Correlating student interest and high school preparation with learning and performance in an introductory university physics course

Jason J. B. Harlow, David M. Harrison,[*] and Andrew Meyertholen

Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7 Canada

(Received 30 August 2013; published 7 April 2014)

We have studied the correlation of student performance in a large first year university physics course with their reasons for taking the course and whether or not the student took a senior-level high school physics course. Performance was measured both by the Force Concept Inventory and by the grade on the final examination. Students who took the course primarily for their own interest outperformed students who took the course primarily because it was required, both on the Force Concept Inventory and on the final examination; students who took a senior-level high school physics course outperformed students who did not, also both on the Force Concept Inventory and on the final exam. Students who took the course for their own interest and took high school physics outperformed students who took the course because it was required and did not take high school physics by a wide margin. However, the normalized gain on the Force Concept Inventory was the same within uncertainties for all groups and subgroups of students.

DOI: 10.1103/PhysRevSTPER.10.010112 PACS numbers: 01.40.Fk

I. INTRODUCTION

As part of a larger study, we have collected data on interest, background, and performance of students in our large (900 student) first year university physics course, PHY131. The course was given in the Fall term of 2012, and concentrates on classical mechanics. Almost 90% of the students in this course are or are intending to major in the life sciences. We surveyed the students about their reasons for taking our course, and whether or not the students took a senior-level high school physics course. We then correlated these factors with student performance as measured by the Force Concept Inventory (FCI) diagnostic instrument and by their scores on the final examination in the course. PHY131 is not intended for physics majors or specialists, or for engineering science students, who have their own courses.

The FCI has become a common tool for assessing students' conceptual understanding of mechanics, and for assessing the effectiveness of instruction in classical mechanics. The FCI was introduced by Hestenes, Wells, and Swackhammer in1992 [1], and was updated in 1995 [2]. The FCI has now been given to many thousands of students at a number of institutions worldwide. A common methodology is to administer the FCI at the beginning of a course, the "precourse," and again at the end, the "postcourse," and looking at the gain in performance. Our students were given one-half a point (0.5% of 100%) towards their final grade in the course for answering all 30 questions on the precourse FCI, regardless of what they answered, and another one-half point for answering all 30 questions on the postcourse FCI, also regardless of what they answered. Below, all FCI scores are in percent.

PHY131 is the first of a two-semester sequence, is calculus based, and the textbook used is by Knight [3] Two of us (J. J. B. H. and A. M.) were the lecturers. Research-based instruction is used throughout the course. Clickers, *Peer Instruction* [4], and interactive lecture demonstrations [5] are used extensively in the classes. There are 2 hours of class every week.

In addition, traditional tutorials and laboratories have been combined into a single active learning environment, which we call practicals [6]; these are inspired by physics education research tools such as McDermott's Tutorials in Introductory Physics [7] and Laws' Workshop Physics [8]. In the practicals, students work in teams of four on conceptually based activities using a guided discovery model of instruction. Whenever possible, the activities use a physical apparatus or a simulation. Some of the materials are based on activities from McDermott and Laws. There are 2 hours of practicals every week.

II. METHODS

The FCI was given during the practicals, the precourse one during the first week of classes and the postcourse one during the last week of classes. There is a small issue involving the values to be used in analyzing both the precourse and postcourse FCI numbers. In our course, 868 students took the precourse FCI, which was over 95% of the students who were currently enrolled, and 663 students took the postcourse FCI, which was over 95% of the students who were still enrolled at that time. Between the

[*] david.harrison@utoronto.ca

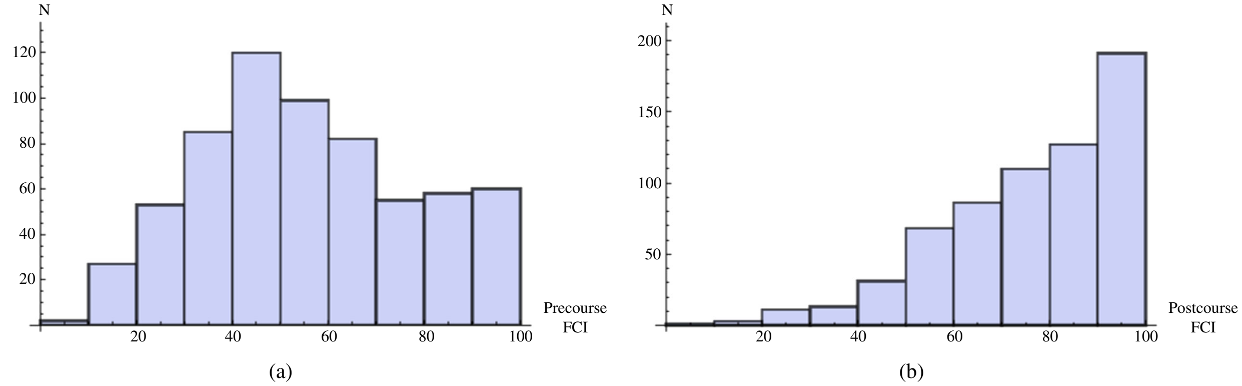

FIG. 1. (a) Precourse and (b) postcourse FCI scores for “matched” students.

