Association between fundamental skills on physics pre-course assessment and post-course test performance

Cherish Bauer-Reich, University of Jamestown

Cherish Bauer-Reich is an Associate Professor at the University of Jamestown in Jamestown, ND. She is also a senior member of IEEE.

Jakob DeLong, University of Jamestown

Jakob DeLong is currently an assistant teaching professor in the Electrical Engineering department at Missouri University of Science and Technology. He received a BS in Electrical Engineering from West Virginia University and a PhD in Electrical Engineering from The Ohio State University, while working as a GRA at the ElectroScience Lab. He previously taught in the engineering department at the University of Jamestown and assisted in course development and revision.

Dr. Wesley Newton, University of Jamestown

Wesley Newton, PhD, retired after 35 years working as a statistical consultant, analyst, and researcher for various federal agencies and universities. Wesley is currently an adjunct Math Instructor within the Engineering, Mathematics, and Environmental Sciences program at the University of Jamestown, Jamestown, ND, teaching mathematics and statistics courses.

Association between fundamental skills on physics pre-course assessment and post-course test performance

Abstract:

Early intervention is beneficial to student outcomes in any course. Determining which skills are most critical for student success may enable more rapid identification and intervention for students who require additional assistance. In this study, the authors used two assessment tools: the Force Concept Inventory (FCI) and the Conceptual Survey of Electricity and Magnetism (CSEM). These tests were administered in calculus-based physics courses to gauge student knowledge at entry and exit. While these tests are useful for obtaining information about student growth, they have typically not been useful as predictive tools. For this study, the authors examined specific fundamental skills assessed on these tests to determine if entrance performance on questions assessing those skills was an indicator of future performance. The authors divided questions into three categories: verbal, diagrammatic, and complex (requiring understanding of both language and image interpretation). Data from tests administered between 2017 to 2024 were analyzed to determine if any predictive relationship existed between entry and exit scores. Further, the authors assessed whether any relationships were associated with demographic variables like gender, major, and grade level. Verbal and complex pre-test scores accounted for some of the outcome on post-test scores on the FCI but were not predictive of raw or normalized gain. Diagrammatic questions did not seem to contribute to the change between pre- and post-test scores, although this may be a result of the low number of questions in that category. All three categories contributed to post-test measurements on the CSEM and accounted for some of the change seen in all three types of post-test measurements.

Introduction

One of the most commonly used assessment tools is the Force Concept Inventory (FCI). [1] Since its introduction in 1992, it has been used as a learning assessment tool for physics classes worldwide, while its use has been heavily studied. During this time, researchers have evaluated the tool to understand whether there are some questions on the test that may be biased due to gender or present hurdles to those for whom English is not their primary language. [2] [3] [4]

We have used the FCI for nearly a decade in physics assessment for college- and university-level physics courses, and during this time, have observed some of the difficulties that students have with understanding the wording on the assessment. Observationally, students tend to be intimidated by the test, particularly when the questions appear to involve a significant amount of reading. This prompted questions regarding whether the language used on the exam was a barrier to doing well on the assessment as well as whether the FCI may also be assessing

language skills and not just physics understanding. This question is particularly relevant in light of recent assessments indicating that, in the USA, reading scores have been dropping nationwide. [5]

To answer this question, we chose to evaluate both the FCI and the Conceptual Survey of Electricity and Magnetism (CSEM) in terms of how much the questions depend on language, graphical representations, and a combination of the two. We hypothesized that a strong dependence on language use in the assessment would result in measurable differences in student performance in each category. We also hypothesized that these differences would result in post-test scores or gains that would depend more heavily on questions from the verbal category, i.e. questions that did not contain graphical representations. Finally, we anticipated that gender and grade level may result in differences in student performance in one or more of the categories.

Method and Results

Force Concept Inventory

We used test results from both the FCI and the CSEM. These were administered in Physics I and Physics II courses, respectively, between the years 2017 and 2024. These courses are calculus-based physics courses taken primarily by students in science and engineering majors at the University of Jamestown. Both courses were taught by the same instructor during this time. Physics I used a flipped-course format and Physics II used a lecture format.

