Analyzing the measurement error from false positives 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

Abstract. We analyze the measurement error from false positives of the Force Concept Inventory (FCI) focusing on the four questions (Q.5, Q.6, Q.7, and Q.16). We determine whether or not a correct response to a given FCI question is a false positive using subquestions. Using the data of 1145 university students in Japan from 2015 to 2017, we find that 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 mid-level student. We consider what degree these errors influence the statistics that are used with the test (e.g. *t* statistics).

1 Introduction

The Force Concept Inventory (FCI) is one of the most widely-used tests in physics education [1]. The FCI is a 30-item, five-choice survey to probe student conceptual understanding of Newtonian mechanics. Although the validity of the FCI had been evaluated in various ways, there remains room for discussion of the false positives, which are the responses of answering a question correctly without understanding the physics concept being tested in the question. For example, several validation studies identified that question 16 is particularly prone to false positives [2, 3]. Specifically, a number of students get the correct answer to Q.16 by incorrectly applying Newton's first law, although Q.16 should be solved with Newton's third law. The measurement error from false positives tends towards scores that are higher than what would be measured without this error. 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. We build upon this body of research and present here a method to analyze the measurement error due to false positives.

2 Methodology

We judge that a respondent understands the physics content being tested in the question if the respondent correctly answers a corresponding set of subquestions, which we designed and inserted into the survey [3]. 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 the concepts required to answer the corresponding FCI question. 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

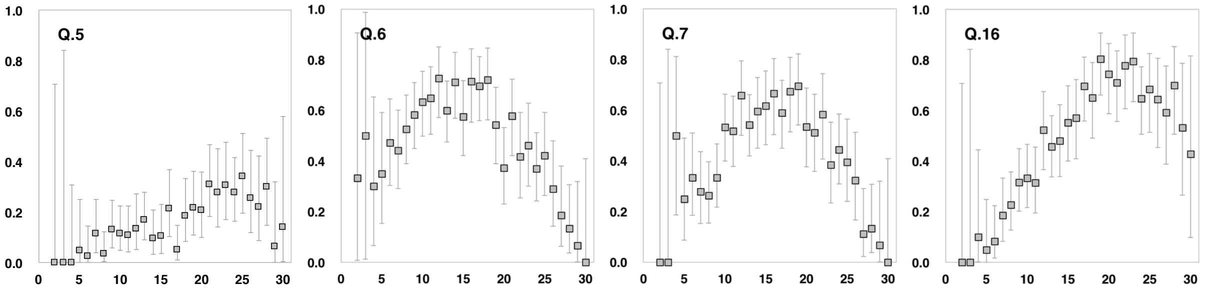

Fig. 1 P_{fp}^{i} (vertical axis) of Q.5, 6, 7, and 16 for each group of students with a given S_{raw} (horizontal axis) [4].

answers a certain FCI question correctly but answers even one subquestion incorrectly, we treat that student's response to be a "false positive."

We created subquestions for only FCI questions 5, 6, 7, and 16, where we had found respondents answering Q.6, Q.7, and Q.16 correctly with clearly erroneous reasoning in our previous interview study [2]. We compared the measurement error from these three questions with the error from Q.5, for which no clearly erroneous reasoning had been found [2]. The survey instrument used in this study consisted of a total of 40 questions, the 30 original Japanese-language FCI questions, 10 subquestions 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.

The measurement error from false positives of each question is defined as follows. First, we define the random variable of raw score, true score, and measurement error of the *i*th question on the FCI as $S_{\mathrm{raw}}^{i}, S_{\mathrm{true}}^{i}$, and E_{fp}^{i}, respectively. Each random variable can be either 0 or 1. For example, $S_{\mathrm{raw}}^{i} = 1$ if a student's response to the *i*th question is a raw positive and $S_{\mathrm{raw}}^{i} = 0$ if not. For their expected values, it follows from classical test theory, $\langle S_{\mathrm{raw}}^{i} \rangle = \langle S_{\mathrm{true}}^{i} \rangle + \langle E_{\mathrm{fp}}^{i} \rangle$. Since each random variable follows a Bernoulli distribution, it follows $P_{\mathrm{raw}}^{i} = P_{\mathrm{true}}^{i} + P_{\mathrm{fp}}^{i}$, where, P_{raw}^{i}, P_{true}^{i}, and P_{fp}^{i} are the probability that a student's *i*th response is a raw positive, true positive, false positive, respectively. The measurement error of each question corresponds to P_{fp}^{i}.

3 Result

In Fig. 1, P_{fp}^{i} of Q.5, Q.6, Q7, and Q.16 estimated from our data are plotted as a function of S_{raw} [4]. The measurement errors of Q.6, Q.7, and Q.16 are much larger than that of the Q.5 in the middle score range. This finding is consistent with the results of our previous interview study [2]. We also consider their combined effect. We analyze R_{fp}, which is defined as, for a given raw score, the ratio of the sum of the measurement errors of Q.5, Q.6, Q.7, and Q.16 to that raw score. We found, for $10 \le S_{\mathrm{raw}} < 20$, R_{fp} is roughly constant as the raw score increases, beginning at $R_{\mathrm{fp}} \sim 0.16$ at $S_{\mathrm{raw}} = 10$ and dropping only marginally to $R_{\mathrm{fp}} \sim 0.12$ at $S_{\mathrm{raw}} = 19$. Beyond this range, R_{fp} decreases more rapidly as S_{raw} increases to a final value of $R_{\mathrm{fp}} \sim 0.02$ when $S_{\mathrm{raw}} = 30$. We can say that at least in the middle score range, the size of the measurement error from false positives on these four questions combined is roughly 10% of the raw score. Although the total systematic error from false positives on all 30 FCI questions must be even larger than this, in the process of calculating the difference of the scores, the measurement errors may be canceled. Therefore, in this study, we also consider what degree these errors influence the statistics that are used with the test (e.g. *t* statistics).

References

[1] D. Hestenes, M. Wells, and G. Swackhamer, Phys. Teach. **30**, (1992) 141.
[2] J. Yasuda, H. Uematsu, and H. Nitta, J. Phys. Educ. Soc. Jpn 59, 90 (2011).
[3] J. Yasuda and M. Taniguchi, Phys. Rev. ST. Phys. Educ. Res. 9, 10113 (2013).
[4] J. Yasuda, N. Mae, M. Hull and M. Taniguchi, Phys. Rev. Phys. Educ. Res. (to be published).