Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability

Vincent P. Coletta[a)] and Jeffrey A. Phillips
Department of Physics, Loyola Marymount University, Los Angeles, California 90045

(Received 2 April 2004; accepted 9 September 2005)

We examined normalized gains and preinstruction scores on the force concept inventory (FCI) for students in interactive engagement courses in introductory mechanics at four universities and found a significant, positive correlation for three of them. We also examined class average FCI scores of 2948 students in 38 interactive engagement classes, 31 of which were from the same four universities and 7 of which came from 3 other schools. We found a significant, positive correlation between class average normalized FCI gains and class average preinstruction scores. To probe this correlation, we administered Lawson's classroom test of scientific reasoning to 65 students and found a significant, positive correlation between these students' normalized FCI gains and their Lawson test scores. This correlation is even stronger than the correlation between FCI gains and preinstruction FCI scores. Our study demonstrates that differences in student populations are important when comparing normalized gains in different interactive engagement classes. We suggest using the Lawson test along with the FCI to measure the effectiveness of alternative interactive engagement strategies. © *2005 American Association of Physics Teachers.*
[DOI: 10.1119/1.2117109]

I. INTRODUCTION

The force concept inventory[1] (FCI) is widely used as a measure of student understanding of introductory mechanics. It is usually given at the beginning and at the end of a course.[2] Students tend to score higher on the test when it is taken the second time, following instruction. Normalized gain G is defined as the change in score divided by the maximum possible increase

$$G = \frac{\text{postcore}\,\% - \text{prescore}\,\%}{100 - \text{prescore}\,\%} \tag{1}$$

In 1998, Hake[3] published the results of an extensive survey of class average gains for 6542 students in 62 introductory physics courses in high schools, colleges, and universities. Hake showed that the class average data for all courses [traditional and interactive engagement (IE)] combined, showed no significant correlation between normalized gain and preinstruction scores.

The importance of Hake's work cannot be overemphasized. Normalized gain provides a readily accessible, objective measure of learning in introductory mechanics. Research[3,4] demonstrates the superiority of IE methods to traditional methods of instruction. However, we will show that the uncritical use of G as a sole measure of relative effectiveness of alternative IE methods across diverse student populations may not be justified. For example, the lack of correlation between G and preinstruction scores for Hake's entire data set does not mean that there is no correlation between these quantities. It is possible that such correlations exist for subsets of the population considered by Hake or in other populations not included in his data set.

One of our purposes is to show that it is important to consider differences in student populations when comparing the normalized gains of different classes. For example, it might be incorrect to conclude that teaching methods used in an IE class with a normalized gain of 0.6 are necessarily more effective than those that produce a gain of 0.3 in a different class. The backgrounds of the students in the two classes could be a more important factor than the specific IE methods used in the classes.

An independent way to probe the background of a student population is the Lawson Classroom Test of Scientific Reasoning. This multiple-choice test includes questions on conservation, proportional thinking, identification of variables, probabilistic thinking, and hypothetico-deductive reasoning.[5,6] The test can be used to identify a student's reasoning level. Maloney showed that Lawson test score averages vary among different populations of college students.[7] In his study in the calculus and algebra based physics courses for science majors at Creighton University, nearly 2/3 of the students were rated as having reached the highest reasoning level, while in the courses that served education and health science majors, barely 1/3 of the students had reached this stage. Given these results, we looked for a possible correlation between Lawson test scores and normalized FCI gains.

In Sec. II we analyze both individual student and class average FCI data. In Sec. III we discuss correlations between G and scores on the Lawson test (see Appendix). Our conclusions are presented in Sec. IV.

II. FCI NORMALIZED GAIN AND PREINSTRUCTION SCORES

We analyzed individual normalized gains for students at Loyola Marymount University (LMU), Southeastern Louisiana University (SLU), University of Minnesota (UM), and Harvard University (HU). These schools employed IE methods in courses with a significant lecture component. The size of the student sample and the class size varied widely: 285 students in 11 classes at LMU, 86 students in two classes at SLU, 1648 students in 14 classes at UM, and 670 students in 4 classes at HU.

The Harvard classes were taught by Eric Mazur, Michael Aziz, William Paul, and Bob Westervelt, using the method described in Mazur's book, *Peer Instruction*.[2] Peer instruc-

tion classes consist of lectures that are divided into short segments each of which is followed by conceptual, multiple-choice questions. Students are first asked to answer the question individually and report their answers to the instructor through a computerized classroom response system or flashcards. When a significant portion of the class obtains the wrong answer, students are instructed to discuss their answers with their partners and, if the answers differ, to try to convince the partners of their answer. After this discussion, students report their revised answers, usually resulting in many more correct answers and much greater confidence in those answers. The class average values of G at HU are unusually high, typically about 0.6. Peer instruction was also used by Kandiah Manivannan at SLU.

The UM classes were taught by various instructors using the same general approach, called "cooperative group problem solving."[8,9] The majority of the class time is spent by the lecturer giving demonstrations and modeling problem solving before a large number of students. Some peer-guided practice, which involves students' active participation in concept development, is accomplished using small groups of students. In the recitation and laboratory sections, students work in cooperative groups, with the teaching assistant serving as coach. The students are assigned specific roles (manager, skeptic, and recorder) within the groups to maximize their participation. The courses utilize context-rich problems that are designed to promote expert-like reasoning, rather than the superficial understanding often sufficient to solve textbook exercises.

Of the 285 LMU students, 134 were taught by one of us (Coletta), using a method in which each chapter is covered first in a "concepts" class. These classes are taught in a Socratic style very similar to peer instruction. The material is then covered again in a "problems" class. The other author (Phillips) taught 70 students in lectures, interspersed with small group activities, using conceptual worksheets, short experiments, and context-rich problems. The other LMU professors, John Bulman and Jeff Sanny, both lecture with a strong conceptual component and frequent class dialogue.

The value of each student's normalized gain G was plotted versus the student's preinstruction score [see Figs. 1(a), 2(a), 3(a), and 4(a)]. Three of the four universities showed a significant, positive correlation between preinstruction FCI scores and normalized gains. Harvard was the exception.

