Towards quantification of the FCI's validity: the effect of false positives

Michael M. Hull,¹ Jun-ichiro Yasuda,² Masa-aki Taniguchi,³ and Naohiro Mae⁴

¹Department of Physical Sciences and Mathematics, Wayne State College, 1111 Main St, Wayne, NE, 68787

²Institute of Arts and Sciences, Yamagata University, Yamagata, Yamagata 990-8560, Japan

³Center for Teacher Education, Meijo University, Nagoya, Aichi 468-8502, Japan

⁴Center for Development of Instructional System of Science and Engineering, Kansai University, Suita, Osaka 564-8680, Japan

We are interested in quantifying the systematic error of the Force Concept Inventory (FCI). A modified version of the FCI was administered to 500 university students in Japan in 2015. In addition to the 30 original questions, subquestions were introduced for three questions that, according to prior research, elicit false positives from students (6, 7, and 16) as well as for question 5. Using logistic regression with the results of question 5 and its subquestions, we estimate the systematic error arising from the remaining 26 questions. Our results indicate that FCI true score can be less than half of the FCI raw score for Japanese students.

I. INTRODUCTION

When it comes to assessing student understanding of Newtonian mechanics, the first survey that comes to the minds of many instructors and education researchers is the Force Concept Inventory (FCI) [1]. The multiple choice instrument was first published in *The Physics Teacher* in 1992 and has since been translated into over 25 languages to be used internationally. Perhaps in part because it is so widely-used, legitimate concerns and criticisms about the FCI have been expressed. For example, question 16 of the FCI (Q.16) presents students with the situation of a car pushing a truck at constant speed and asks how the force from the truck on the car compares with the force from the car on the truck. The correct answer is "(A) the amount of force with which the car pushes on the truck is equal to that with which the truck pushes back on the car", and the correct explanation is that one force is the reaction of the other, and by Newton's Third Law, is hence of equal magnitude. However, some interviewed students chose option A for the wrong reasoning: *since they are moving at a constant speed, the force from the truck on the car must cancel out the force from the car on the truck* [2]. Since the two forces act on different bodies, this is an incorrect application of Newton's First Law that is frequently seen with students taking introductory physics. This "false positive"—obtaining the correct answer despite not correctly understanding the content tested by the question—on Q.16 was observed by the FCI creators as well as other researchers validating the FCI [3–5]. False positives are a source of systematic error, inflating student scores on the FCI. Regarding this issue, the FCI authors have acknowledged that "Newtonian choices for non-Newtonian reasons were fairly common. Therefore... the Inventory score should be regarded as an upper bound on a student's Newtonian understanding." [1]. They do not, however, quantify to what degree the Inventory score (described in this paper as "raw score") is higher than a student's Newtonian understanding ("true score"). Our research aims to quantify that difference.

Much research has explored the validity of the FCI, including its wording and diagrams [1], its distractors [6, 7], and whether or not the specific question content influences how

students answer [8]. Item response theory has been used to quantify the tendency of students to get an individual question correct, as a function of their overall understanding. For example, it was found that students of little understanding still got Q.16 correct [9]. However, IRT has not been used to estimate student true score given raw score—indeed, it might not be suitable for such a task. We build upon this body of research and present here the first published report of a method to estimate a student's true score given raw score.

II. METHODOLOGY

A. Subquestions

Previous validation studies of the Japanese version of the FCI showed that, similar to the studies mentioned above, some students answer Q.16 correctly by inappropriately utilizing Newton's First Law instead of Third. Furthermore, it was also observed that some students answered Q.6 and Q.7 correctly with incorrect reasoning as well [2]. Question 7 involves a ball being swung around in a circle in a horizontal plane. Students must choose the direction that the ball will travel when the string breaks. The correct answer is (B), which is an arrow drawn tangent to the circle. The correct reasoning involves recognizing that, when the string breaks, there will no longer be a force acting on the ball in the horizontal plane. However, several students who chose the correct answer incorrectly explained that at that particular moment of release, the force being exerted on the ball is in the direction of motion, and that force is what would make it continue to go in a straight path. Question 6 is identical to Q.7, except that a ball is leaving a circular track instead of being on a cut string.

