

The Gender Gap on the FCI – Question by Question

by

Richard D. Dietz, Matthew R. Semak, Courtney W. Willis
Physics, University of Northern Colorado, Greeley, CO 80639

Abstract: The existence of a gender gap in performance on the Force Concept Inventory has been established by several studies. Those studies have focused on the total score achieved by students on the FCI and subsequent gains in the total score when the instrument is administered again after instruction. Here we analyze the gender gap for each of the 30 questions in the FCI to determine if particular questions present greater difficulty for females.

Introduction: At the start of the 2007 fall semester two introductory physics classes (n=114) at UNC took the FCI and the Lawson. Both groups also took the FCI again at the end of the semester. Information about each student's gender was collected. Of the original 114 students tested, complete results were available for only 90 students, so their results make up the data set we analyzed. The distribution by gender is males = 42, females = 48.

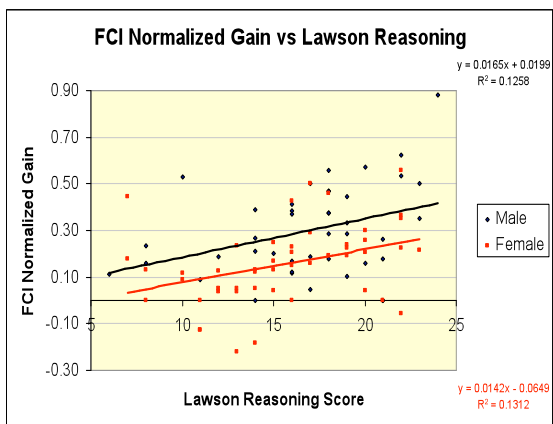
Overview: When we consider only the total scores on the three tests we find the following results averaged by gender. The number presented is the fraction of the possible questions answered correctly. A perfect score on a test would be represented by 1.0. We adopt the definition of Normalized Gain advocated by Bao (2006) defined as follows:

$$\text{If Post} \geq \text{Pre then } G = (\text{Post} - \text{Pre}) / (1 - \text{Pre})$$

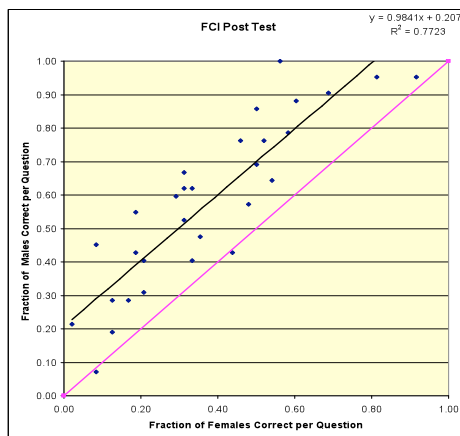
$$\text{If Post} < \text{Pre then } G = (\text{Post} - \text{Pre}) / (\text{Pre})$$

	FCI Pre	FCI Post	Gain	Lawson
Female	0.25	0.38	0.16	0.65
Male	0.41	0.58	0.26	0.71

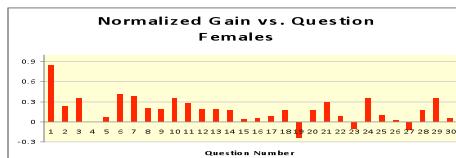
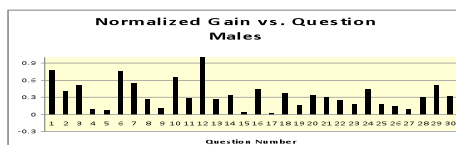
Reasoning: The graph below shows the normalized FCI gain vs. the Lawson Reasoning score. Each dot represents a student's score. The positive correlation between the FCI Gain and Lawson Score is consistent with the results reported by Colette and Phillips (2005).



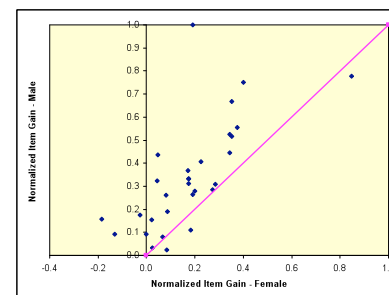
Performance: The graph below compares male/female performance for each question on the FCI when it was administered as a post-test. The red line is a plot of $x = y$, and each dot represents a question. Twenty-eight of the thirty dots lie above the red line. If we compare lying to one side or the other of the red line to the flipping of a fair coin, the likelihood of obtaining the result by chance is one in 2.5 million. The slope (~1.0) of the line of best fit suggests that the gender gap is independent of proficiency.



Gains: The bar graphs below show the male and female normalized gains for each of the 30 questions on the FCI.



The graph below shows how male/female normalized gains compare for each of the 30 questions on the FCI. We have analyzed the distribution of normalized gains using the non-parametric Mann-Whitney U test and find that we may reject ($p < 0.01$) the null hypothesis that gain is unrelated to gender.



Question Difficulty: Do males and females agree on which questions on the FCI are easiest or hardest? Below we rank the 30 questions from hardest to easiest based on the fraction of males or females who answered the question correctly on the FCI post-test.

Males		Females		Males		Females		Males		Females	
Question	Fraction	Question	Fraction	Question	Fraction	Question	Fraction	Question	Fraction	Question	Fraction
17	0.07	26	0.02	30	0.45	22	0.29	8	0.69	8	0.50
5	0.19	17	0.08	2	0.48	14	0.31	3	0.76	10	0.50
26	0.21	30	0.08	28	0.52	20	0.31	27	0.76	3	0.52
4	0.29	5	0.13	23	0.55	28	0.31	16	0.79	29	0.54
25	0.29	25	0.13	21	0.57	11	0.33	10	0.86	12	0.56
15	0.31	4	0.17	22	0.60	19	0.33	24	0.88	16	0.58
11	0.40	18	0.19	19	0.62	2	0.35	7	0.90	24	0.60
13	0.40	23	0.19	20	0.62	9	0.44	1	0.95	7	0.69
9	0.43	13	0.21	29	0.64	27	0.46	6	0.95	6	0.81
18	0.43	15	0.21	14	0.67	21	0.48	12	1.00	1	0.92

Two lists like these may be compared in the following way. If the number of items n in the list is an even number (such as 30) then the maximum number of shifts of the elements in the second list to bring it into agreement with the first is $n/2$ which in the case of 30 items is 450. Say that m shifts are required to bring the lists into agreement. We then define the degree of agreement, A , to be $1 - m/(n/2)$. By counting shifts we find that for the two lists above, $A = 0.81$ on a scale where 1.0 would mean perfect agreement. It appears that there is general agreement between males and females on the degree of difficulty of the individual questions on the FCI.