Evaluating students' performance on the FCI at a minority serving institution

Qing X. Ryan
Department of Physics and Astronomy, California State Polytechnic University Pomona, 3801 W. Temple Ave., Pomona, CA, 91768 USA

Darwin Agunos, Armando Villasenor, Homeyra Sadaghiani and Alexander Small
Department of Physics and Astronomy, California State Polytechnic University Pomona, 3801 W. Temple Ave., Pomona, CA, 91768 USA

As part of an effort to provide evidence for the reproducibility of educational studies for a variety of student bodies, we collected a year-long data set in introductory physics courses at Cal Poly Pomona (a primarily undergraduate and Hispanic-serving institution) to understand factors that affect students' performance on the FCI (force concept inventory). This study also allows us to gain insights into possible gender or racial gaps in students' performance. In this paper, we discuss background variables that predict students' FCI scores at the end of the term. Weak correlation is found between students' SAT score and FCI normalized gain, suggesting education research findings can be population dependent. Gender and racial gaps are found in students' FCI performance, both at initial preparation and overall gain. There is a gender gap of 16% FCI pre-test and 17% of FCI post-test with a strong effect size (d=0.81). Caucasian students outperform Asian and Hispanic students on FCI pre, post and gain. No significant interaction is found between gender and race/ethnicity after controlling for course grade.

2019 PERC Proceedings edited by Cao, Wolf, and Bennett; Peer-reviewed, doi.org/10.1119/perc.2019.pr.Ryan

I.INTRODUCTION

Research efforts in many disciplines have led to significant progress in understanding how students learn, identifying fundamental principles of learning, and using these principles to guide effective teaching practices [1]. In the field of Physics Education Research (PER) and other discipline-based research, one of the important goals is to measure students' learning outcomes and make comparisons between different pedagogical methods. However, most of the PER research has been done in R1 (research one) universities. Less research is done in primarily undergraduate institutions (PUI) or minority serving institutions (with 30-45% Hispanic/Latino population) [2]. Conducting such research for a variety of student populations will provide valuable evidence for reproducibility of educational studies [3].

Conducting replication studies for a variety of student population can also be particularly insightful in understanding teaching and learning of underrepresented populations. Among research-based conceptual assessments [4,5,6], the Force Concept Inventory (FCI) [7] is one of the most well-known assessments, and a consistent gender gap in performance on the FCI has been observed [8]. While the gender gap has been extensively studied, little research has been done to explore whether similar gaps extend to underrepresented racial groups. In a recent study by Henderson and Stewart [9], gender and racial gaps were investigated in the first-semester, calculus-based mechanics course at a midwestern land-grant university. Differences in FCI posttest performance were found, with Caucasian students outperforming African-American students (14%) and Hispanic students (6%). After controlling for course performance measured by physics grade, the differences narrowed but still exist. As pointed out by the authors, this work was performed at one institution and the results may be population dependent. The work also relied on self-reported race where mixed race might be miscounted. Studies at another institution will contribute to these findings.

Motivated by the goal of doing replication studies at a PUI and further exploring gender and racial gaps, we conducted a year-long study in introductory mechanics (first-term calculus-based introductory physics for science and engineering) at Cal Poly Pomona (CPP) to evaluate student performance on the FCI, as well as probing possible gender and racial gaps in students' performance. We collected data in this course for an entire year, covering all terms to give a complete sample of student population: Winter 2017, Spring 2017, Summer 2017 and Fall 2017. The data we collected include: FCI pre-test and post-test scores, students' math pre-requisite GPA (grade point average), SAT scores [10], etc.

Research questions of this study include the following: 1) Understand local population: how background variables (such as SAT, GPA, etc) are related to students' performance on the FCI at our institution? Answers to this question helps to serve as a baseline and reference for any future pedagogical reforms. 2) Evidence for reproducibility: one previous research study at Loyola Marymount University suggests that students' SAT scores correlate with their normalized gain on the FCI [11]. Can we replicate the strong correlation between FCI gain and SAT? 3) Gender and racial gaps: Are there any gender and racial gaps in students' performance in FCI at our institution?