There are, of course, many factors affecting an individual's value of G, and so there is a broad range of values of G for any particular prescore. The effect of prescore on G can be seen more clearly by binning the data, averaging values of G over students with nearly the same preinstruction scores. Binning makes it apparent that very low prescores ($<15\%$) and very high prescores ($>80\%$) produce values of G that do not fit the line that describes the data for prescores between 15% and 80%. For each university, we created graphs based on individual data, using all pretest scores; individual data, using only prescores from 15% to 80%; binned data, using all pretest scores; and binned data, using only prescores from 15% to 80%. Table I gives the correlation coefficients, significance level, and the slopes of the best linear fits to these graphs. Figures 1(b), 2(b), 3(b), and 4(b) show the binned data with prescores from 15% to 80%; the correlation effects are seen most clearly in these graphs. The usual measure of statistical significance is $p \leqslant 0.05$. For LMU and UM, the correlations are highly significant ($p < 0.0001$): the probability that G and prescore are not correlated in these populations is <0.0001. In Sec. III we discuss a possible explanation for the lack of correlation in the Harvard data.

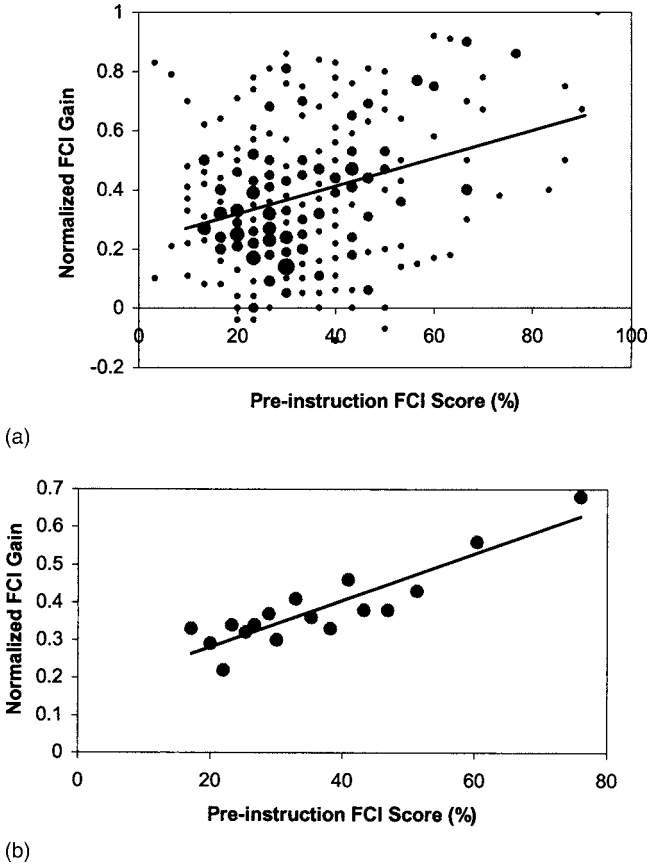
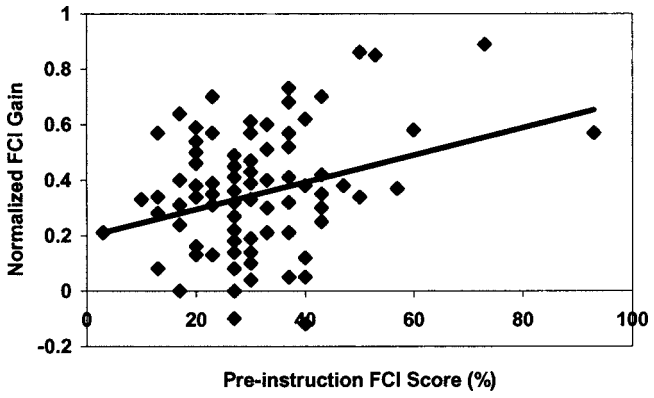


Fig. 1. (a) Plot of individual students' normalized FCI gain versus preinstruction FCI scores for 285 LMU students. Slope of best-fit line $s = 0.0047$, $r = 0.33$, $p < 0.0001$. (b) Plot of normalized FCI gains versus preinstruction FCI scores for LMU prescores between 15% and 80%, with individual student data averaged within 17 bins; $s = 0.0062$, $r = 0.90$, and $p < 0.0001$. The standard errors for G range from 0.03 to 0.06.

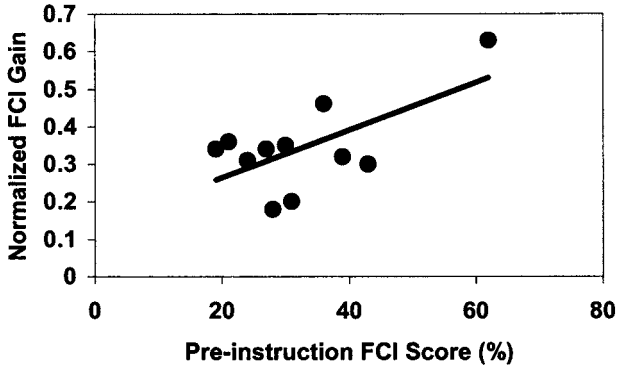
We chose bins with approximately the same number of students in each bin. Ideally, we want the bins to contain as many students as possible to produce a more meaningful average. However, we also want as many bins as possible. We chose the number of bins to be roughly equal to the square root of the total number of students in the sample, so that the number of bins and the number of students in each bin are roughly equal. Varying the bin size had little effect on the slope of the fit.

Class average preinstruction scores and normalized gains were also collected for seven classes at three other schools where peer instruction is used: Centenary College, the University of Kentucky, and Youngstown State University. These data and class average data for the 31 classes from the other four universities are shown in Fig. 5. The correlation coefficient is 0.63 and $p < 0.0001$. The linear fit gives a slope of 0.0049, close to the slopes in Figs. 1 and 2 and corresponds to $G = 0.34$ at a preinstruction score of 25% and a $G = 0.56$ at a preinstruction score of 70%.