precourse and postcourse FCI dates, 223 students had dropped the course; this dropout rate of about 25% is typical for this course. In addition, 22 students added the course later or for another reason did not take the precourse FCI, but did take the postcourse FCI. With one exception that is noted below, all data and analysis below use “matched” values, i.e., the 641 students who took both the precourse and the postcourse FCI. In all cases, the difference between using raw data or matched data is only a few percent. These small differences between matched and unmatched data are consistent with a speculation by Hake for courses with an enrollment >50 students [9].

Figure 1(a) shows the precourse FCI scores. The distribution is not well modeled by a Gaussian, due to the tendency for scores to flatline at higher values. Figure 1(b) shows the postcourse FCI scores, which do not conform to a Gaussian distribution at all. Therefore, in the analysis below we will use the medians and quartiles instead of computed means and standard deviations to characterize the FCI results. Appendix A lists the values of the quartiles for the data shown in Fig. 1, plus all other quartile values discussed below.

The final exam in the course was 2 hours long. It had 14 conventional problems (3 algebraic and 11 numeric), conceptual questions which included only words, figures, and/or graphs, and one question on uncertainty analysis. The exam had 12 multiple-choice questions worth 5 points each and 2 long-answer questions which were marked in detail with some part marks available. On the multiple choice section, 8 of the questions were traditional problems; in the long-answer section 12 of the available 20 points were traditional problems. Table I shows the overall relative weighting of these questions. We should emphasize that the “conceptual” questions were more tightly focused than the typical question on the FCI, and in no case were the questions on the exam based on FCI ones. Also, note that the majority of the exam was testing conventional problems [10].

TABLE I. Questions on the final exam.

Type of question	Weight
Conventional problems	72%
Conceptual	23%
Uncertainty analysis	5%

Six hundred sixty-eight students wrote the final exam. Figure 2 shows the grade distribution. It can be approximately modeled as a Gaussian, so we use the mean and standard deviation to characterize the distribution. Here the value for the mean is 68 and the standard deviation is 18. At the University of Toronto, a grade of 68 is a C+.

We asked the students six questions about their reason for taking the course and some background information about themselves. We collected these data during the second week of classes with clickers. Appendix B lists the questions and percentage of student answers. The only factors that gave statistically significant differences in student performance were their reason for taking the course, question 2, and whether or not they had taken a senior-level high school physics course, question 4. For some other questions, such as question 6 on whether the student has previously started but dropped the course, the lack of a correlation may be due to the fact that the percentage of students who had previously dropped the course was so small that the uncertainties in the results were overwhelming.

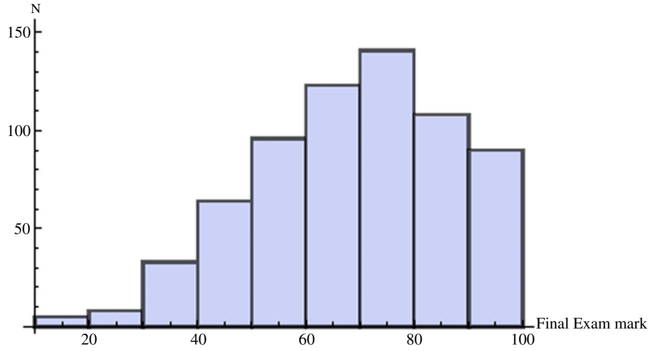

FIG. 2. Final exam scores for “matched” students.

Students receive a small number of points towards their grade for answering clicker questions in class. However, only about 75% of the matched students answered these questions. These comparatively low numbers surprised us. Perhaps some students had not yet gotten their clickers, or had not remembered to bring them to class, or did not bother to answer these questions. We note that this unfortunate loss of nearly 25% of our sample size could have been avoided if we had included these questions on the precourse FCI. Nonetheless, we believe that using the data for students who did answer these questions gives us a reasonable profile of the class.

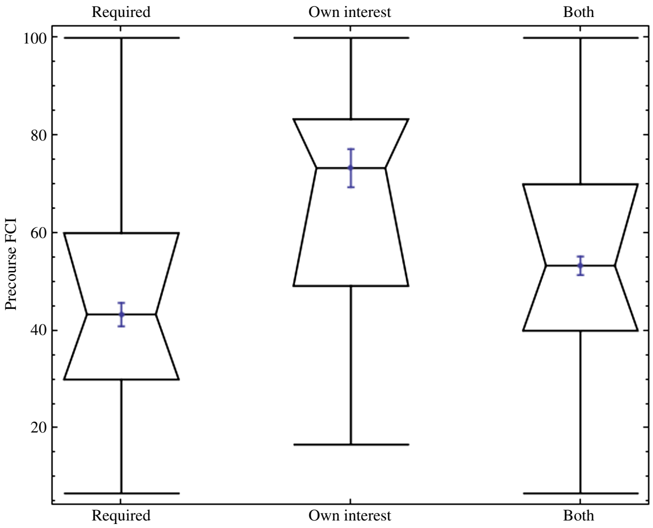

FIG. 3. Boxplots of the precourse FCI scores for different reasons for taking PHY131.