We analyzed the questions in both examinations and broke them into three categories: verbal, diagrammatic, and complex. Verbal questions were those that only required the written text to solve. Some verbal questions included diagrams, but these problems were considered verbal if the diagram did not provide any information that was not explicitly stated in the text. Diagrammatic questions were questions which had minimal verbiage requiring the student to interpret a diagram. Complex questions were a mixture of both: the questions included some information in the text and some in illustrations, and the problem could not be solved without information from both the text and illustrations. Because of how these are defined, each question can only belong to a single category. Table 1 shows the categorization of questions for both the FCI and CSEM.

Table 1: Categorization of FCI and CSEM questions

Category	FCI Questions	CSEM Questions
Verbal	1-4, 13, 15-17, 25-30	1-5, 10, 11, 13, 14, 16, 21, 24
Diagrammatic	6, 7, 12, 14	6, 12, 15, 22, 23, 26, 28
Complex	5, 8-11, 18-24	7-9, 17-20, 25, 27, 29, 31, 32

Subtotal scores were determined by adding single points for correct answers in that subsection and dividing by the total number of questions in the subsection. For example, if a

student correctly answered two of the four diagrammatic questions on the FCI, their total diagrammatic score would be 0.5.

We used a generalized linear modeling approach, beta regression, to assess the relationship between post-assessment scores with pre-assessment scores. We assumed the distribution of the post-scores followed a beta distribution because the values are percents summarized as proportions and constrained in the interval 0 to 1 exclusive of 0 and 1. [6] We used a logit link function to transform the post-scores prior to modeling. Because multicollinearity also can be an issue among predictor variables in beta regression, we developed a suite of a priori plausible models that excluded within each model pre-assessment scores that showed strong simple correlations ($r>0.7$). [7] The models chosen are shown in Table 2.

Table 2: A priori models developed avoiding potential multicollinearities among predictor variables and subject matter expertise.

Model	Predictor Variables
Null	Null
1	Verbal
2	Diagrammatic
3	Complex
4	Verbal + Diagrammatic
5	Verbal + Complex
6	Diagrammatic + Complex
7	Pretest
8	Diagrammatic + Gender + Gender×Diagrammatic
9	Complex + Gender + Gender×Complex
10	Verbal + Gender + Gender×Verbal
11	Diagrammatic + Level + Level×Diagrammatic
12	Complex + Level + Level×Complex
13	Verbal + Level + Level×Verbal
14	Pretest + Gender + Gender×Pretest
15	Pretest + Level + Level×Pretest

The predictor variables included the FCI verbal pretest score (Verbal), FCI diagrammatic pretest score (Diagrammatic), FCI complex pretest score (Complex), FCI total pretest score (Pretest), Gender (Gender), and grade level (Level). The FCI total pretest score was positively correlated with verbal, diagrammatic, and complex pretest scores, so none of the models included a linear combination of the total pretest score with any of the subscores. Figure 1 shows the bivariate scatterplot among the four pretest scores. These show the relationship between the pretest scores, indicating that the FCI total score should not be used in the same

model as the category pretest scores. No significant differences were detected between males and females, nor among grade levels, for any of the predictor variables (all p >0.05). Hence, we included two-way interactions between gender by each predictor variable and level by each predictor variable. Small sample sizes precluded using gender and level in the same model.