We also examined data from Ref. 3. For Hake's entire data set, which includes high school and traditional college classes, as well as IE college and university classes, the correlation coefficient was only 0.02, indicating no correlation.

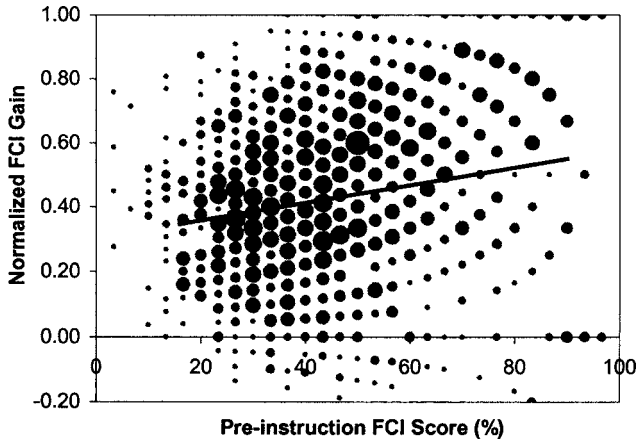


(a)

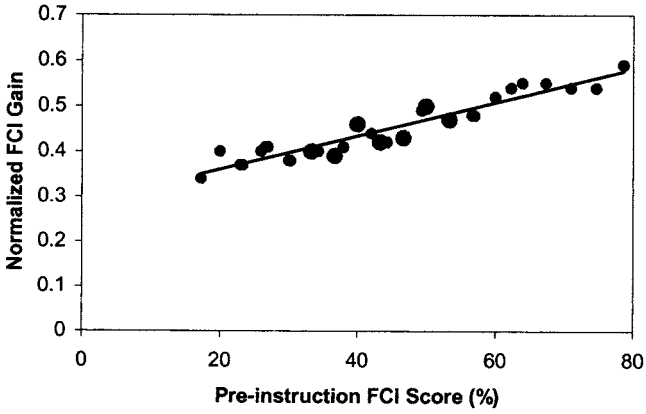


(b)

Fig. 2. (a) Plot of individual students' normalized FCI gain versus preinstruction FCI scores for 96 SLU students; $s=0.0049$, $r=0.30$, $p=0.003$. (b) Plot of normalized FCI gains versus preinstruction FCI scores for SLU prescores between 15% and 80%, with individual student data averaged within 11 bins; $s=0.0063$, $r=0.63$, and $p=0.04$. The standard errors for G range from 0.05 to 0.15.



(a)



(b)

Fig. 3. (a) Plot of individual students' normalized FCI gain versus preinstruction FCI scores for 1648 UM students; $s=0.0023$, $r=0.15$, and $p<0.0001$. (b) Plot of normalized FCI gains versus preinstruction FCI scores for UM prescores between 15% and 80%, with individual student data averaged within 38 bins; $s=0.0037$, $r=0.94$, and $p<0.0001$ Standard errors for G range from 0.03 to 0.05.

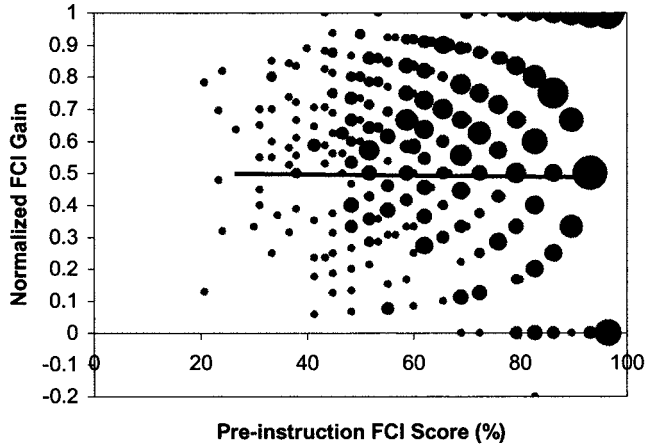
When we analyzed separately the 38 IE college and university classes in his study, we found some correlation, although not enough to show significance ($r=0.25$, $p=0.1$); the slope of the linear fit was 0.0032. We then combined Hake's college and university data with the data we collected. Hake's data provided only 35 additional classes, because 3 HU classes in Hake's data set were also contained in our data set. For the entire set of 73 IE colleges and universities, we found a slope of the linear fit close to what we had found for our collected data alone (0.0045). The coefficient was 0.39, significant for this size data set ($p=0.0006$).

III. FCI AND SCIENTIFIC REASONING ABILITY

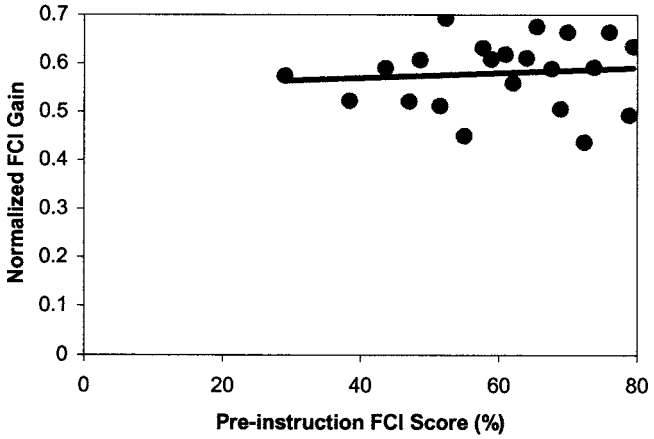
Recently, Meltzer published the results of a correlation study[10] for introductory electricity and magnetism. He used normalized gain on a standardized exam, the conceptual survey of electricity (CSE), to measure the improvement in conceptual understanding of electricity. He found that individual students' normalized gains on CSE were not correlated with their CSE preinstruction scores. Meltzer did find a significant correlation between normalized gain on the CSE and scores on math skills tests[11] for three out of four groups, with $r=0.30$, 0.38, and 0.46. Meltzer also reviewed other studies that have shown some correlation between math skills and success in physics. Meltzer concluded that the correlation shown by his data is probably not due to a causal relation between math skills and normalized gain and that students' level of performance on the math test and their normalized gains on the CSE may both be functions of one or more "hidden variables." He mentioned several candidates for such variables: general intelligence, reasoning ability, and study habits. We have come to similar conclusions regarding the correlation between G and the preinstruction score we found in our data set.