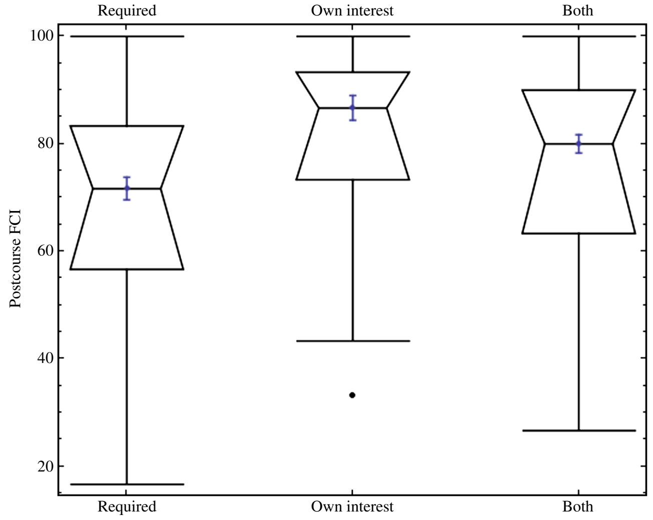

FIG. 4. Boxplots of the postcourse FCI scores for different reasons for taking PHY131.

TABLE II. Final examination performance for different reasons for taking the course.

Reason for taking PHY131	Final examination mean
Because it is required	65.5 ± 1.5 (C)
For their own interest	74.4 ± 2.0 (B)
Both because it is required and for their own interest	70.0 ± 1.1 (B−)

III. STUDENT REASONS FOR TAKING PHY131

As shown in Appendix B, the question that we asked the students about their reasons in taking our course and the percentage of the students in each category, in parentheses, was as follows:

What is the main reason you are taking PHY131?

A. It is required (32%)

B. For my own interest (16%)

C. Both because it is required and because of my own interest (52%)

Figure 3 shows boxplots of the precourse FCI scores for each category of student interest. The "waist" on the boxplot is the median, the "shoulder" is the upper quartile, and the "hip" is the lower quartile. The vertical lines extend to the largest (smallest) value less (greater) than a heuristically defined outlier cutoff [11]. Also shown in the figure are the statistical uncertainties in the value of the medians [12].

As seen in Fig. 4, the same correlation with student interest was seen in student performance on the postcourse FCI, although the overall median score was higher for the postcourse test (77%) than the precourse one (53%). The dot represents a data point that is considered to be an "outlier."

The different student reasons for taking PHY131 were also reflected in the final examination grades in the course, as shown in Table II. The errors are the standard error of the mean $\sigma_m \equiv \sigma/\sqrt{N}$, where σ is the standard deviation and N is the number of students. Also shown in parentheses are the corresponding letter grades of the means according to University of Toronto standards.

Appendix C discusses the p values for these distributions plus the two groups of the next section.

IV. SENIOR-LEVEL HIGH SCHOOL PHYSICS

In Ontario, the senior-level high school physics course is commonly called "grade 12 physics." Grade 12 physics or an equivalent course is recommended but not required for PHY131. As shown in Appendix B, 75% of our students took grade 12 physics and 25% did not.

There have been surprisingly few studies of high school physics and later performance in university physics. Champagne and Klopfer studied 110 University of Pittsburgh students and looked at many factors that might

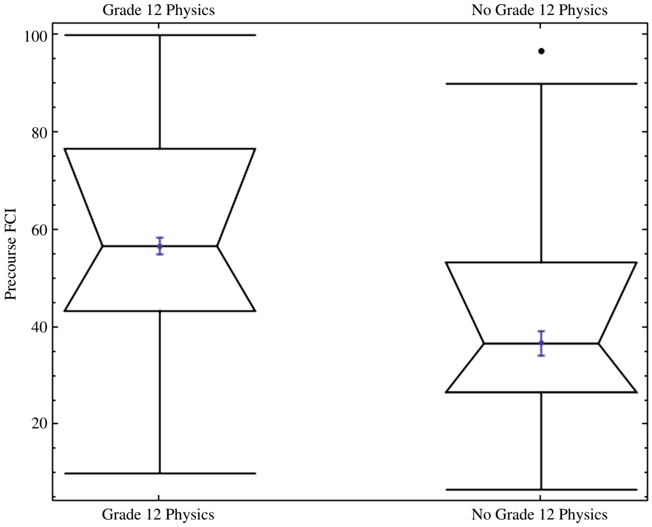

FIG. 5. Precourse FCI scores for students with and without senior level high school physics.