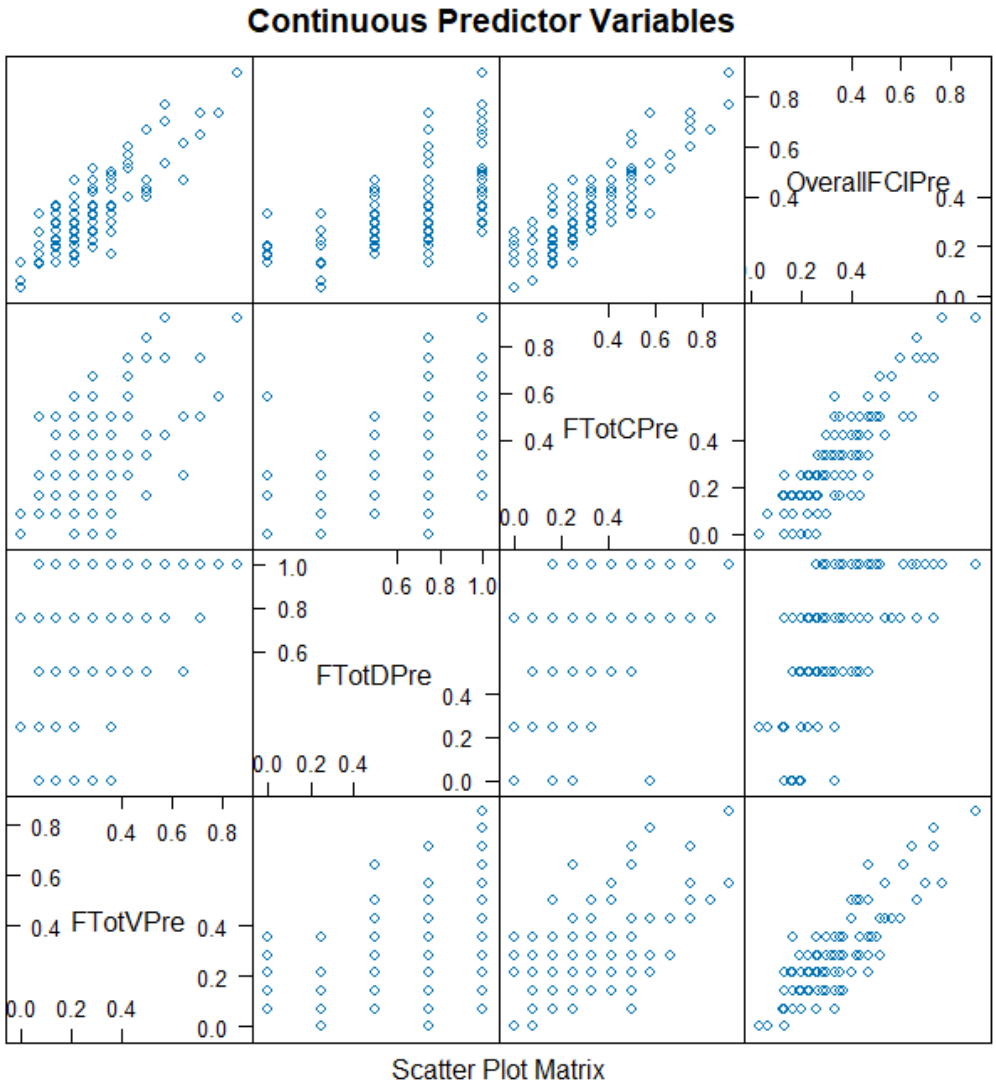

Figure 1: Bivariate scatter plot among continuous predictor variables.

We used the model selection approach prescribed in [8] to rank the models using Akaike's Information Criteria (AIC_C). For the models with the lowest AIC_C values, we computed the correlation between the predicted post-scores from the model and the actual post-scores to assist with interpreting model fit. [9] Evaluation of the selected models was repeated using the SPSS Automatic Linear Modeling function to examine simple linear regression models. The results of this analysis are shown in Table 3. It should be noted that models 14 and

15 did not find significance in the interactions when modeled using simple linear models and reduced to model 7. Therefore, a comparison was only made among models 5-7.

Table 3: Summary of models and AIC_C model selection results using beta regression and comparison with simple linear models (last two columns) for prediction of FCI post-test scores. Models with $\Delta AIC_C > 15.47$ were excluded.