Piaget's model of cognitive development may provide some insight into differences among students in introductory physics. According to Piaget, a student progresses through discrete stages, eventually developing the skills to perform scientific reasoning.[12] When individuals reach the penultimate stage, known as concrete operational, they can classify objects and understand conservation, but are not yet able to form hypotheses or understand abstract concepts.[13] In the final stage, known as formal operational, an individual can think abstractly. Only at this point is an individual able to control and isolate variables or search for relations such as proportions.[14] Piaget believed that this stage is typically reached between the ages of 11 and 15.

Contrary to Piaget's theoretical notion that most teenagers reach the abstract thinking stage, educational researchers



(a)



(b)

Fig. 4. (a) Plot of individual students' normalized FCI gain versus preinstruction FCI scores for 670 HU students; $s=-0.0007$, $r=0.037$, $p=0.34$. (b) Plot of normalized FCI gains versus preinstruction FCI scores for HU prescores between 15% and 80%, with individual student data averaged within 22 bins; $s=0.0002$, $r=0.04$, and $p=0.87$. Standard errors for G range from 0.03 to 0.09.

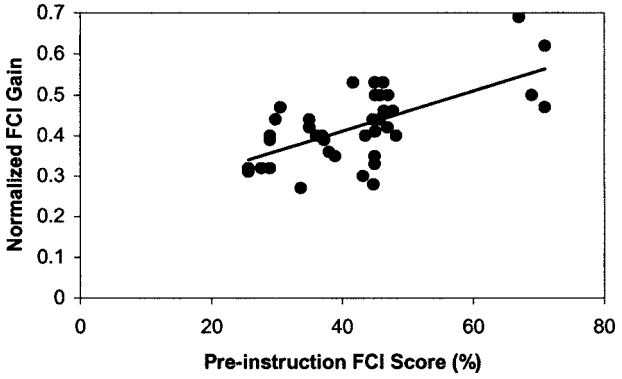


Fig. 5. Graph of class average normalized FCI gain versus average preinstruction FCI scores for all data collected in this study ($N=38$); $s=0.0049$, $r=0.63$, and $p<0.0001$.

have shown that many high school students, as well as college students, have not reached the formal operational stage.[15,16] Arons and Karplus claimed that only 1/3 of college students have reached the formal reasoning stage,[17] and that the majority of students either remain confined to concrete thinking or are only capable of partial formal reasoning, often described as transitional. In other studies focusing on physics students, including the work of Maloney, similar results have been seen.[7,18–20] Formal reasoning skills are necessary for the study of physics. For example, students who lack the ability to understand abstract concepts will struggle with Newton's second law.[21,22]

In 2003 we began to administer Lawson's Classroom Test of Scientific Reasoning as well as the FCI to LMU students to probe the relation between scientific reasoning ability and normalized gain on the FCI. Of the 285 LMU students tested with the FCI, 65 also took the Lawson test. We found a highly significant, positive correlation between students' normalized FCI gains and their Lawson test scores (see Fig. 6). With the slope of the linear fit of 0.0069, and $r=0.51$ ($p<0.0001$), this correlation is stronger than the correlation between FCI gains and preinstruction FCI scores either for these students alone (slope=0.0034, $r=0.26$) or in any of the other samples. Figure 7 shows the average value of G for each quartile in Lawson test scores. The 16 students with the highest Lawson scores (the top quartile) had an average Lawson score of 93% and an average G of 0.59 ± 0.07 (standard error), while the 16 students with the lowest Lawson scores (the bottom quartile) averaged 48% on the Lawson test, with an average G of 0.26 ± 0.04.

To compare the correlation between Lawson test scores and G with the correlation between FCI prescores and G, we divided the 65 student sample into two groups, those with FCI prescores $\leqslant 33\%$ ($N=33$) and those with FCI prescores $>33\%$ ($N=32$). We then divided each of these groups into two parts based on their Lawson test scores. Thus we obtained four groups: (1) 16 students with low FCI scores (23% average) and Lawson test scores $<60\%$ (48% average); (2) 17 students with low FCI scores (21% average) and Lawson test scores $\geqslant 60\%$ (76% average); (3) 15 students with high FCI scores (45% average) and Lawson test scores $<80\%$

Table I. Correlation of the normalized FCI gains and preinstruction scores for individual students and groups of students. The slope of the best-fit straight line, correlation coefficient r, and significance level p are given for each university.

		All Prescores			Prescores 15% to 80%			All data binned			Binned data, 15% to 8%		
	N	Slope	r	p	Slope	r	p	Slope	r	p	Slope	r	p
LMU	285	0.0047	0.33	<0.0001	0.0058	0.35	<0.0001	0.0049	0.81	<0.0001	0.0062	0.90	<0.0001
SLU	96	0.0049	0.30	0.003	0.0060	0.29	0.006	0.0045	0.51	0.09	0.0063	0.63	0.04
UM	1648	0.0023	0.15	<0.0001	0.0039	0.23	<0.0001	0.0026	0.78	<0.0001	0.0037	0.94	<0.0001
HU	670	−0.0007	0.037	0.34	0.0001	0.008	0.84	−0.0006	0.10	0.67	0.0002	0.04	0.87

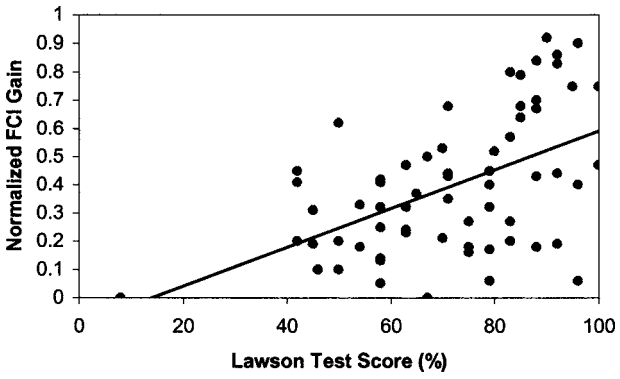


Fig. 6. Graph of individual students' normalized FCI gain versus Lawson test scores for 65 LMU students; $s=0.0069$, $r=0.51$, and $p<0.0001$.