II.METHODS

The calculus-based introductory physics course for science and engineering majors at CPP is taught in many different sections, with regular class sizes being 60 students and large classes having 110 students. Both tenure-track faculty and adjunct lecturers teach these classes but lecturers teach a majority of the sections. The classes meet for ten weeks total (one quarter) and each class is typically three 50-minute lectures or two 75-minute lectures. A three-hour laboratory is a co-requisite for the lecture but is taught separately.

For logistical reasons, we collected FCI data in the laboratory course since there is usually more free time in the first and last week of the lab. We explained to the students the survey will not affect their grade but will help us gather information about student learning and inform curriculum and instruction transformations. Students were asked to give their best efforts and offered the opportunity to talk about their performance in person with one of the faculty leading the research afterwards. Roughly 30 minutes were given for both pre and post tests. Other than the FCI, we also collected background information from IRAR (Institutional Research and Academic Resources), including SAT scores, instructor's name of the lecture section, students' GPA at enrollment of the class, high school GPA, students' GPA in the math pre-requisite class (Calculus) and students course grade at the end of the term.

To capture a complete picture of the entire student population, we collected data in all four quarters throughout the year: Winter 2017, Spring 2017, Summer 2017 and Fall 2017. No special pedagogical interventions were conducted in this academic year, all instructors involved were teaching the way they normally would teach. Some instructors give traditional lectures and some teach with more student interaction. We categorize the type of their instructions into a three-point Likert scale: IT0--traditional lecture, IT--somewhat interactive, IT2--more interactive. This ranking was assigned collaboratively by two tenure-track faculty members who are also authors of this paper. One faculty member spent 10+ years in the department and had conducted many classroom observations for other instructors and therefore had a good understanding of the instructors' approaches. There are two ways to compute normalized gain $=(post\% - pre\%)/(1 - pre\%)$: one is gain of the averages (<g>)—average pre and post scores of the entire class, one is average of the gains (G)—computing gain for each student and take the overall average if needed. We use both types of gains in this analysis and we use different symbols (<g> and G) to differentiate them. The difference between these two calculations is not significant for large classes and the gain of averages is the official definition given by Hake [6]. There are many concerns and discussions on the appropriate use of normalized gain [12,13], one ought to be careful when interpreting the results. Since we are interested in comparing results with what was previously reported, for this first round of analysis, we used the typical definition. Gain of averages(<g>) was used when we compare to the commonly reported gains in Hake's study [6]. When we need fine-grained data in order to compute correlations between gain and student variables (e.g. SAT scores), individual gain must be calculated and G is reported.

III. RESULTS

A. How background variables relate to FCI?

To understand how background variables (SAT, GPA, etc) are related to students' FCI scores, we first investigate the correlation between background variables and students' FCI scores. We also used hierarchical linear regression to investigate if FCI scores differ for different instructional approaches, with other variables controlled for.

Table I shows the Pearson correlation coefficients of students' FCI post-test scores with the following variables: FCI pre-test, SAT math, SAT total (verbal+math), grade in the math pre-requisite course, overall GPA at the time of enrollment in the physics course, and high school GPA. Since we don't have all pieces of data for all of the students, different numbers of sample sizes are reported in Table I.

We can see that FCI pre-score correlates very strongly with FCI post score, predicting almost 65% of the variance ($r = 0.80$). The next strong predictor is SAT total, which correlates with FCI post score with a correlation coefficient of 0.52.

TABLE I. Correlation of FCI post with other variables.

r	FCI pre	SAT total	SAT math	High school GPA	Total GPA	Math prereq. grade
N	700	488	488	635	617	276
FCI post	0.80	0.52	0.50	0.22	0.20	0.15

Based on the correlation coefficient, we use hierarchical linear regression methods to build a model to predict students' FCI post-test, with the intention to see if this dependent variable is affected by instruction type. Hierarchical linear regression is a technique to determine if the addition of an independent variable significantly improves percentage of the variance in the dependent variable. The statistical package R was used to conduct the analysis and instruction type (IT0, IT1, IT2) was converted to categorical variable in R.

