Analyzing the measurement error from false positives in the Force Concept Inventory

J Yasuda¹, N Mae², M M Hull³ and M Taniguchi⁴

¹ Institute of Arts and Sciences, Yamagata University, Yamagata, Yamagata 990-8560, Japan
² Faculty of Engineering Science, Kansai University, Suita, Osaka 564-8680, Japan
³ Austrian Educational Competence Centre Physics, University of Vienna, Vienna, 1090, Austria
⁴ Center for Teacher Education, Meijo University, Nagoya, Aichi 468-8502, Japan

Abstract. We analyze the measurement error due to false positives of the Force Concept Inventory (FCI) focusing on 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 systematic error from the false positives of 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 the error influences the measures of the effectiveness of a course, namely, the average normalized gain and Cohens’ d. Using a set of simulated data, we show that Cohens’ d is less sensitive to the systematic error due to false positives than the average normalized gain.

1. Introduction

The Force Concept Inventory (FCI) is a widely used assessment test to probe student conceptual understanding of Newtonian mechanics [1, 2]. The test has 30 items with five choices whose 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 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, 2]. In addition, Stewart et al. used a context-modified test to investigate context sensitivity of the FCI and they confirmed that the average test score is not particularly context-dependent [6]. Recently, DeVore et al. examined the effects of testwiseness in the FCI [7]. 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 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 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 [2], 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 [8 - 16]. 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 [10], 8 in 10 correct responses), although Q.16 should be solved with Newton’s third law. 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 [17]. The systematic error from false positives results in 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.

Recently, Yasuda et al. developed a method to estimate the systematic error due to false positives of Q.5, Q.6, Q.7, and Q.16 of the FCI [18]. They found that the sum of the 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. For example, if a student scores 20 out of 30 points on the FCI, two of those 20 points on average are coming from false positives from those four questions alone. It follows that the total measurement error from false positives on all 30 FCI questions must be even larger than this 10%.

When one considers that the FCI is used primarily for assessments of physics courses by comparing the scores of the pre and post test, one might wonder whether this systematic error makes any difference [3]. Namely, if the error is 10% on both the pre-test and the post-test, is it possible that the effects cancel each other out? Certainly, it depends upon how the pre and post-test scores are used in the assessment. In this paper, we build upon the body of research studying the validity of the FCI and present here a result of analyzing how systematic error affects the statistics for the effectiveness of a course. In particular, we focus on the average normalized gain [3] and Cohen’s d [19] in our analysis.

2. Methodology

2.1. Survey design
We classify a student’s response to a particular FCI question as being one of the four options shown in table 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. Hestenes et al. considered false positives and false negatives on the FCI and wrote that false positives “were fairly common” [1] but that “the probability of a false negative [is] certainly less than ten percent” [2]. 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 this paper.

| FCI Question | True attribute of a respondent | Understanding | Not understanding |
|--------------|--------------------------------|---------------|------------------|
| Correct      | True positive                  | False positive| True negative    |
| Incorrect    | False negative                 |               | True negative    |
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 \([12, 13, 18]\). 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.

As an example, we show Q.7 in figure 1, and the corresponding subquestions in figure 2. For 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.” In accordance with this finding, we created two subquestions for this FCI question, one asking about the force on the ball and the other asking about the velocity of the ball, as shown in figure 2.

**Figure 1.** Question 7 (Q.7) of the FCI: 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 in the figure, which path would the ball most closely follow after the string breaks?

**Figure 2.** Subquestions of the FCI Q.7: 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

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 2.

**Table 2.** Contingency table of answers with subquestions

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

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. 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 \([10]\). As these 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 as a result of clearly incorrect reasoning have not been found. In order to analyze the systematic error that arises from guessing on the remaining 27 questions, we chose Q.5 as a representative question because we could create subquestions for it in a straightforward manner.

In total, the survey instrument used in this study consisted of 40 questions, the 30 original Japanese-language FCI questions, 4 subquestions for Q.5, 2 subquestions for Q.6, 2 subquestions for Q.7, and 2 subquestions for Q.16 (see [18] for all subquestions). 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.

2.2. 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, resulting in 1110 valid responses. Most of the respondents were first-year students. Students were from different departments, mainly the departments of science, technology, and agriculture.

2.3. Analysis method
We calculate the systematic error of the FCI by subtracting the “true score” from the “raw score.” The raw score is the number of correct answers. The true score in this study is taken to be the number of raw score, true score, and systematic error of the FCI questions, 4 subquestions for Q.5, 2 subquestions for Q.6, 2 subquestions for Q.7, and 2 subquestions for Q.16 (see [18] for all subquestions). 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.