Realizing that the FCI does not allow us to distinguish between students who answer Q.6, Q.7, and Q.16 correctly for correct reasons vs. incorrect reasons or guessing, we wished to introduce a minimum number of additional questions to the FCI that would allow us to make this distinction. To assess whether students were answering Q.16 correctly with the correct reasoning, we introduced two questions to the survey, one

asking what force "balances" the force of the car on the truck (the correct selection being #2, the force of air resistance and friction on the truck) and the other asking what force is the "reaction" to the force of the car on the truck (#1, the force of the truck pushing on the car)[10]. Students who answer Q.16 correctly with the incorrect reasoning detected in interviews should incorrectly answer these "subquestions". In particular, we would expect them to select #1 for the first subquestion, thinking that the force of the truck pushing on the car balances the force of the car on the truck. Students who answered Q.16 correctly, but answered at least one subquestion incorrectly, were coded as false positives for Q.16. Students who answered Q.16 correctly and got both subquestions correct were coded as true positives for Q.16. A similar process and analysis was conducted for Q.6 and Q.7, using three subquestions instead of two (one asking about the force on the ball, one asking about the ball's acceleration, and one asking about the ball's velocity). Only students who answered all three subquestions correctly were coded as true positives.

We did not observe false positives for the remaining 27 FCI questions in validation interviews and so we consider them, in contrast to the other 3 questions, to be "valid-like" with comparably small systematic errors. Nevertheless, we considered that false positives could still exist on these questions, and we wished to find a way to estimate the size of their effect as well. Adding subquestions to all remaining FCI questions would make the inventory exceedingly long for students to complete. Thus, as a first step, we selected one question, Q.5, and repeated the process of making subquestions and checking for false positives.

We selected Q.5 to represent the remaining 26 questions because we could create subquestions for it in a straightforward manner and because the guessing parameter is relatively small [9]. Since the guessing parameters of the remaining 26 questions tend to be higher, choosing Q.5 prevents over-estimating the systematic error and thus constitutes a conservative approach.

In Q.5, students must select, among a limited set of options, a plurality of forces acting on a ball traveling in a passage. Therefore, in the subquestions, we broke down the combinations of the plurality of forces and asked students to determine whether each force was working on the ball or not. If a student answered Q.5 correctly by correctly understanding that each of the forces in their selection is acting, they should get the subquestions correct. If, on the other hand, they had guessed randomly, had used process of elimination, etc., then we would expect them to get at least one of the subquestions wrong.

We placed the subquestions in such a way as to avoid giving students hints to the correct answers to the FCI. We put the subquestions for Q.6, 7, and 16 after the 30 FCI questions. On the other hand, because Q.5 subquestions did not contain any cues to the correct answer to the FCI question, we put these subquestions before the 30 FCI questions. Students were instructed to answer questions sequentially on the FCI and to not return to already-answered questions.

We ensured 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. The survey instrument used in this study consisted of a total of 42 questions, the 30 original Japanese-language FCI questions [11], and a total of 12 subquestions (4 for Q.5, 3 for Q.6, 3 for Q.7, and 2 for Q.16).

B. 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. These four universities are middle-rank universities in Japan. The total number of survey responses was 513. 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 503.

C. Analysis method

As described above, we are interested in estimating a student's true score on the FCI, given their raw score on the assessment. The raw score, S_{raw} , is what the FCI measures directly—it is the number of correct answers on the 30 FCI questions. The true score, S_{true} , in this study is taken to be the number of true positives, i.e. the questions answered correctly and with correct reasoning. For Q.5, Q.6, Q.7, and Q.16, we directly calculate the number of true positives using subquestions, as described above. For the remaining 26 questions, we estimate the true positives with the method we will soon describe. Before that, however, let us represent the true score for the 30 questions (S_{true} , or, written more explicitly, S_{true}^{30}) of an individual student as,

$$S_{\text{true}}^{30} = S_{\text{true}}^4 + S_{\text{raw}}^{26} \cdot r_{\text{tp}} \quad (1)$$

S_{true}^4 is the true score for the 4 questions (Q.5, Q.6, Q.7, Q.16) for which we created the subquestions. S_{raw}^{26} is the raw score for the 26 questions for which we did not create the subquestions. r_{tp} is the true positive ratio for a respondent, which is defined by

$$r_{\text{tp}} = \frac{S_{\text{true}}^{26}}{S_{\text{raw}}^{26}} \quad (2)$$

where S_{true}^{26} is the true score for the remaining 26 questions, which we are not measuring directly.