Since the variables that correlate most strongly with FCI post are "FCI pre" and "SAT total", when running linear regression, we include these two variables and "Instruction Type (IT)" as predictors, adding one variable at a time. However, we lose 30% of the data by using "SAT total" (N=488). Since "FCI pre" variable itself predicts most of the variance in FCI post, it is also possible to use only "FCI pre" and "IT" to utilize the larger sample size. We run the regression in two ways:

1. Use only "FCI pre" and "IT" as the predictor (N=700).
2. Use "FCI pre", "SAT total" score and "IT" (N=488).

Both methods give similar results and below we list statistics for method 1 only.

TABLE II. Hierarchical linear regression analysis predicting FCI posttest percentage. B is the regression coefficient, SE the standard error, and the adjusted R^2 indicate the significance of the improved fit of the model over the model in which it is nested. "*" denotes p < 0.05, "**" denotes p < 0.01, and "***" denotes p < 0.001.

		B	SE	Adjusted R^2
Model 1	FCI pre	0.84 ***	0.02	0.64 ***
Model 2	FCI pre	0.85 ***	0.02	0.65 ***
	IT1	−0.41	0.43	
	IT2	1.03 **	0.39	

Model 1 shows that students' FCI pre-test is a significant predictor of the post-test. Model 2 use both FCI pre and instructor type as predictors and overall it is a statistically significant model. It shows that after controlling for students' differences in their preparation measured by FCI pre, instructor type still made a significant difference. Instruction Type 0 form the baseline for the regression and the regression coefficient measures the change with respect to this baseline. We can see the statistically significant difference occurs between IT2 (more interactive engagement) and IT0 (traditional) with the regression coefficient of IT2 being statistically significant but not IT1 (see table II). When using IT1 as the baseline, the regression coefficient of IT2 is still statistically significant. There is no significant difference between IT0 and IT1, while the difference only occurs when comparing IT2 to others.

However, despite being statistically significant, adding IT only added an extra 1% of the variance in FCI post score, indicating the effect size is small. Eta-squared [14] is used to characterize the effect size of certain variables in regression and the eta-squared for IT is 0.01. The table below listed the normalized gain for each group, with classes taught with more interactive type pedagogy (IT2) achieving higher gain than classes taught less interactive (IT0 and IT1). However, the normalized gain is relatively low across all sections, indicating there is still a lot of room for improvement.

TABLE III. FCI pre, FCI post and normalized gain (<g>) for different type of instructions.

	Total	IT0	IT1	IT2
N	700	139	205	356
FCI pre	12.5	12.0	13.6	12.1
FCI post	15.9	15.0	16	16.2
<g>	0.19	0.17	0.15	0.23

B. Correlation with SAT

In a previous study [10], pre-instruction SAT math scores and normalized gains (G) on the FCI were examined for individual students in interactive engagement courses in introductory mechanics at one high school (Edward Little High School, N=335) and one university (Loyola Marymount University (LMU), N=292), and strong, positive correlations were found for both populations (r=0.57 and r=0.46, respectively). It was suggested that these correlations are likely due to the importance of cognitive skills and abstract reasoning in learning physics. Normalized gain for each individual students (G) was computed in order to obtain correlation. Below we list correlation between SAT math, verbal and total score with normalized gain.

TABLE IV. Correlation (r) between FCI pre, post and G and SAT.