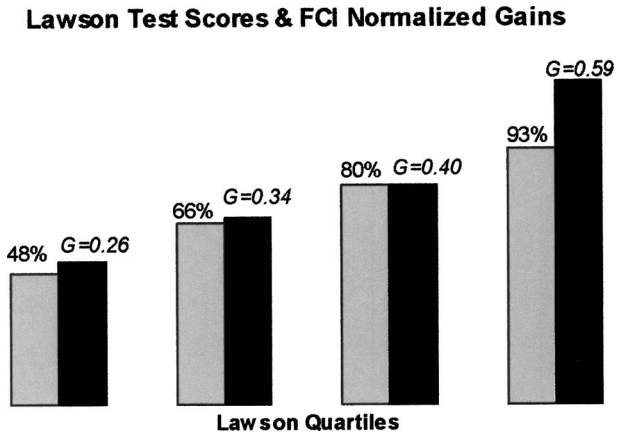


Fig. 7. Comparison of average G and average Lawson Test scores for quartiles of Lawson scores.

(69% average); and (4) 17 students with high FCI scores (58% average) and Lawson test scores $\geqslant$80% (91% average). The results in Table II indicate a stronger relation between G and Lawson test scores than between G and FCI prescores. For example, we see that, even though group 3 has a much greater average FCI prescore than group 2, group 3 has a lower average G (0.30 versus 0.44), consistent with the lower average Lawson test score (69% versus 76%).

As a final test that is relevant to our discussion of the Harvard data in Sec. IV, we examined data from the 16 students who scored highest on the Lawson test, the top quartile. In comparing FCI prescores and Lawson test scores for these students, we found no correlation ($r=0.005$). There was also no significant correlation between G and FCI prescores ($r=0.1$).

Our study indicates that Lawson test scores are highly correlated with FCI gains for LMU students. This correlation may indicate that variations in average reasoning ability in different student populations are a cause of some of the variations in the class average normalized gains that we observe. In other words, we believe that scientific reasoning ability is a "hidden variable" affecting gains, as conjectured by Meltzer.[10]

IV. CONCLUSIONS

Based on the data analyzed in this study, we conclude the following. (1) There is a strong, positive correlation between individual students' normalized FCI gains and their preinstruction FCI scores in three out of four of the populations tested. (2) There is a strong, positive correlation between class average normalized FCI gain and class average FCI preinstruction scores for the 38 lecture style interactive engagement classes for which we collected data, and nearly as strong a correlation between G and prescore when Hake's data from his 1998 study are included. (3) A sample of 65 students showed a very strong positive correlation between individual students' normalized FCI gain and their scores on Lawson's classroom test of scientific reasoning. The correlation between G and FCI prescores among these students is far less significant than the correlation between G and Lawson test scores.

Why does the Harvard data show no correlation between G and FCI prescores, while the other three schools show significant correlations? And why are there variations in the slopes? For LMU and SLU the slopes are 0.0062 and 0.0063, respectively, whereas the UM slope is 0.0037, and the class average slope is 0.0049. A possible answer to both questions is that these differences are caused by variations in the compositions of these populations with regard to scientific reasoning ability. We expect that a much higher fraction of Harvard students are formal operational thinkers and would score very high on Lawson's test. We found that among the top LMU Lawson test quartile, there is no correlation between FCI prescores and Lawson test scores, and no correlation between G and FCI prescores. It is reasonable to assume that a great majority of the Harvard student population tested would also show very high scientific reasoning ability and no correlation between scientific reasoning ability and FCI prescore; 75% of all Harvard students score 700 or higher on the math SAT, and math SAT scores have been shown to correlate with formal operational reasoning.[23,24] In contrast, less than 10% of LMU's science and engineering students have math SAT scores $\geqslant$700. If scientific reasoning ability is a hidden variable that influences FCI gains, we would expect to see no correlation between G and FCI prescore for very high reasoning ability populations.

Why should there be any correlation between G and FCI prescores for other populations in which a significant number of students are not formal operational thinkers? When students score low on FCI as a pretest in college, there are many possible reasons, including the inability to grasp what they were taught in high school due to limited scientific reasoning ability and lack of exposure to the material in high school. Those students whose lack of scientific reasoning ability limited their learning in high school are quite likely to have limited success in their college physics course as well. But those students who did not learn these concepts in high

Table II. Comparison of FCI prescores, Lawson test scores, and values of the FCI normalized gain G for four groups of students described in text. (s.e. is standard error).

Group	Average FCI prescore (%)	Average Lawson score (%)	Average $G\pm$s.e.
1	23	48	0.25±0.04
2	21	76	0.44±0.05
3	45	69	0.30±0.04
4	58	91	0.59±0.06

school for some other reason, but who do have strong scientific reasoning ability, are more likely to score high gains in an interactive engagement college or university physics class. We believe it is the presence of those students with limited scientific reasoning ability, present in varying proportions in different college populations, that is primarily responsible for the correlation between G and prescore that we have observed.

What, if anything, can be done about poor scientific reasoning ability? One indication that remedial help is possible is the work of Karplus. He devised instructional methods for improving proportional reasoning skills of high school students who had not learned these skills through traditional instruction. He demonstrated strong improvement, both short-term and long-term, with a great majority of those students.[25]

We hope to soon have all incoming students in the College of Science and Engineering at LMU take the Lawson test, so that we can identify students who are at risk for learning difficulties in physics and other sciences, and we have begun to develop instructional materials to help these students.

We hope that other physics instructors will begin to use the Lawson test in their classrooms. It would be especially meaningful if physics education researchers report interactive engagement methods that produce relatively high normalized FCI gains in populations that do not have very high Lawson test scores.