Since r_{tp} includes S_{true}^{26} , we cannot calculate this value directly—rather, we estimate it with the method that we now describe. The 26 questions for which estimation is necessary are the questions for which we have not found respondents answering correctly with clearly erroneous reasoning. Therefore, we modelled the tendencies of these 26 questions to induce false positives after Q.5. We can think of r_{tp} of Eq. (2)

as representing the likelihood of a respondent to get true positives on the 26 questions. At the same time, we can also think of the likelihood of respondents to get a true positive on Q.5 to be represented by R_{tp}^{Q5} , where R_{tp}^{Q5} is defined by

$$R_{tp}^{Q5} = \frac{N \text{ true positives for Q.5}}{N \text{ raw positives for Q.5}} \quad (3)$$

where N represents the number of respondents. Note that r_{tp} is for a given respondent whereas R_{tp}^{Q5} is calculated from a group of respondents.

We assume that within a given group of students who all have the same S_{raw}^{26} , the true positive ratio r_{tp} of an individual respondent can be approximated by the calculated value of true positive ratio $\hat{R}_{tp}^{Q5}$ of that student's group. Namely,

$$r_{tp}(S_{raw}^{26}) \sim \hat{R}_{tp}^{Q5}(S_{raw}^{26}) \quad (4)$$

In Eq. (4), we show the dependence on S_{raw}^{26} explicitly, in order to show that we can use the approximation only within each of the 27 groups (0, 1, 2, ... 26) corresponding to the possible values of S_{raw}^{26} . In other words, the S_{raw}^{26} of the individual respondent (l.h.s of Eq. (4)) must be the same as of the group (r.h.s). The $\hat{\wedge}$ added to R_{tp}^{Q5} denotes that we will be using the predicted value obtained by regression analysis, as described below. This is an assumption that requires validation. Future research could attend to this, for example, by introducing subquestions for a random subset of the remaining 26 questions.

III. RESULTS

In Fig. 1, the percentage of students who obtained a true positive on Q.5, R_{tp}^{Q5} , is plotted as a function of S_{raw}^{26} . Note that at higher values of S_{raw}^{26} , the true positive ratio increases, as we would expect. The curve passing through the data, $\hat{R}_{tp}^{Q5}(S_{raw}^{26})$, was obtained by performing weighted logistic regression analysis with SPSS. A Hosmer and Lemeshow test of goodness-of-fit was performed. We found our model's predictions fit the data at an acceptable level, $\chi^2(df = 8, N = 152) = 4.79, p > .05$ [12]. The function is

$$\hat{R}_{tp}^{Q5}(S_{raw}^{26}) = \frac{e^{\beta_0 + \beta_1 S_{raw}^{26}}}{1 + e^{\beta_0 + \beta_1 S_{raw}^{26}}} \quad (5)$$

, where $\beta_0 = -1.30 \pm .50$, $p < .05$; and $\beta_1 = .09 \pm .03$, $p < .05$. We chose to use weighted logistic regression analysis because the dependent variable, coded as either a true positive or a false positive, is binary [13].

Substituting this function for $r_{tp}(S_{raw}^{26})$ in Eq. (1) allows us to calculate values of S_{true}^{30} for each of our data points (recall that we are measuring S_{true}^{4} directly by use of the subquestions designed for those four questions). These values of S_{true}^{30} are plotted in Fig. 2 as a function of that respondent's S_{raw}^{30} . There are multiple values of S_{true}^{30} for a given S_{raw}^{30} because, in part, S_{true}^{4} for a student can take on values 0 – 4.

FIG. 1. R_{tp}^{Q5} for each group of students with a given S_{raw}^{26} . The trendline is $\hat{R}_{tp}^{Q5}(S_{raw}^{26})$. The error bars show the standard error of the dependent variable. The radius of the bubble corresponds to the number of students in each of the 26 groups who answered Q.5 correctly.