influence physics performance. They found that there was a positive correlation between taking high school physics and performance on university physics course tests and exams, although their methodology, perhaps wisely, did not attempt to quantify the size of the effect [13]. In 1993 Hart and Cottle reported that taking high school physics correlated with a mean 6.02 ± 1.09 increase in the final grade in university-level introductory physics for 508 students at Florida State University [14], and in 2001 Sadler and Tai reported a 3.49 ± 0.57 increase in a study of 1933 students at a variety of U.S. universities [15]. The differences between the values reported by Hart and Cottle versus Sadler and Tai are not well understood. However, Hazari, Tai, and Sadler in a massive study reported in 2007 showed that there are correlations between university physics course grades and the details of the curriculum of the high school physics course that the students took [16]. This result indicates that there is perhaps at least a small causal relationship between taking high school physics and university physics performance.

Figure 5 shows the precourse FCI scores for our students who did and did not take grade 12 physics. The boxplots for the postcourse FCI scores looked similar except for an overall upward shift in the median values, so are not shown.

Table III shows the final examination results for these two groups of students. Again, the students with grade 12 physics outperformed students without.

TABLE III. Final examination performance and whether the student took senior-level high school physics.

Took grade 12 physics?	Final examination mean
Yes	71.4 ± 0.9 (B−)
No	62.4 ± 1.6 (C−)

V. COMBINING INTEREST AND BACKGROUND

When we compare students who are primarily taking PHY131 for their own interest *and* who took grade 12 physics (61 students) with students who are primarily taking PHY131 because it is required *and* did not take grade 12 physics (48 students), the differences are quite dramatic, as shown in Figs. 6 and 7 and Table IV. Note in Fig. 6 that the interquartile ranges do not even overlap.

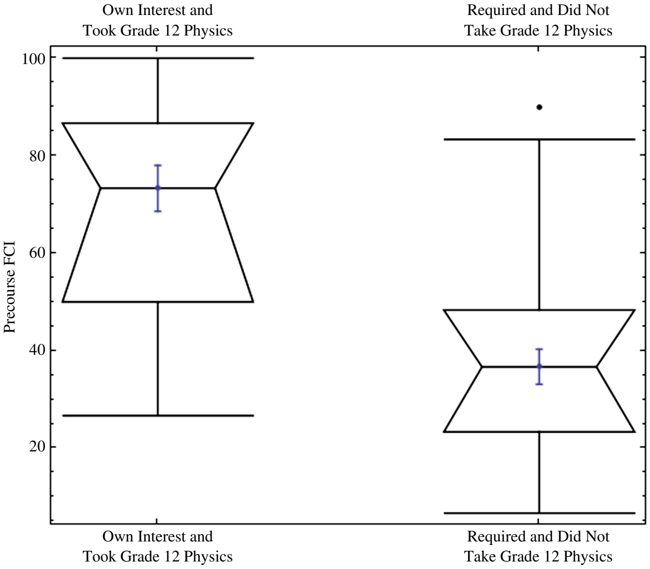

FIG. 6. Two different categories of students and their precourse FCI scores.

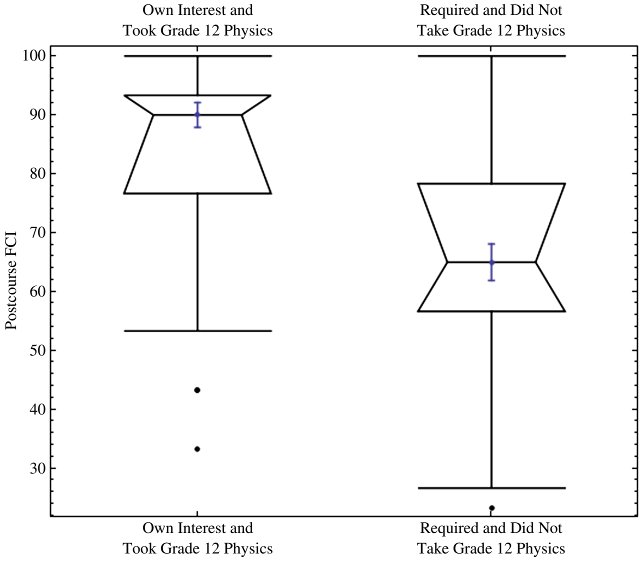

FIG. 7. Two different categories of students and their postcourse FCI scores.

TABLE IV. Two different categories of students and their final examination grades.