It is ironic that much of the improved average gains seen in interactive engagement classes is likely due to greatly improved individual gains for the best of our students, the most formal operational thinkers. This leaves much work to be done with students who have not reached this stage.

ACKNOWLEDGMENTS

We thank Anton Lawson for permission to reproduce in this paper his classroom test of scientific reasoning. We also wish to thank all those who generously shared their student FCI data, particularly those who provided individual student data: John Bulman and Jeff Sanny for their LMU data, David Meltzer and Kandiah Manivannan for SLU data, Catherine Crouch and Eric Mazur for Harvard data, and Charles Henderson and the University of Minnesota Physics Education Research Group for the Minnesota data, which was originally presented in a talk by Henderson and Patricia Heller.[26]

APPENDIX: Lawson's Classroom Test Of Scientific Reasoning: *Multiple Choice*

Version

Directions: This is a test of your ability to apply aspects of scientific and mathematical reasoning to analyze a situation to make a prediction or solve a problem. Make a dark mark on the answer sheet for the best answer for each item. If you do not fully understand what is being asked in an item, please ask the test administrator for clarification.

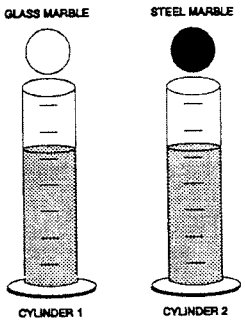
1. Suppose you are given two clay balls of equal size and shape. The two clay balls also weigh the same. One ball is flattened into a pancake-shaped piece. *Which of these statements is correct?*
   a. The pancake-shaped piece weighs more than the ball
   b. The two pieces still weigh the same
   c. The ball weighs more than the pancake-shaped piece

2. *because*
   a. the flattened piece covers a larger area.
   b. the ball pushes down more on one spot.
   c. when something is flattened it loses weight.
   d. clay has not been added or taken away.
   e. when something is flattened it gains weight.

3. To the right are drawings of two cylinders filled to the same level with water. The cylinders are identical in size and shape.

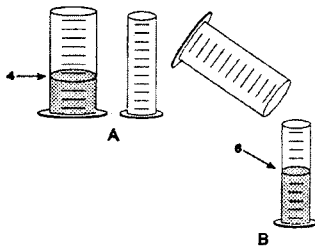
   Also shown at the right are two marbles, one glass and one steel. The marbles are the same size but the steel one is much heavier than the glass one.

   When the glass marble is put into Cylinder 1 it sinks to the bottom and the water level rises to the 6th mark. *If we put the steel marble into Cylinder 2, the water will rise*
   a. to the same level as it did in Cylinder 1
   b. to a higher level than it did in Cylinder 1
   c. to a lower level than it did in Cylinder 1



4. *because*
   a. the steel marble will sink faster.
   b. the marbles are made of different materials.
   c. the steel marble is heavier than the glass marble.
   d. the glass marble creates less pressure.
   e. the marbles are the same size.

5. To the right are drawings of a wide and a narrow cylinder. The cylinders have equally spaced marks on them. Water is poured into the wide cylinder up to the 4th mark (see A). This water rises to the 6th mark when poured into the narrow cylinder (see B).



Both cylinders are emptied (not shown) and water is poured into the wide cylinder up to the 6th mark. *How high would this water rise if it were poured into the empty narrow cylinder?*

a. to about 8
b. to about 9
c. to about 10
d. to about 12
e. none of these answers is correct

6. *because*

a. the answer can not be determined with the information given.
b. it went up 2 more before, so it will go up 2 more again.
c. it goes up 3 in the narrow for every 2 in the wide.
d. the second cylinder is narrower.
e. one must actually pour the water and observe to find out.

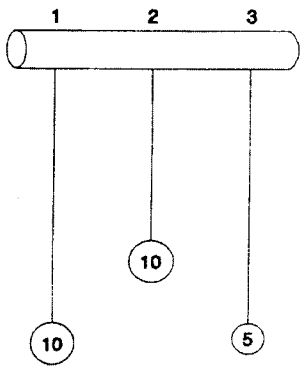
7. Water is now poured into the narrow cylinder (described in Item 5 above) up to the 11th mark. *How high would this water rise if it were poured into the empty wide cylinder?*

a. to about 7 1/2
b. to about 9
c. to about 8
d. to about 7 1/3
e. none of these answers is correct

8. *because*

a. the ratios must stay the same.
b. one must actually pour the water and observe to find out.
c. the answer can not be determined with the information given.
d. it was 2 less before so it will be 2 less again.
e. you subtract 2 from the wide for every 3 from the narrow.

9. At the right are drawings of three strings hanging from a bar. The three strings have metal weights attached to their ends. String 1 and String 3 are the same length. String 2 is shorter. A 10 unit weight is attached to the end of String 1. A 10 unit weight is also attached to the end of String 2. A 5 unit weight is attached to the end of String 3. The strings (and attached weights) can be swung back and forth and the time it takes to make a swing can be timed.

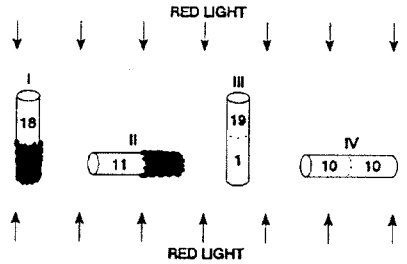


Suppose you want to find out whether the length of the string has an effect on the time it takes to swing back and forth. *Which strings would you use to find out?*

a. only one string
b. all three strings
c. 2 and 3
d. 1 and 3
e. 1 and 2

10. *because*
    a. you must use the longest strings.
    b. you must compare strings with both light and heavy weights.
    c. only the lengths differ.
    d. to make all possible comparisons.
    e. the weights differ.