N=488	SAT math	SAT verbal	SAT total
FCI pre	0.49***	0.39***	0.50***
FCI post	0.49***	0.43***	0.52***
G	0.11*	0.16***	0.16***

Both FCI pre and post scores have a medium to strong, positive correlation with SAT scores. However, normalized gain correlates much less strongly with SAT, which is different than the previous study at LMU. There are several possible reasons for this difference. In the LMU study, it was suggested that the weaker correlation between SAT score and G for college students than high schoolers is likely due to the greater time delay between taking the SAT exam and the beginning of introductory mechanics for college students (either almost 2 years or about 3 and 1/2 years). It is possible that during that long delay the developmental experiences of students would vary which makes SAT a less accurate indicator of their initial state in a physics class. Based on this idea, one possible hypothesis for a much weaker correlation for CPP students is that students tend to take longer than four years to graduate since many of them also have to work part-time or full-time during college years (average 4-year graduate rates is only 25% for year 2012, 2013 and 2014). Therefore, there would be an even longer delay between the time SAT was taken and the time to take introductory physics. It is also possible that the stronger correlation between SAT and G only applies to that particular group of student body at LMU. For example, the class at LMU was taught using interactive engagement while we have a mix of instructional methods. It is possible the correlation weakens in larger, more heterogeneous student bodies. Results from educational studies can be population dependent and it is understandable this correlation would be different for another student body.

C. Gender and Ethnicity

To explore possible gender and racial gaps in FCI scores, we ask three sub-questions and below we organize results accordingly. Below we plotted students' FCI pre, FCI post and normalized gain (G) for different gender and ethnic groups.

Sub-question 1: Are there differences in FCI pre, FCI post and normalized gain between male and female?

As shown in the figures, there is a gender gap of 16% in initial preparation (measured by FCI pre) between male and female. This gap persists and increases slightly (17%) on the post test. There is a gap of 5% in normalized gain (G): average of the normalized gain). average of the normalized gain). The difference between male and female gains was calculated by Welch's two-sample t-test. Effect size was characterized by Cohen's d [15]: according to Cohen's convention d = 0.2 is considered a small effect, d = 0.5 is a medium effect, and d = 0.8 is a large effect. The difference between male and female on both FCI pre and post is statistically significant and the effect size is large. This is also on the larger side compared with the average gender gap (12%) that was previously reported [16].

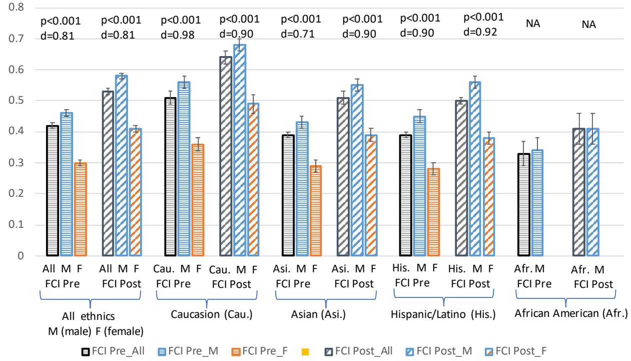

FIGURE I. FCI pre, post for different gender and ethnic groups.

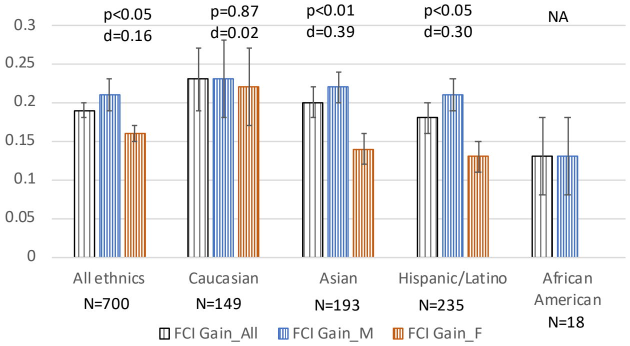

FIGURE II. FCI gain (G) for different gender and ethnic groups. P-value and Cohen's d are listed on the top of the graph.

This gender gap in both FCI pre and post persists across different races and ethnicities. The gender gap in initial preparation (FCI pre) is largest for Caucasians, while the largest gender gaps in the average normalized gain occur among Asian and Hispanic/Latino students. Due to the small sample size, we can't make meaningful conclusions for African-American students.