Category	Final examination mean
Taking PHY131 for their own interest and took grade 12 physics	75.9 ± 2.1 (B)
Taking PHY131 because it is required and did not take grade 12 physics	58.6 ± 2.6 (D+)

VI. GAINS ON THE FORCE CONCEPT INVENTORY

The standard way of measuring student gains on the FCI is from a seminal paper by Hake [17]. It is defined as the gain divided by the maximum possible gain, often called the normalized gain G:

$$G = \frac{(\text{postcourse\%} - \text{precourse\%})}{(100 - \text{precourse\%})}. \tag{1}$$

Clearly, G cannot be calculated for students whose precourse% score was 100. For our course, 9 students got perfect scores on the precourse FCI and no value of G was calculated. In addition to these 9 students, there were 10 students whose precourse% was over 80%, and whose G was less than -0.66. Somewhat arbitrarily, we classified these 10 students as outliers and ignore their G values below: perhaps they were survey fatigued and did not try to do their best on the postcourse FCI.

One hopes that the students' performance on the FCI is higher at the end of a course than at the beginning. The standard way of measuring the gain in FCI scores for a class is called the average normalized gain, to which we give the symbol $\langle g\rangle_{\text{mean}}$, and was also defined by Hake in Ref. [17]:

$$\langle g\rangle_{\text{mean}} = \frac{\langle \text{postcourse\%}\rangle - \langle \text{precourse\%}\rangle}{100 - \langle \text{precourse\%}\rangle}, \tag{2}$$

where the angle brackets indicate means. However, since the histograms of FCI scores such as Fig. 1 are not well approximated by Gaussian distributions, we believe that the median is a more appropriate way of characterizing the results. We will report $\langle g\rangle_{\text{mean}}$ since it is standard in the literature, but will also report the normalized gain using the medians $\langle g\rangle_{\text{median}}$, which is also defined by Eq. (2) except that the angle brackets on the right-hand side indicate the medians.

Recall that our study uses only "matched" FCI scores; the 10 student outliers are also excluded from our calculations of $\langle g\rangle$. The overall normalized gain for PHY131 was $(\langle g\rangle_{\text{mean}}, \langle g\rangle_{\text{median}}) = (0.45 \pm 0.02,\ 0.50 \pm 0.03)$. The stated uncertainties are the propagated standard error of the means for the average normalized gain and the interquartile ranges divided by $\sqrt{N}$ for the median normalized gain. The value of the average normalized gain is consistent with other courses that, like ours, make extensive use of research-based "reformed" pedagogy.

The normalized gains for all the categories and subcategories of students discussed above were consistent with being the same as the overall value for the course. Table V summarizes. Since the uncertainties in the values of the average normalized gains are the propagated standard errors of the mean of the average precourse and postcourse scores, 2 times the given values corresponds to a 95% confidence interval; interpreting the uncertainties in the median normalized gain is less direct. The conclusion that the normalized gains are the same is consistent with a p value of 0.39 for the values of G for the various groups of students, as discussed in Appendix C.

To the extent that that the normalized gain $\langle g\rangle$ measures the effectiveness of instruction, the data indicate that the pedagogy of PHY131 is equally effective for all groups and subgroups of students. As the saying goes, "A rising tide lifts all boats."

VI. DISCUSSION

Our goal was to determine if a student's interest in physics and/or involvement in a senior-level high school physics course had any effect on student success in a large Canadian university physics course. To our knowledge, this is the first time this has been attempted in such an institution. Although our results may be applicable to other institutions in other countries, we are not aware of any data

TABLE V. Average and median normalized gains for various student categories.

Student category	$(\langle g\rangle_{\text{mean}}, \langle g\rangle_{\text{median}})$
Taking the course because it is required	$(0.43 \pm 0.03, 0.50 \pm 0.04)$
Taking the course for their own interest	$(0.48 \pm 0.07, 0.50 \pm 0.11)$
Taking the course both because it is required and for their own interest	$(0.49 \pm 0.03, 0.57 \pm 0.04)$
Took grade 12 physics	$(0.45 \pm 0.03, 0.54 \pm 0.03)$
Did not take grade 12 physics	$(0.50 \pm 0.03, 0.55 \pm 0.04)$
Taking the course for their own interest *and* took grade 12 physics	$(0.44 \pm 0.08, 0.63 \pm 0.10)$
Taking the course because it is required *and* did not take grade 12 physics	$(0.46 \pm 0.05, 0.45 \pm 0.06)$

to support this except for the correlation with whether or not the student took high school physics (Refs. [13–16]).

We found evidence that taking physics for their own interest and having taken a senior-level high school physics course were both indicators for success on the final exam. Although the precourse and postcourse FCI scores were different for these groups and subgroups of students, neither interest nor background correlated within experimental uncertainties with the normalized gains on the FCI.

However, as shown in Table V, the highest performing group of students, those who took the course for their own interest and took grade 12 physics, also had the highest median normalized gain of 0.63 ± 0.10, while the lowest performing group, those who took the course because it was required and did not take grade 12 physics, had the lowest median normalized gain of 0.45 ± 0.06. The difference between these two values is 0.18 ± 0.12, which is perhaps suggestive of a nonzero value, but the difference from zero is not statistically significant.