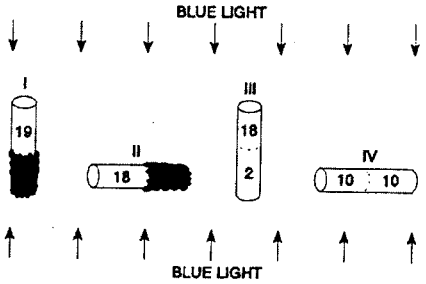
11. Twenty fruit flies are placed in each of four glass tubes. The tubes are sealed. Tubes I and II are partially covered with black paper; Tubes III and IV are not covered. The tubes are placed as shown. Then they are exposed to red light for five minutes. The number of flies in the uncovered part of each tube is shown in the drawing.



This experiment shows that flies respond to (respond means move to or away from)
    a. red light but not gravity
    b. gravity but not red light
    c. both red light and gravity
    d. neither red light nor gravity

12. *because*
    a. most flies are in the upper end of Tube III but spread about evenly in Tube II.
    b. most flies did not go to the bottom of Tubes I and III.
    c. the flies need light to see and must fly against gravity.
    d. the majority of flies are in the upper ends and in the lighted ends of the tubes.
    e. some flies are in both ends of each tube.

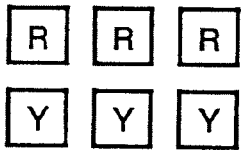
13. In a second experiment, a different kind of fly and blue light was used. The results are shown in the drawing.



These data show that these flies respond to (respond means move to or away from):
    a. blue light but not gravity
    b. gravity but not blue light
    c. both blue light and gravity
    d. neither blue light nor gravity

14. *because*
    a. some flies are in both ends of each tube.
    b. the flies need light to see and must fly against gravity.
    c. the flies are spread about evenly in Tube IV and in the upper end of Tube III.
    d. most flies are in the lighted end of Tube II but do not go down in Tubes I and III.
    e. most flies are in the upper end of Tube I and the lighted end of Tube II.

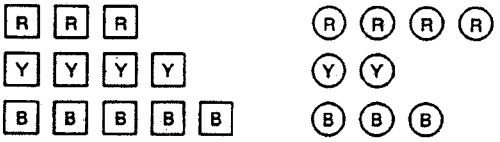
15. Six square pieces of wood are put into a cloth bag and mixed about. The six pieces are identical in size and shape, however, three pieces are red and three are yellow. Suppose someone reaches into the bag (without looking) and pulls out one piece. *What are the chances that the piece is red?*



a. 1 chance out of 6
b. 1 chance out of 3
c. 1 chance out of 2
d. 1 chance out of 1
e. can not be determined

16. *because*
a. 3 out of 6 pieces are red.
b. there is no way to tell which piece will be picked.
c. only 1 piece of the 6 in the bag is picked.
d. all 6 pieces are identical in size and shape.
e. only 1 red piece can be picked out of the 3 red pieces.

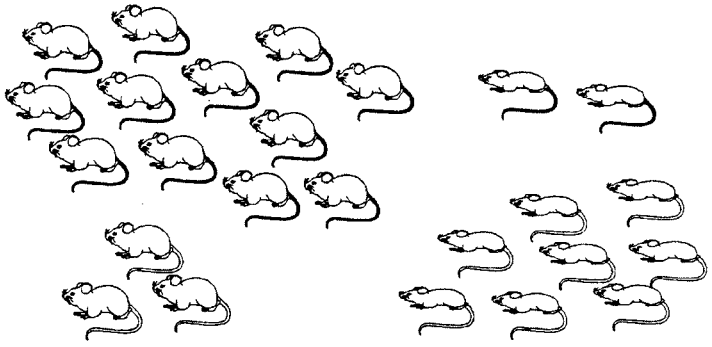
17. Three red square pieces of wood, four yellow square pieces, and five blue square pieces are put into a cloth bag. Four red round pieces, two yellow round pieces, and three blue round pieces are also put into the bag. All the pieces are then mixed about. Suppose someone reaches into the bag (without looking and without feeling for a particular shape piece) and pulls out one piece.



What are the chances that the piece is a red round or blue round piece?
a. can not be determined
b. 1 chance out of 3
c. 1 chance out of 21
d. 15 chances out of 21
e. 1 chance out of 2

18. because
a. 1 of the 2 shapes is round.
b. 15 of the 21 pieces are red or blue.
c. there is no way to tell which piece will be picked.
d. only 1 of the 21 pieces is picked out of the bag.
e. 1 of every 3 pieces is a red or blue round piece.

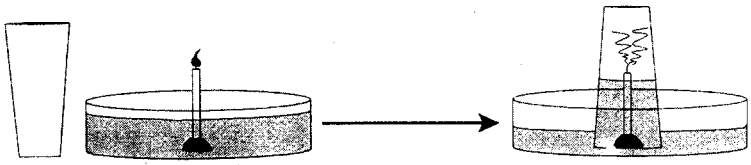
19. Farmer Brown was observing the mice that live in his field. He discovered that all of them were either fat or thin. Also, all of them had either black tails or white tails. This made him wonder if there might be a link between the size of the mice and the color of their tails. So he captured all of the mice in one part of his field and observed them. Below are the mice that he captured.



Do you think there is a link between the size of the mice and the color of their tails?
a. appears to be a link
b. appears not to be a link
c. can not make a reasonable guess

20. *because*
a. there are some of each kind of mouse.
b. there may be a genetic link between mouse size and tail color.
c. there were not enough mice captured.
d. most of the fat mice have black tails while most of the thin mice have white tails.
e. as the mice grew fatter, their tails became darker.

21. The figure below at the left shows a drinking glass and a burning birthday candle stuck in a small piece of clay standing in a pan of water. When the glass is turned upside down, put over the candle, and placed in the water, the candle quickly goes out and water rushes up into the glass (as shown at the right).



This observation raises an interesting question: Why does the water rush up into the glass?

Here is a possible explanation. The flame converts oxygen into carbon dioxide. Because oxygen does not dissolve rapidly into water but carbon dioxide does, the newly-formed carbon dioxide dissolves rapidly into the water, lowering the air pressure inside the glass.