Model$_i$	**K$_i$**	**AIC$_C$**	**Δ_i**	**r$_i$**	**RSS**	**Linear r$_i$**	**Linear AIC$_C$**	**Linear Δ_i**
5	4	-121.54	1.77	0.62	2.237	0.60	-485.84	2.17
6	4	-107.47	15.47	0.49	3.038	0.61	-472.43	15.58
7	3	-123.31	0.00	0.62	2.733	0.61	-488.01	0
13	5	-105.08			3.052			
14	5	-119.90	3.41		2.721			
15	5	-119.98	3.33		2.725			

Both methods of analysis implied that model 7 was preferred model. Model 5 accounted for a similar amount of variance but required more parameters. The comparison between methods indicated that the simple linear models were sufficiently accurate for the purposes of this study. The generalized linear model accounts for 38% of the change from pre- to post-test scores, while the simple linear model accounts for 37% of the change. These values match other studies such as [10] that indicate that other factors such as motivation account for the rest of the change pre-test to post-test scores.

The simple linear regression for model 7 predicts FCI Post-test Score = 0.250 + 0.796 × FCI Pre-test Score. The significance of both coefficients was $p<0.001$ with n=128. For the model, $F_{1, 126}=75.643$ and $R^2=0.37$. Figure 2 shows the comparison of post-test predicted scores, based on this model, with actual post-test scores. Simple linear regression of Model 5 predicted FCI Post-test Score = 0.268 + 0.425 × Complex pre-test score + 0.405 × Verbal pre-test score. The significance of all three coefficients was $p<0.001$ with n=128 with $F_{2, 125}=37.470$ and $R^2=0.365$. It was assumed that the constant value was a result of instruction as both models had similar constants.

Similar procedures were followed to determine whether pretest scores and subtest scores were predictor variables for the FCI gain, defined as the difference between post- and pre-test scores, and normalized gain. [11] These relationships did not appear to be sufficiently strong, so these were not considered in the remainder of the analysis.

We next attempted to determine if similar relationships existed between particular concept categories and the final post-test score. To do this, we examined the concepts tested by the FCI and determined which problems applied to each of the categories. The concepts tested by the FCI include Newtons first law (FL), Newton's second law (SL), Newton's third law (TL), kinematics (lumped masses undergoing translation or rotation) (Kin), superposition of forces and

fields (SP), and kinds of forces (gravitational, mechanical, etc) (For). The problems assigned to each category are given in Table 1. The resulting problem distribution is shown in Table 4. It should be noted that not all categories are represented in the FCI, and some problems fall into more than one category.

There were four categories of problems that were found to be significant predictors of post-test FCI scores at the 95% confidence level: complex kinematics ($p<0.001$), verbal third law ($p=0.002$), verbal kinds of forces ($p=0.013$), and complex first law ($p=0.021$). The regression predicted that FCI Post-scores = 0.277 + 0.236 × complex kinematics + 0.161 × verbal third law + 0.205 × verbal kinds of forces + 0.127 × complex first law. The model had n=128, $R^2=0.378$, and $F_{4, 123}=24.386$.