FIG. 2. Calculated values of S_{true}^{30} for each respondent vs that respondent's S_{raw}^{30} . The trendline is the estimated $\hat{S}_{true}^{30}(S_{raw}^{30})$ calculated by regression analysis.

We performed one final regression analysis so as to obtain a function from the data shown in Fig. 2 to allow us to estimate a student's true score given his or her raw score. Using SPSS to fit the data to the function

$$\hat{S}_{true}^{30} = S_{raw}^{30} \frac{e^{\beta_0 + \beta_1 S_{raw}^{30}}}{1 + e^{\beta_0 + \beta_1 S_{raw}^{30}}} \quad (6)$$

, we find that $\beta_0 = -1.88 \pm .12$; and $\beta_1 = .097 \pm .006$ for the mean logistic regression coefficients. Notice that the true score tends to be around 50% of the raw score. For example, $S_{true}^{30} \simeq 10$, when $S_{raw}^{30} = 20$.

IV. DISCUSSION AND CONCLUSIONS

In this paper, we described a novel approach to addressing invalidity of the FCI and of surveys in general. Rather than discarding the survey or a particular question on the survey (especially when so much data has already been universally collected!), the effect of statistical errors caused by false positives and the like can be calculated and corrected for after the fact with the method of subquestions.

However, there are still several issues with this method with which we are wrestling. There remains a need to find a systematic process for creating subquestions. Here, we only analyzed the effect of false positives and estimated the reduction of a true value from a raw value, namely, the negative part of the systematic error. It is necessary to analyze the effect of false negatives as well. By considering them, the reduction of the raw value could be eased to some extent. However, we do not expect that the effect of false negatives fully counters the effect of false positives, as Hestenes et al. wrote that

the false negatives are "certainly less than ten percent" [14]. Nevertheless, for a still more accurate measure of a learner's understanding of Newtonian mechanics, subquestions could be created and used in future research to calculate the effect of false negatives.

Despite these limitations, we feel that this method of subquestions is useful for instructors in calculating more accurately the degree to which their students are Newtonian thinkers. Currently our data is restricted to four universities in Japan, and there is need to determine how similar other populations are for a more accurate calculation of true score. We welcome educators interested in administering our modified FCI to contact the first author.

ACKNOWLEDGMENTS

This work is supported in part by KAKENHI 16K00948.

-
- [1] D. Hestenes, M. Wells, and G. Swackhamer, *Phys. Teach.* **30**, 141 (1992).
  - [2] J. Yasuda, H. Uematsu, and H. Nitta, *Phys. Educ. Soc. Jpn.* (in Japanese) **59**, 90 (2011). <http://ci.nii.ac.jp/naid/110008673371>. Retrieved 6/21/2017.
  - [3] J. Yasuda and M. Taniguchi, *Phys. Rev. ST: PER.* **9**, 010113 (2013).
  - [4] T. F. Scott, D. Schumayer, and A. R. Gray, *Phys. Rev. ST: PER.* **8**, 020105 (2012).
  - [5] R. K. Thornton, D. Kuhl, K. Cumings and J. Marx, *Phys. Rev. ST: PER.* **5**, 010105 (2009).
  - [6] N. Rebello and D. Zollman, *Am. J. Phys.* **72**, 116 (2004).
  - [7] G. A. Morris et. al., *Am. J. Phys.* **74**, 449 (2006).
  - [8] J. Stewart, H. Griffin, and G. Stewart, *Phys. Rev. ST: PER.* **3**, 010102 (2007).
  - [9] J. Wang and L. Bao, *Am. J. Phys.* **78**, 1064 (2010).
  - [10] Questions introduced to the FCI are in the appendix: <https://sites.google.com/view/hull2017>
  - [11] M. Ishimoto, H. Uematsu, K. Tsukamoto, H. Nitta, H. Lang (2011). <http://modeling.asu.edu/R&E/Research.html>. Retrieved 6/21/2017.
  - [12] Also, no DFBetas had a magnitude greater than $0.16(2/\sqrt{152})$, where 152 is the number of students answering Q5 correctly, and so all data was included.
  - [13] L. Ding and M. D. Caballero, *Phys. Rev. ST: PER.* **10**, 020125 (2014).
  - [14] D. Hestenes and I. Halloun, *Phys. Teach.* **33**, 502, (1995).