, H. Griffin, and G. Stewart, Context sensitivity in the force concept inventory, Phys. Rev. ST Phys. Educ. Res. **3**, 010102 (2007).
[8] S. DeVore, J. Stewart, and G. Stewart, Examining the effects of testwiseness in conceptual physics evaluations, Phys. Rev. Phys. Educ. Res. **12**, 020138 (2016).
[9] 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).
[10] J. Wang and L. Bao, Analyzing force concept inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).
[11] J.-i. Yasuda, H. Uematsu, and H. Nitta, Validating a Japanese version of Force Concept Inventory, J. Phys. Educ. Soc. Jpn. **59**, 90 (2011).
[12] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, Phys. Rev. ST Phys. Educ. Res. **8**, 020105 (2012).
[13] J.-i. Yasuda and M.-a. Taniguchi, Validating two questions in the Force Concept Inventory with subquestions, Phys. Rev. ST Phys. Educ. Res. **9**, 010113 (2013).
[14] M.-a. Taniguchi and J.-i. Yasuda, Quantitative validation of Japanese translation of Force Concept Inventory using subquestions, J. Phys. Educ. Soc. Jpn **62**, 226 (2014).
[15] K. F. Wilson and D. J. Low, On second thoughts…": Changes of mind as an indication of competing knowledge structures, Am. J. Phys. **83**, 802 (2015).

[16] D. J. Low and K. F. Wilson, The role of competing knowledge structures in undermining learning: Newton's second and third laws, Am. J. Phys. **85**, 54 (2017).
[17] T. F. Scott and D. Schumayer, Conceptual coherence of non-Newtonian worldviews in Force Concept Inventory data, Phys. Rev. Phys. Educ. Res. **13**, 010126 (2017).
[18] R. Taylor and John, *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, 1st ed. (University Science Books, Mill Valley, CA, 1982), p. 82.
[19] R. J. Gregory, *Psychological Testing History, Principles, and Applications*, 7th ed. (Pearson, London, 2014), pp. 100–102.
[20] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.14.010112 for all subquestions.
[21] M. Ishimoto, H. Uematsu, K. Tsukamoto, H. Nitta, and H. Lang, https://www.physport.org/assessments/ (Retrieved 11/11/2017).
[22] M. M. Hull, J.-i. Yasuda, M.-a. Taniguchi, and N. Mae (to be published).
[23] 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).
[24] 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).
[25] C. J. Clopper and E. S. Pearson, The use of confidence or fiducial limits illustrated in the case of the binomial, Biometrika **26**, 404 (1934).
[26] R. G. Newcombe, *Confidence Intervals for Proportions and Related Measures of Effect Size*, 1st ed. (CRC Press, Boca Raton, FL, 2012).
[27] T. Fagan, Exact 95% confidence intervals for differences in binomial proportions, Computers in Biology and Medicine **29**, 83 (1999).
[28] G. Y. Zou, On the estimation of additive interaction by use of the four-by-two table and beyond, American Journal of Epidemiology **168**, 212 (2008).
[29] J. Stewart and G. Stewart, Correcting the normalized gain for guessing, Phys. Teach. **48**, 194 (2010).
[30] R. E. Kirk, *Statistics: An Introduction*, 5th ed. (Cengage Learning, Boston, MA, 2007).
[31] E. Brewe, J. Bruun, and I. G. Bearden, Using module analysis for multiple choice responses: A new method applied to Force Concept Inventory data, Phys. Rev. Phys. Educ. Res. **12**, 020131 (2016).