There are, of course, other variables that correlate with physics performance for which we have not collected data; these include gender, socioeconomic background, etc. Hazari, Tai, and Sadler discuss many of these factors in Ref. [16]. However, there is one factor that we have not studied which has been shown to have a measurable impact on FCI performance: the ability of students to think in a scientific way. Lawson has developed a Classroom Test of Scientific Reasoning (CTSR) [18] that is based on Piagetian taxonomy [19]. Coletta and Phillips studied the correlation of CTSR performance with the average normalized gain G (not $\langle g \rangle$) and found a positive correlation for students at Loyola Marymount University, but in an indirect argument propose that there is no such correlation for students at Harvard [20]. Coletta, Phillips, and Steinert added data on a positive correlation for students at Edward Little High School [21], Diff and Tache found a positive correlation for students at Santa Fe Community College [22], and Nieminen, Savinainen, and Viiri found a positive correlation for high school students in Finland [23]. Since the groups and subgroups of students we studied have essentially the same median normalized gains $\langle g \rangle_{\text{median}}$, these CTSR G studies lead to some very interesting questions. One is, do the various groups and subgroups of students that we have studied have similar ability to reason in a scientific, formal operational way? Another related question is, are our students more like Harvard students than they are like students at, say, Loyola? Lacking data, we cannot answer either of these questions.

We should caution that when looking at the correlation between student performance and whether or not they took grade 12 physics, one should beware of assigning a cause-and-effect relationship to the data. For example, a student who knows (or perhaps just believes) that he or she is naturally weak in physics will tend to avoid taking grade 12 physics in order to keep a higher average grade. So is the student's ability to do well in physics determined by whether he or she took grade 12 physics, or perhaps vice versa?

Furthermore, the two questions about student interest and high school background are not independent. The students who avoid high school physics will also tend to be the students who are taking PHY131 mainly because it is required, and a higher percentage of students who voluntarily take high school physics will also tend to be taking PHY131 mainly for their own interest.

Considering the correlations of student background and interest with performance, measured with either the FCI or the course final examination, it is tempting to think of separating these widely divergent student populations. In 2002, Henderson looked at the idea of using FCI precourse results for this purpose, and his data show that this is not appropriate: the FCI score does not do a good job of predicting success or failure in the class [24].

The ultimate failures in our course are the 25% of the students who dropped it, although the "failure" may be ours, not the students. These are not "matched" students since they did not take the postcourse FCI. The quartiles of their performance on the precourse FCI were $(27, 40.0 \pm 2.0, 57)$, which are not radically lower than the matched students' quartiles of $(37, 53.3 \pm 1.3, 70)$. These dropouts had a similar profile of their reasons for taking the course, but 45% of them did not take a senior-level high school course compared to 25% of the matched students.

For the students who completed the course, 13 did not take a senior-level high school course, were taking our course mainly because it was required, and scored less than 25% on the precourse FCI. Over half of these students, 7 out of 13, ended up passing the final examination and 2 of them received letter grades of B; these two students received final course grades of B+ and A−; these two students also achieved normalized gains G on the FCI of 0.50 and 0.65, respectively. There was also one student in this group who got a C+ on the final exam, a final course grade of B−, and scored an amazing normalized gain G of 0.85 on the FCI, improving their FCI score from 13.3% to 86.7%. We certainly do not want to have excluded these good students from our course.

Our data are based on students self-reporting with clickers on their main reason for taking the course, and whether or not they took a senior-level high school physics course. All surveys have a problem with the fact that the people being surveyed have a tendency to answer what they believe the surveyor wishes to hear, and our clicker-based one probably has the same problem. We are unaware of any reason why a clicker-based survey may be more or be less biased than a paper-based one, a web-based one, or an in-person interview. Although in principle we could check the answers for whether or not the student took grade 12 physics, in fact the state of the databases at our university makes this extremely difficult; checking the question about interest in the course is probably impossible in principle.

TABLE VI. Quartiles of FCI scores for various categories of students.

Category	Lower quartile	Median	Upper quartile
All students, precourse FCI	37	53.3 ± 1.3	70
All students, postcourse FCI	63	76.7 ± 1.1	90
Taking the course because it is required,precourse FCI	30	43.3 ± 2.4	60
Taking the course because it is required,postcourse FCI	57	71.7 ± 2.1	83
Taking the course for their own interest,precourse FCI	49	73.3 ± 3.9	83
Taking the course for their own interest,postcourse FCI	73	86.7 ± 2.3	93
Taking the course both because it is required and for their own interest, precourse FCI	40	53.3 ± 1.9	70
Taking the course both because it is required and for their own interest, postcourse FCI	63	80.0 ± 1.7	90
Took grade 12 physics, precourse FCI	43	56.7 ± 1.7	77
Took grade 12 physics, postcourse FCI	67	80.0 ± 1.2	90
Did not take grade 12 physics,precourse FCI	27	36.7 ± 2.5	53
Did not take grade 12 physics,postcourse FCI	60	71.7 ± 2.5	87
Taking the course for their own interest and took grade 12 physics, precourse FCI	50	73.3 ± 4.7	87
Taking the course for their own interest and took grade 12 physics, postcourse FCI	77	90.0 ± 2.1	93
Taking the course because it is required and did not take grade 12 physics, precourse FCI	23	36.7 ± 3.6	48
Taking the course because it is required and did not take grade 12 physics, postcourse FCI	57	65.0 ± 3.1	78