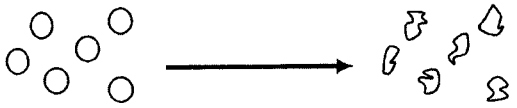
Suppose you have the materials mentioned above plus some matches and some dry ice (dry ice is frozen carbon dioxide). *Using some or all of the materials, how could you test this possible explanation?*

a. Saturate the water with carbon dioxide and redo the experiment noting the amount of water rise.
b. The water rises because oxygen is consumed, so redo the experiment in exactly the same way to show water rise due to oxygen loss.
c. Conduct a controlled experiment varying only the number of candles to see if that makes a difference.
d. Suction is responsible for the water rise, so put a balloon over the top of an open-ended cylinder and place the cylinder over the burning candle.
e. Redo the experiment, but make sure it is controlled by holding all independent variables constant; then measure the amount of water rise.

22. What result of your test (mentioned in #21 above) would show that your explanation is probably wrong?

a. The water rises the same as it did before.
b. The water rises less than it did before.
c. The balloon expands out.
d. The balloon is sucked in.

23. A student put a drop of blood on a microscope slide and then looked at the blood under a microscope. As you can see in the diagram below, the magnified red blood cells look like little round balls. After adding a few drops of salt water to the drop of blood, the student noticed that the cells appeared to become smaller.



Magnified Red Blood Cells After Adding Salt Water

This observation raises an interesting question: Why do the red blood cells appear smaller?

Here are two possible explanations: I. Salt ions (Na+ and Cl-) push on the cell membranes and make the cells appear smaller. II. Water molecules are attracted to the salt ions so the water molecules move out of the cells and leave the cells smaller.

To test these explanations, the student used some salt water, a very accurate weighing device, and some water-filled plastic bags, and assumed the plastic behaves just like red-blood-cell membranes. The experiment involved carefully weighing a water-filled bag in a salt solution for ten minutes and then reweighing the bag.

What result of the experiment would best show that explanation I is probably wrong?

a. the bag loses weight
b. the bag weighs the same
c. the bag appears smaller

24. *What result of the experiment would best show that explanation II is probably wrong?*

a. the bag loses weight
b. the bag weighs the same
c. the bag appears smaller

[a]Electronic mail: vcoletta@lmu.edu

[1] David Hestenes, Malcolm Wells, and Gregg Swackhamer, "Force concept inventory," Phys. Teach. **30**, 141–158 (1992).
[2] Eric Mazur, *Peer Instruction: A User's Manual* (Prentice-Hall, Upper Saddle River, NJ, 1997).
[3] Richard R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," Am. J. Phys. **66**, 64–74 (1998).
[4] Edward F. Redish and Richard N. Steinberg, "Teaching physics: Figuring out what works," Phys. Today **52**, 24–30 (1999).
[5] A. E. Lawson, "The development and validation of a classroom test of formal reasoning," J. Res. Sci. Teach. **15**(1), 11–24 (1978).
[6] A. E. Lawson, *Classroom Test of Scientific Reasoning*, revised ed. ⟨http://lsweb.la.asu.edu/alawson/LawsonAssessments.htm⟩
[7] D. P. Maloney, "Comparative reasoning abilities of college students," Am. J. Phys. **49**(8), 784–786 (1981).
[8] P. Heller, R. Keith, and S. Anderson, "Teaching problem solving through cooperative grouping. Part 1: Groups versus individual problem solving," Am. J. Phys. **60**, 627–636 (1992).
[9] P. Heller and M. Hollabaugh, "Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups," Am. J. Phys. **60**, 637–645 (1992).
[10] David E. Meltzer, "The relationship between mathematics preparation and conceptual learning in physics: A possible 'hidden variable' in diagnostic pretest scores," Am. J. Phys. **70**, 1259–1268 (2002).
[11] Meltzer used two different math skills tests: the ACT Mathematics for two of the four groups and, for the other two, Hudson's Mathematics Diagnostic Test. H. Thomas Hudson, *Mathematics Review Workbook for College Physics* (Little, Brown, 1986), pp. 147–160.
[12] J. W. Renner and A. E. Lawson, "Piagetian theory and instruction in physics," Phys. Teach. **11**(3), 165–169 (1973).
[13] 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).
[14] A. E. Lawson, "The generality of hypothetico-deductive reasoning: Making scientific thinking explicit," Am. Biol. Teach. **62**(7), 482–495 (2000).
[15] D. Elkind, "Quality conceptions in college students," J. Soc. Psychol. **57**, 459–465 (1962).
[16] J. A. Towler and G. Wheatley, "Conservation concepts in college students," J. Gen. Psychol. **118**, 265–270 (1971).
[17] A. B. Arons and R. Karplus, "Implications of accumulating data on levels of intellectual development," Am. J. Phys. **44**(4), 396 (1976).
[18] H. D. Cohen, Donald F. Hillman, and Russell M. Agne, "Cognitive level and college physics achievement," Am. J. Phys. **46**(10), 1026–1029 (1978).
[19] J. W. McKinnon, and J. W. Renner, "Are colleges concerned with intellectual development?," Am. J. Phys. **39**(9), 1047–1051 (1971).
[20] A. E. Lawson, and J. W. Renner, "A quantitative analysis of responses to piagetian tasks and its implications for curriculum," Sci. Educ. **58**(4), 545–559 (1974).
[21] J. W. Renner and A. E. Lawson, "Promoting intellectual development through science teaching," Phys. Teach. **11**(5), 273–276 (1973).
[22] J. W. Renner, "Significant physics content and intellectual development-cognitive development as a result of interacting with physics content," Am. J. Phys. **44**(3), 218–222 (1976).
[23] SAT data from The College Board web site, 2004, ⟨www.collegeboard.com⟩
[24] G. O. Kolodiy, "Cognitive development and science teaching," J. Res. Sci. Teach. **14**(1), 21–26 (1977).
[25] B. Kurtz and R. Karplus, "Intellectual development beyond elementary school vii: teaching for proportional reasoning," Sch. Sci. Math. **79**, 387–398 (1979).
[26] C. Henderson and P. Heller, "Common Concerns about the FCI," *Contributed Talk, American Association of Physics Teachers Winter Meeting*, Kissimmee, FL, January 19, 2000.