We define the random variable of raw score, true score, and systematic error of the ith question on the FCI as $S_{raw}^i$, $S_{true}^i$, and $E_{fp}^i$, respectively. With these notations, equation (3) for question i becomes

$$\langle S_{raw}^i \rangle = \langle S_{true}^i \rangle + \langle E_{fp}^i \rangle$$

We define the probability that each random variable takes 1 as $P_{raw}^i$, $P_{true}^i$, and $P_{fp}^i$. For example, $P_{raw}^i$ is the probability that a student answers the ith question of the FCI correctly. Since each random variable $S_{raw}^i$, $S_{true}^i$, and $E_{fp}^i$ follows a Bernoulli distribution, it follows that $\langle S_{raw}^i \rangle = P_{raw}^i$, $\langle S_{true}^i \rangle = P_{true}^i$ and $\langle E_{fp}^i \rangle = P_{fp}^i$. Therefore, we rewrite equation (4) as

$$P_{raw}^i = P_{true}^i + P_{fp}^i$$

The systematic error due to the false positives corresponds to the size of $P_{fp}^i$. After calculating $P_{raw}^i$ and $P_{true}^i$, $P_{fp}^i$ is calculated using equation (5).

3. Result

3.1. Systematic error of Q.5, Q.6, Q.7, and Q.16
First, we analyze the behavior of the systematic error due to the false positives on Q.5, Q.6, Q.7, and Q.16. In this analysis, $P_{raw}^i(S_{raw})$ and $P_{true}^i(S_{raw})$ for each of the 31 possible values of $S_{raw}$ (0, 1, 2, ..., 30) are estimated using the following equations,
\[ p_{\text{raw}}^i(S_{\text{raw}}) = \frac{N_{\text{raw}}^i(S_{\text{raw}})}{N(S_{\text{raw}})} \] 
\[ p_{\text{true}}^i(S_{\text{raw}}) = \frac{N_{\text{true}}^i(S_{\text{raw}})}{N(S_{\text{raw}})} \]

where \( N(S_{\text{raw}}) \) is the number of students with a given \( S_{\text{raw}} \), \( N_{\text{raw}}^i(S_{\text{raw}}) \) is the number of those students who answered the \( i \)th question correctly, and \( N_{\text{true}}^i(S_{\text{raw}}) \) is the number of those students who also answered the subquestions of that \( i \)th question correctly.

In figure 3, \( p_{\text{raw}}^i \) and \( p_{\text{true}}^i \) of Q.5, Q.6, Q.7, and Q.16 are plotted as a function of \( S_{\text{raw}} \). The error bars indicate Clopper-Pearson 95% confidence intervals [21] and are gray with caps for \( p_{\text{raw}}^i \) and black without caps for \( p_{\text{true}}^i \). From the graphs, we can see the tendency that as \( S_{\text{raw}} \) increases, \( p_{\text{raw}}^i \) and \( p_{\text{true}}^i \) also increase as we expected.

The systematic error of each question corresponds to the difference between \( p_{\text{raw}}^i \) and \( p_{\text{true}}^i \). In figure 4, \( p_{\text{fp}}^i \) of Q.5, Q.6, Q.7, and Q.16 are plotted as a function of \( S_{\text{raw}} \). Note that the systematic 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 [10] where we found respondents choosing the correct response with clearly erroneous reasoning for Q.6, Q.7, and Q.16, but we did not for Q.5.

![Figure 3](image_url)

**Figure 3.** \( p_{\text{raw}}^i \) (white circle) and \( p_{\text{true}}^i \) (black triangle) of Q.5, Q.6, Q.7, and Q.16 for each group of students with a given \( S_{\text{raw}} \) [18]. The error bars (\( p_{\text{raw}}^i \): gray with caps, \( p_{\text{true}}^i \): black without caps) indicate the Clopper-Pearson 95% confidence intervals of the dependent variable.
Figure 4. $P_{fp}^i$ of Q.5, 6, 7, and 16 for each group of students with a given $S_{raw}$ [18]. The error bars indicate the Clopper-Pearson 95% confidence intervals of the dependent variable.

Figure 5. (a) The sum of the systematic errors of Q.5, Q.6, Q.7, and Q.16 and (b) its ratio to a given raw score $S_{raw}$ for each group of students with that $S_{raw}$ [18]. The error bars indicate the square-and-add intervals of the dependent variable for each question. In figure 5(a), $\sum_{i=5,6,7,16} P_{fp}^i$ is fitted with a quadratic function.

3.2. Combined effect of systematic errors

From the above data, it is straightforward to estimate a minimum value for the systematic error due to false positives on the entire FCI. We do this by calculating the sum of the systematic errors $P_{fp}^i(S_{raw})$ for $i=5, 6, 7, \text{ and } 16$ and then examining its size relative to the raw score. As an equation, this ratio, $R_{fp}(S_{raw})$, is calculated by
\[ R_{fp}(S_{raw}) = \frac{\sum_{i=5,6,7,16} P_{fp}(S_{raw})}{S_{raw}} \]  

(8)

Figure 5 shows the dependency of a) \( \sum_{i=5,6,7,16} P_{fp}(S_{raw}) \) upon \( S_{raw} \) and b) \( R_{fp} \) upon \( S_{raw} \). The error bars indicate the square-and-add intervals of the dependent variable for each question [22, 23]. From the figure 5(b), 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 must be even larger than this.