Sub-question 2: Are there differences in FCI pre-test, post-test scores and normalized gain for different ethnic groups?

Asian and Hispanic/Latino students are very similar in all measures (FCI pre, post, and gain). The average scores for the Caucasian group are higher than Asian/Hispanic students: 12% on FCI pre, 13% (Asian) and 14% (Hispanic) respectively on FCI post, and 3% (Asian) and 5% (Hispanic) on normalized gain (G). The largest racial gap occurs between Caucasian and African American students (18% difference on FCI pre, 20% on FCI post, and 10% on G). This pattern persists for both male and female.

Sub-question 3: If a gender difference exists in FCI posttest scores, is this gender difference the same for students of all races and ethnicities?

From the two figures we see there is a gender gap for all ethnicities, to further explore this question, we run a hierarchical linear regression.

TABLE V. Hierarchical linear regression analysis predicting FCI posttest percentage. Male formed the baseline and female was compared to male. Likewise, Caucasian formed the baseline and other ethnicities were compared to Caucasian. See Table II caption for definitions of B, SE, and Adj. R^2 (adjusted R^2).

		B	SE	Adj. R^2
Step1	Course Grade	0.08***	0.01	0.17***
Step2	Course Grade Female	0.07*** -0.16***	0.01 0.02	0.28***
Step3	Grade Female Asian Hispanic/Latino African American	0.06*** -0.16*** -0.08** -0.08** -0.14*	0.01 0.02 0.03 0.02 0.06	0.30***
Step4	Grade Female Asian Hispanic/Latino African American Female*Asian Female*Hispanic Female*African American	0.06*** -0.19*** -0.09** -0.08** -0.15* 0.03 0.03 NA	0.01 0.05 0.03 0.03 0.06 0.06 0.06 NA	0.29***

We converted the letter grade students received in this physics course to numerical grade in this analysis as follows: A (4.0), A- (3.7), B+ (3.3), B (3.0), B- (2.7), C+ (2.3), C (2.0), C- (1.7), D+ (1.3), D (1.0), F (0.0). Course grade alone predicts 17% of the variance in FCI post score (Step1). Female students score 16% lower on the FCI post than males, even after controlling for course grade (Step2). Adding gender improves the model significantly with 28% of the variances accounted for (Step2). After controlling for both course grade and gender, the racial gaps decrease a bit (compared to earlier results in sub-question 2) with Asian and Hispanic/Latino both score 8% lower than the baseline (Caucasian). African American scores 14% lower than the baseline, although the statistical power is limited with the small sample size (N=18). Even though adding ethnicity to the model explained only an additional 2% of variance in posttest, Step3 was a significantly better model than Step2 ($p < 0.001$).

Step4 introduced interactions between gender and ethnicity. This model did not explain significantly more variability in posttest average. The interaction terms in this model were also not statistically significant. After controlling for course grade, the gender gap was the same for Caucasian, Asian and Hispanic students. We could not list any analysis for African American because we only had one female student in the sample.