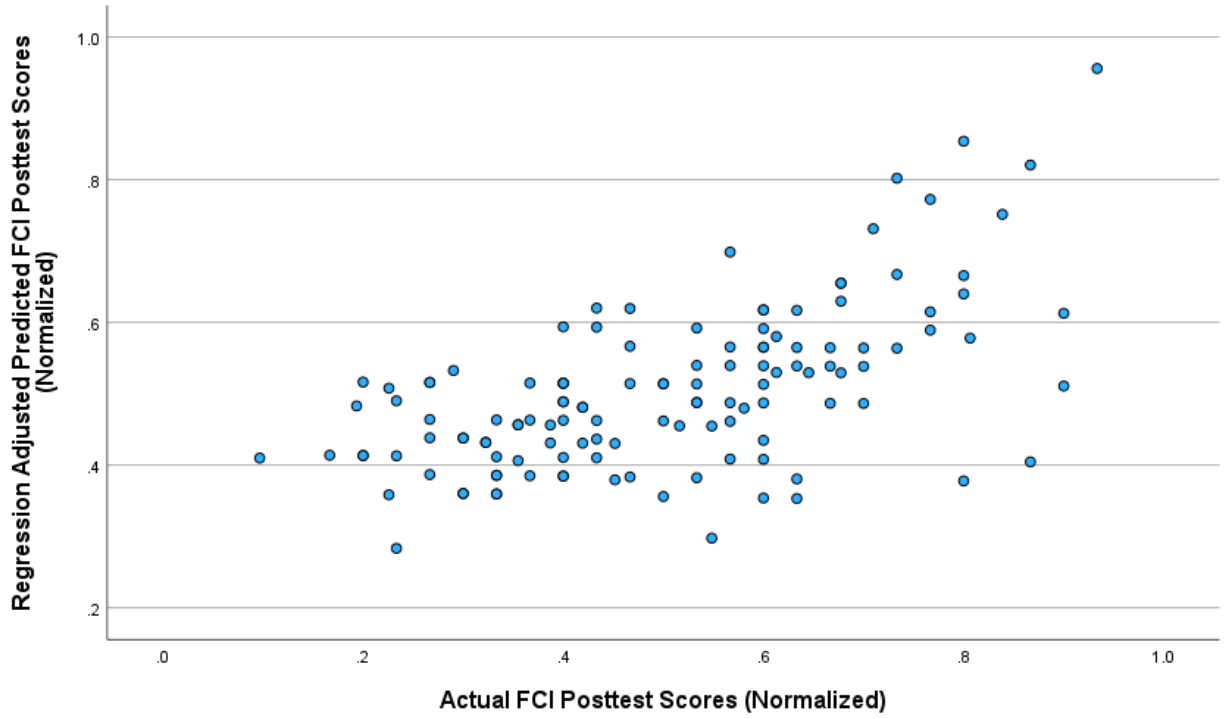

Figure 2: Scatterplot showing the predicted scores from regression vs actual post-test scores from model 7.

Conceptual Survey of Electricity and Magnetism (CSEM)

To evaluate CSEM scores, three sets of analyses were performed. Regressions were performed with final post-test score, raw gain, and normalized gain as the dependent variables. For each of the dependent variables, three sets of predictors were examined. The first was the pre-test final score, the second included the pre-test category scores, and the final looked at the conceptual categories. The categorization of the CSEM questions into verbal, diagrammatic, and complex was shown in Table 1. Table 5 shows how these were sorted into concepts. The results of this analysis are shown in Table 6 for n=67.

Table 4: Problems associated with each concept at category on the FCI. Significant predictors are highlighted.

Concept	Category	Problems
First Law (FL)	verbal	17, 25
	diagrammatic	6, 7
	complex	8, 10, 11, 23, 24
Second Law (SL)	verbal	26
	complex	8, 9, 21, 22
Third Law (TL)	verbal	4, 15, 16, 28
Kinematics (Kin)	diagrammatic	12, 14
	complex	9, 19-22
Superposition (SP)	verbal	17, 25
	complex	8, 9, 11
Kinds of forces (For)	verbal	1-3, 13, 17, 27, 29, 30
	diagrammatic	12, 14
	complex	5, 11, 18

Of note is that some of the predictor variables are inversely related to the dependent variable. This is generally considered to be the result of guessing. This occurs when students guess correctly on the pretest but then choose the wrong answer accidentally on the post-test. Some analyses can remove these answers when assessing normalized gain. However, they were not removed for this study.

Discussion

As has been noted in previous literature, the FCI is not a good predictive tool for future performance on the FCI. In particular, we noted that the FCI pre-test did account for some of the change in score, as was seen in Models 5 and 7. However, we did not find that there was any relationship between pre-test scores and raw or normalized gains. When looking at both the models incorporating the category and subfield pre-test scores, the verbal and complex categories seemed to provide the best prediction for the post-test scores as none of the predictor variables was associated with the diagrammatic category. It should be noted that these three categories comprised 17 of the 30 questions in the examination. There was no significant relationship between any of these variables and either gender or grade-level.