3.3. Corrections to average normalized gain and Cohens’ d

We now consider what degree the error reported above influences the measures of the effectiveness of a course, namely, the average normalized gain and Cohens’ d. To that end, we create two sets of raw score data, each with \( N = 10,000 \) data points, shown in figure 6. The first data set, representing the pre-test score, is created to follow a normal distribution with mean of 12.2 and standard deviation of 5.9 (values calculated from actual data). Including the condition that no negative raw scores be allowed results in the data set being shifted to the right a small amount, to a mean of 12.4. The second data set, representing the post-test data, is created to follow a normal distribution with mean of 15.7 and standard deviation of 6.2 (values calculated from a different set of actual data). Then, the sum of the systematic errors of Q.5, Q.6, Q.7, and Q.16 for each raw score in each data set is calculated using the fitted function shown in figure 5(a), and the true score for each raw score is calculated by subtracting the sum of the systematic errors from the raw score based on equation (5). From figure 6, we can see that the distribution of the true scores is shifted to the left side of the distribution of the raw scores, as expected, with the mean of the true scores at 10.8 for the pre-test data and 13.9 for the post-test data.

![Figure 6](image_url) Simulated data (\( N = 10,000 \) each) for the pre and post test. The data is created assuming a normal distribution.

As we mentioned above, even if the systematic error has a large effect on the raw scores, in considering student growth from pre to post-test (for purposes of assessing the effectiveness of instruction), it is possible that the systematic errors may cancel [3]. In that case, the effect of the systematic error due to false positives may be actually negligible. To investigate this question, we utilize our simulated data sets to see what (if any) influence the decrease from raw to true score has on statistics that measure the effectiveness of a course. Here, we focus on the average normalized gain [3] and Cohen’s \( d \) [19]. The average normalized gain \( \langle g \rangle \) is defined by,
As this conclusion in this paper is based only upon simulated data, we plan to conduct more comprehensive analyses in subsequent research.

References
[1] Hestenes D, Wells M and Swackhamer G 1992 Phys. Teach. 30 141
[2] Hestenes D and Halloun I 1995 Phys. Teach. 33, 502
[3] Hake R 1998 Am. J. Phys. 66, 64
[4] Redish E F, Saul J M and Steinberg R N 1997 Am. J. Phys. 65, 45
[5] Redish E F 2003 Teaching Physics with the Physics Suite (Hoboken: Wiley) p 96
[6] Stewart J, Griffin H and Stewart G 2007 Phys. Rev. ST Phys. Educ. Res. 3, 010102
[7] DeVore S, Stewart J, Stewart G 2016 Phys. Rev. Phys. Educ. Res. 2, 20138
[8] Thornton R K, Kuhl D, Cummings K and Marx J 2009 Phys. Rev. ST Phys. Educ. Res. 5, 010105
[9] Wang J and Bao L 2010 *Am. J. Phys.* **78**, 1064
[10] Yasuda J, Uematsu H and Nitta H 2011 *J. Phys. Educ. Soc. Jpn* **59**, 90
[11] Scott T F, Schumayer D and Gray A R 2012 *Phys. Rev. ST. Phys. Educ. Res.* **8**, 20105
[12] Yasuda J and Taniguchi M 2013 *Phys. Rev. ST. Phys. Educ. Res.* **9**, 10113
[13] Taniguchi M and Yasuda J 2014 *J. Phys. Educ. Soc. Jpn* **62**, 226
[14] Wilson K F and Low D J 2015 *Am. J. Phys.* **83**, 802
[15] Low D J and Wilson K F 2016 *Am. J. Phys.* **85**, 54
[16] Scott T F and Schumayer D 2017 *Phys. Rev. Phys. Educ. Res.* **13**, 10126
[17] Taylor J R 1982 *An Introduction to Error Analysis* (Mill Valley: University Science Books) p 82
[18] Yasuda J, Mae N, Hull M and Taniguchi M 2018 *Phys. Rev. Phys. Educ. Res.* **14**, 010112
[19] Kirk R E 2007 *Statistics an Introduction* 5th ed. (Belmont: Thomson Wadsworth) pp 281-2
[20] Gregory R J 2015 *Psychological Testing* 7th ed. (Edinburgh: Pearson) pp 100-102
[21] Clopper C and Pearson E S 1934 *Biometrika* **26**, 404–13
[22] Fagan T 1999 *Comput. Biol. Med.* **29**, 83–7
[23] Zou G 2008 *Am. J. Epidemiol* **168**, 212–24