IV. DISCUSSIONS AND CONCLUSIONS

In order to provide valuable evidence for reproducibility of educational studies for a variety of student populations and explore gender and racial gaps, we conducted a year-long research at CPP. We collected FCI data for all four terms in introductory physics for science and engineering majors. First of all, FCI pre is found to be the strongest predictor of FCI post scores. Using a linear regression analysis, we showed that instruction type (IT) is also a statistically significant predictor of FCI post, while the difference only occurs when comparing the more interactive class (IT2) to other less interactive ones (IT0 or IT1). The model with IT as a variable is statistically a better model, with the more interactive type of instruction results in a higher FCI gain than the traditional lectures. Given the typical gains reported with active-learning classes [17-18], there is still a lot of room for improvement at our institution. The second research question is to investigate correlation between FCI gain and SAT scores. Unlike the stronger correlation found in the previous study (r=0.46) at LMU, we found normalized FCI gain correlates weakly with SAT total (r=0.16). This result can be explained by the more diverse population we have, where we included students from all four quarters, all types of instructions and different years in their undergraduate career. Last but not least, gender and racial gaps were found in students' FCI performance, both at initial preparation and overall gain. There is a gender gap of 16% FCI pre-test and 17% of FCI post-test with a strong effect size (d=0.81). This is consistent with what was previous reported (12% on average) but on the large side. There is a gap of 5% in normalized gain (G: average of the gain) with a small effect size (d=0.16) without controlling for course grade. The last analysis of hierarchical linear regression allows us to investigate the gender gap with course grade controlled for. Gender gap stays the same even after controlling for course grade, racial gaps decreases a bit but still exists after controlling for course grade and gender (Table V). This suggests that great caution is needed when using these diagnostic tools given the persistent gender and racial gaps. Compared to previous research [8], our results followed a similar pattern while showing larger gender and racial gaps overall. Future work can include further investigation about equity and the effectiveness of possible pedagogical interventions at our institution.

The authors wish to thank all instructors and students involved and the department of physics and astronomy at Cal Poly Pomona. Special thanks to department chair Dr. Hector Mireles and staff member Hector Maciel.

[1] Bransford, J., Brown, A., Cocking, R. (Eds) (2000). How People Learn: Brain, Mind, Experience, and School, National Academy Press, Washington, DC.

[2] CPP IRPA, https://www.cpp.edu/~data/cpp-facts.shtml

[3] Loken, E. and Gelman, A., (2017). Measurement error and the replication crisis, *Science* 10 Feb 2017: Vol. 355, Issue 6325, pp. 584-585

[4] R. K. Thornton and D. R. Sokoloff, Assessing student learning of newtons laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, *Am. J. Phys.* 66, 338 (1998).

[5] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, *Phys. Rev. ST Phys. Educ. Res.* 2, 010105 (2006).

[6] 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).

[7] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, *Phys. Teach.* 30, 141 (1992).

[8] A. L. Traxler, X. C. Cid, J. Blue, and R. Barthelemy, Enriching gender in physics education research: A binary past and a complex future, *Phys. Rev. Phys. Educ. Res.* 12, 020114 (2013).

[9] R. Henderson and J. Stewart, Racial and ethnic bias in the Force Concept Inventory, *Physics Education Research Conference* 2017, Cincinnati, OH: July 26-27, 2017, Pages 172-175

[10] Coletta, V. P., Phillips, J. A., & Steinert, J. J., Interpreting force concept inventory scores: Normalized gain and SAT scores. *Phys. Rev. ST Phys. Educ. Res.* 3, 010106 – Published 23 May 2007

[11] "2018 SAT Suite of Assessments Annual Report" *College Board.* Retrieved October 28, 2018.

[12] Willoughby,. S., Exploring Gender differences with different gain calculations in astronomy and biology. *AJP*, 77, 651, 2009.

[13] James Day, Jared B. Stang, N. G. Holmes, Dhaneesh Kumar, and D. A. Bonn, Gender gaps and gendered action in a first-year physics laboratory, *Phys. Rev. Phys. Educ. Res.* 12, 020104 – Published 1 August 2016

[14] Richardson, J.T.E., Eta squared and partial eta squared as measurements of effect size in educational research. *Educational Research Review, 6*, 135-147.2011

[15] J. Cohen, Statistical Power Analysis for the Behavioral Sciences (Routledge, NY, '88), 2nd ed.

[16] A. Madsen, S. B. McKagan, and E. C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Phys. Rev. Phys. Educ. Res.* 9, 020121 (2013)

[17] Bransford, J., Brown, A., Cocking, R. (Eds) (2000). How People Learn: Brain, Mind, Experience, and School, National Academy Press, Washington, DC,

[18] L. McDermott and E. Redish. (1999) Resource Letter: PER-1: Physics Education Research, Am. J. Phys. 67, 755 .