These results may suggest that students' language skills may be a factor in their performance, and the FCI may be measuring, through a change in score, how well a student adapts to the language and terminology used in physics. This may be a function of background exposure to scientific language prior to enrollment in the course, but it may also be a function of students' ability or desire to learn this language. If this is the case, then improving student language skills may be one way to improve student outcomes on the FCI.

The CSEM pre-test scores, on the other hand, accounted for some level of variation in post-test scores, raw gain, and normalized gain in most of the models. In particular, most models indicated that diagrammatic pretest scores are positively correlated with performance on the post-test while language and complex skills on the pretest were generally negatively correlated. We did not examine gender or grade-level for the CSEM because of the small sample size.

Table 5: Problems associated with each concept category on the FCI.

Concept	Category	Problems
E-field force (EFor)	Verbal	10, 11
	Diagrammatic	12, 15
	Complex	19, 20
Work, potential, field (WP)	Verbal	11, 16
	Complex	17-20
Charge distribution (CD)	Verbal	1, 2, 13
Coulomb's law (CL)	Verbal	3-5
E-field superposition (ESup)	Complex	7-9
	Diagrammatic	6
Induced charge (IC)	Verbal	13, 14
	Complex	25, 27
Magnetic force (MF)	Verbal	21
	Diagrammatic	22
	Complex	25, 27, 31
Magnetic field from current (BFC)	Verbal	24
	Diagrammatic	23, 26, 28
Magnetic field superposition (BS)	Diagrammatic	23, 28
Faraday's Law (CFL)	Diagrammatic	30
	Complex	29, 31, 32
Third Law (CTL)	Verbal	4, 5, 24

An important difference between students who took the CSEM versus those who took the FCI is that one must finish Physics I to complete Physics II, where the CSEM is administered, but the requirement for entry to Physics I is co-registration in the first calculus class. Therefore, the bar to entry is higher for Physics II than to Physics I, and the larger variance in student background and/or ability may explain why the FCI is not as good at predicting how students do on the post-test. This is not a trivial issue because these factors are very difficult to incorporate into a predictive model for student success, and it is clear that those factors play a bigger role in

student performance on these exams than the physics understanding or categorical abilities measured on the FCI. [10]

Table 6: Regression models to predict post-test score, raw gain, and normalized gain on the CSEM.

Dependent Variable	Model	R^2	AIC_C	Predictors	β	p-level
Post-test score	Pre-test total	0.228	-290.99	Constant	0.181	<0.001
				CSEM Pretest total	0.721	<0.001
	Categories	0.341	-301.50	Constant	0.259	<0.001
				Diagrammatic	0.439	<0.001
	Subfields	0.533	-319.60	Constant	0.802	<0.001
				ESup, Diagrammatic	-0.103	<0.001
				BFC, Diagrammatic	-0.120	0.001
				CL, Verbal	-0.152	0.007
				ESupA, Complex	-0.156	0.027
Raw Gain	Pre-test total	0.053	-294.62	Constant	0.192	<0.001
				CSEM Pretest total	-0.335	.034
	Categories	0.156	-299.99	Constant	0.222	<0.001
				Verbal	-0.267	0.019
				Complex	-0.330	0.014
				Diagrammatic	0.156	0.048
	Subfields	0.448	-327.99	Constant	-0.329	<0.001
				CFL, Complex	0.167	<0.001
				EFor, Verbal	0.218	0.001
				MF, Verbal	0.074	0.001
				CL, Verbal	-0.051	0.033
				Esup, Diagrammatic	-0.052	0.018
				Efor, Complex	0.104	.019
Normalized Gain	Pre-test total			No relationship		
	Categories	0.108	-256.09	Constant	0.222	<0.001
				Diagrammatic	0.291	.009
	Subfields	0.431	-284.93	Constant	-0.316	0.003
				CFL, Complex	0.237	<0.001
				ESup, Diagrammatic	-0.117	<0.001
				MF, Verbal	0.110	<0.001
				EFor, Verbal	0.262	0.003