Nonetheless, even if a fraction of the students answered these questions based on what they thought we wanted to hear, it would be very unlikely to change our conclusions.

VII. FUTURE WORK

Coletta and Phillips [20] showed that there is correlation between precourse FCI scores and the normalized gain G for students at 3 of the 4 schools studied, Loyola Marymount University, Southeastern Louisiana University, and the University of Minnesota, but found no correlation for students at Harvard. They believe that there is a “hidden variable” affecting these correlations: the ability of students to reason scientifically Our data, which are not shown, also show a positive correlation: fitting G versus the precourse FCI scores gave a slope of 0.00212 ± 0.00054, although, as discussed, we have not measured the hidden variable with the CTSR.

Administering the FCI under controlled conditions takes a total of 1 hour of precious time from our practicals, which is about 5% of the total. We are also using the Colorado Learning Attitudes about Science Survey (CLASS) [25], but since we are reluctant to give up more class or practical time, we have made it an online survey. Administering the CTSR under controlled conditions would take even more class or practical time. In addition, we are concerned about inducing “survey fatigue” in our students by giving them too many diagnostic instruments. However, we are considering using the CTSR, perhaps in place of the FCI, and looking at the reasoning ability of the various groups and subgroups of students that we have discussed in this paper.

ACKNOWLEDGMENTS

We particularly thank our teaching assistants, whose great work in the practicals is a vital part of the students' learning. We thank April Seeley in the Department of Physics for assistance in the collecting of the data of this study. We also thank Vatche Deyirmenjian, Stephen W. Morris, and Pierre Savaria in the Department of Physics for assistance in the design of this project. We benefited from discussions of data analysis with Richard Hake, Emeritus of Indiana University. The reviewers of this paper have also made many excellent suggestions that we believe have strengthened it.

APPENDIX A

Table VI provides quartiles of FCI scores for various categories of students.

APPENDIX B

We asked the students to self-report on the reason they are taking the course and some background information about themselves. Here we summarize that data.

1. “What is your intended or current Program of Study (PoST)?”

Answer	Percent
Life Sciences	88%
Physical and Mathematical Sciences	9%
Other/Undecided	4%

2. “What is the main reason you are taking PHY131?”

Answer	Percent
“Because it is required”	32%
“For my own interest”	16%
“Both because it is required and for my own interest”	52%

3. “When did you graduate from high school?”

Answer	Percent
2012	78%
2011	9%
2010	5%
2009	3%
Other/NA	4%

4. “Did you take Grade 12 Physics or an equivalent course elsewhere?”

Answer	Percent
Yes	75%
No	25%

5. “MAT135 or an equivalent calculus course is a co-requisite for PHY131. When did you take the math course?”

Answer	Percent
“I am taking it now”	81%
“Last year”	10%
“Two or more years ago”	9%

6. “Have your previously started but did not finish PHY131?”

Answer	Percent
Yes	4%
No	96%

APPENDIX C

The student’s t-test is well known for testing whether or not two distributions are the same [26]. It typically returns the probability that the two distributions are statistically the same, the p value, which is sometimes referred to just as p. By convention, if the p value is <0.05, the two distributions are considered to be different.

However, the test assumes that the two distributions are both Gaussian, which is not the case for FCI scores. Two alternatives for non-Gaussian distributions are the Mann-Whitney U test [27] and the Kruskal-Wallis one-way analysis of variance [28]. Both of these are based on the median, not the mean. The Kruskal-Wallis analysis is an extension of the one Mann-Whitney and can deal with more than two samples, but assumes that the distributions have the same shape and differ only in the value of the medians. Both typically return p values, which are interpreted identically to the p value of student’s t-test.

We are not aware of better alternatives to these ways of calculating p values for our data, although none are perfect. In practice, for our data all three methods gave similar p values in comparing the various groups and subgroups of students, although our software, *Mathematica*, sometimes complained about the fact that the data do not really match the assumptions of the particular algorithm being used. Table VII summarizes some of the results. Note that for comparing three or more categories of students, we show the results for the only test that accepts such data, the Kruskal-Wallis text, although the assumption of distributions with the same shape is not really correct, except for the G values of the last row.