There is a small number of diagrammatic questions on the FCI when compared with the CSEM, however, and it is possible that a drawback to using the FCI as an assessment tool is a potential overreliance on language-based questions to determine students' understanding. Students in majors who require physics generally have been found to have higher visual-spatial skills than students in other majors. [12] The CSEM assesses this skill area more thoroughly than

the FCI, and we see that these questions contribute to the prediction models developed with these categories. It therefore seems plausible that one way to improve the predictive ability of the FCI would be to incorporate more questions relating to diagrammatic aspects of understanding Newtonian mechanics. Research has already been done on shortened versions of the FCI that map to questions on the full FCI and show comparable assessment value. Neimenin et. al. chose 9 questions (7 verbal and 2 complex) and created representational versions of these questions, while Han et. al. created two shorted versions of the FCI by selecting a subset of questions. [4] [13] Interestingly, the shortened versions used by Han et. al. each consisted of 6 verbal, 6 complex, and 2 diagrammatic questions, so these were approximately the same ratio as represented in the original FCI. Future work on this topic should examine whether additional diagrammatic questions on either the FCI or a shortened version provide better predictions of student performance after taking a course. A modified version of this test could better answer whether the test is measuring all relevant skills accurately or if students need more instruction in the areas that are being assessed with the FCI.

Conclusion

We attempted to determine if student performance on the FCI was affected by performance on questions with a heavy verbal focus. We did this by examining models of student performance using both beta regression and linear regression. Both methods produced similar results. The preferred models seemed to correlate weakly with verbal and complex questions but not with diagrammatic questions. This effect was not seen on the CSEM where pre-test performance on diagrammatic questions were positively correlated with final test scores but complex and verbal had more complicated predictive ability. However, it is unclear if this is a result of the small number of diagrammatic questions on the FCI resulting in an effect too weak to measure or due to students' interaction with language on the verbal and complex questions. Further work using modifications to the FCI may be necessary to answer this question.

Appendix I

Table 7 provides descriptive statistics for pre- and post- test scores on the FCI and CSEM. These statistics represent only matched scores, i.e., those values for which the student has both pre- and post-test scores. Table 8 provides score information for both instruments based on gender and grade level. One student was enrolled as a high school student and omitted from grade level data. For some students, the grade level information on the CSEM was no longer available and therefore was omitted.

Table 7: Statistical information describing test score categories.

Test score	n	Mean	St. Dev.	Minimum	Maximum
FCI					
Verbal Pretest	128	0.271	0.152	0	0.857
Diagrammatic Pretest	128	0.574	0.314	0	1
Complex Pretest	128	0.302	0.186	0	0.916
Overall Pretest Score	128	0.321	0.147	0.033	0.900
Overall Posttest Score	128	0.502	0.187	0.097	0.933
Raw Gain	128	0.185	0.151	-0.133	0.667
Normalized Gain	128	0.275	0.213	-0.200	0.850
CSEM					
Verbal Pretest	67	0.246	0.123	0	0.667
Diagrammatic Pretest	67	0.205	0.175	0	0.750
Complex Pretest	67	0.223	0.101	0	0.417
Overall Pretest Score	67	0.234	0.092	0	0.500
Overall Posttest Score	67	0.248	0.128	0.125	0.688
Raw Gain	67	0.115	0.112	-0.219	0.344
Normalized Gain	67	0.149	0.151	-0.333	0.458

Table 8: Statistical information describing FCI test scores by gender and grade level.

		Mean (SD)						
		Pretest						
Group	n	Verbal	Diagrammatic	Complex	Overall	Posttest Overall	Raw Gain	Normalized Gain
FCI								
Gender								
Male	102	0.288 (0.158)	0.608 (0.308)	0.315 (0.199)	0.339 (0.153)	0.521 (0.185)	0.185 (0.146)	0.282 (0.207)
Female	26	0.203 (0.100)	0.442 (0.311)	0.253 (0.117)	0.253 (0.097)	0.438 (0.178)	0.185 (0.172)	0.247 (0.240)
Grade Level								
Freshmen	78	0.291 (0.163)	0.590 (0.312)	0.312 (0.193)	0.333 (0.155)	0.506 (0.203)	0.177 (0.157)	0.270 (0.228)
Sophomores	27	0.249 (0.135)	0.546 (0.286)	0.262 (0.147)	0.291 (0.116)	0.511 (0.149)	0.220 (0.141)	0.309 (0.188)
Juniors	11	0.214 (0.096)	0.455 (0.270)	0.348 (0.148)	0.299 (0.103)	0.468 (0.162)	0.165 (0.149)	0.236 (0.213)
Seniors	11	0.208 (0.108)	0.614 (0.424)	0.295 (0.176)	0.290 (0.136)	0.483 (0.163)	0.193 (0.127)	0.275 (0.181)