TABLE VII. p values for different categories of students.

Category	Test	p value
Different reasons for taking the course, precourse FCI	Kruskal-Wallis	6×10^{-9}
Different reasons for taking the course, postcourse FCI	Kruskal-Wallis	1×10^{-6}
Different reasons for taking the course, final exam	Kruskal-Wallis	0.0014
Grade 12 physics? Precourse FCI	Mann-Whitney	3×10^{-14}
Grade 12 physics? Postcourse FCI	Mann-Whitney	0.00039
Grade 12 physics? Final exam	Student’s T-Test	2×10^{-6}
Different reasons for taking the course, grade 12 physics? G values	Kruskal-Wallis	0.39

[1] D. Hestenes, M. Wells, and G. Swackhammer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[2] Available from http://modeling.asu.edu/R&E/Research.html.

[3] Randall D. Knight, *Physics for Scientists and Engineers: A Strategic Approach* (Pearson, Toronto, 2013), 3rd ed.

[4] E. Mazur, *Peer Instruction: A User's Manual* (Addison-Wesley, New York, 1996).

[5] D. R. Sokoloff and R. K. Thornton, Using interactive lecture demonstrations to create an active learning environment, Phys. Teach. **35**, 340 (1997).

[6] University of Toronto's practicals Web site, http://www.upscale.utoronto.ca/Practicals/.

[7] L. C. McDermott, P. S. Schaffer, and the Physics Education Group, *Tutorials in Introductory Physics* (Prentice-Hall, Englewood Cliffs, NJ, 2002).

[8] P. W. Laws, *Workshop Physics Activity Guide* (Wiley, New York, 2004).

[9] Richard Hake, Lessons from the physics education reform effort, *Ecol. Soc.* **5**, 28 (2002), http://www.ecologyandsociety.org/vol5/iss2/art28/.

[10] The final examination is at http://www.physics.utoronto.ca/~jharlow/teaching/phy131f12/final.pdf.

[11] There are various conventions for the cutoff definition. We use 1.5 times the interquartile range extending from the upper and lower quartiles, which was proposed in J. D. Emerson and J. Strenio, Boxplots and batch comparison, in *Understanding Robust and Exploratory Data Analysis,* edited by D. C. Hoaglin, F. Mosteller, and J. W. Tukey (Wiley-Interscience, Toronto, 1983), p. 58ff. This cutoff definition is the usual one.

[12] We use the interquartile range divided by the square root of the total number of students in the sample, which is similar in spirit but simpler than a recommendation in B. Iglewicz, Robust Scale Estimators and Confidence Intervals for Location, Ref. 10, pg. 424.

[13] A. B. Champagne and L. E. Klopfer, A causal model of students' achievement in a college physics course, J. Res. Sci. Teach. **19**, 299 (1982).

[14] G. E. Hart and P. D. Cottle, Academic backgrounds and achievement in college physics, Phys. Teach. **31**, 470 (1993).

[15] P. M. Sadler and R. H. Tai, Success in Introductory Physics: The role of high school oreparation, Sci. Educ. **85**, 111 (2001).

[16] Z. Hazari, R. H. Tai, and P. M. Sadler, Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors, Sci. Educ. **91**, 847 (2007).

[17] 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).

[18] A. E. Lawson, The development and validation of a classroom test of formal reasoning, *J. Res. Sci. Teach.* **15**, 11 (1978). Also available from https://modelinginstruction.org/wp-content/uploads/2013/06/LawsonTest_4-2006.pdf.

[19] B. Inhelder and J. Piaget, *The Growth of Logical Thinking From Childhood to Adolescence; An Essay On The Construction of Formal Operational Structures* (Basic Books, New York, 1958).

[20] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).

[21] V. P. Coletta, J. A. Phillips, and J. J. Steinert, Why you should measure your students' reasoning ability, Phys. Teach. **45**, 235 (2007).

[22] K. Diff and N. Tache, From FCI To CSEM to Lawson test: A report on data collected at a community college, http://www.compadre.org/portal/items/detail.cfm?ID=9054&Relations=1.

[23] P. Nieminen, A. Savinainen, and J. Viiri, Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning, Phys. Rev. ST Phys. Educ. Res. **8**, 101223 (2012).

[24] C. Henderson, Common concerns about the Force Concept Inventory, Phys. Teach. **40**, 542 (2002).

[25] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[26] See, for example, E. M. Pugh and G. H. Winslow, *The Analysis of Physical Measurements* (Addison-Wesley, Don Mills, Canada, 1966), pp. 172ff; http://en.wikipedia.org/wiki/Student%27s_t-test (retrieved May 2, 2013).

[27] H. B. Mann and D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. **18**, 50 (1947).

[28] W. H. Kruskal and W. A. Wallis, Use of ranks in one-criterion variance analysis, J. Am. Stat. Assoc. **47**, 583 (1952).