References

[1] D. Hestenes, M. Wells, and G. Swackhamer, “Force concept inventory,” *The physics teacher*, vol. 30, no. 3, pp. 141-158, Mar. 1992.

[2] R. Henderson, J. Stewart, and A. Traxler, “Partitioning the gender gap in physics conceptual inventories: force concept inventory, force and motion conceptual evaluation, and conceptual survey of electricity and magnetism,” *Phys. Rev. Phys. Ed. Res.*, vol.15, May 28, 2019, doi: https://doi.org/10.1103/PhysRevPhysEducRes.15.010131

[3] Y. Shoji, S. Munejiri, and E. Kaga, “Validity of force concept inventory evaluated by students’ explanations and confirmation using modified item response curve,” *Phys. Rev. Phys. Ed. Res.*, vol. 17, Sept. 20, 2021, doi: https://doi.org/10.1103/PhysRevPhysEducRes.17.020120

[4] P. Nieminen, A. Savinainen, and J. Viiri, “Force concept inventory-based multiple-choice test for investigating students’ representational consistency,” *Phys. Rev. Phys. Ed. Res.*, vol. 6, Aug. 25, 2010, doi: https://doi.org/10.1103/PhysRevSTPER.6.020109

[5] C. C. Miller, S. Mervosh, and F. Paris, “Students are making a ‘surprising’ rebound from pandemic closures. But some may never catch up,” New York Times, Jan. 31, 2024. [Online]. Available: https://www.nytimes.com/interactive/2024/01/31/us/pandemic-learning-loss-recovery.html

[6] S.P. Ferrari, and F. Cribari-Neto, “Beta regression for modeling rates and proportions,” *Jour. of App. Stats.*, vol. 31, no. 7, pp. 799-815, 2004.

[7] R. H. Myers, D. C. Montgomery, G.G. Vining, T.J. Robinson, *Generalized Linear Models: with Applications in Engineering and the Sciences, 2nd ed*, Hoboken, NJ, USA: John Wiley and Sons, Inc., 2010.

[8] K.P. Burnham, and D.R. Anderson, “A Practical Information-Theoretic Approach,” in *Model Selection and Multimodel Inference, 2nd ed*. New York, NY, USA: Springer, 2002, pp. 75-117, doi: https://doi.org/10.1007/978-1-4757-2917-7_3

[9] G. Pineiro, S. Perelman, J. P. Guerssham JP, J.M. Paruela. “How to evaluate models: observed vs predicted or predicted vs observed,” *Ecological Modeling*, vol. 216, no. 3-4, pp. 316-322, 2008.

[10] G. Taasoobshirazi and G. M. Sinatra, “A structural equation model of conceptual change in physics,” *Jour. Of Res. In Sci. Teaching*, vol. 48, no. 8, pp. 901-918, June, 2011.

[11] R. Hake, “Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses,” *Am. Jour. Phys.*, vol. 66, no. 1, 1998, pp. 64-74, doi: https://doi.org/10.1119/1.18809

[12] M. Kozhevnikov, M.A. Motes, and M. Hegarty, “Spatial visualization in physics problem solving,” *Cognitive Sci.*, vol. 31, pp. 549-579, 2007.

[13] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, “Dividing the Force Concept Inventory into two equivalent half-length tests,” *Phys. Rev. ST Phys. Educ. Res.*, vol. 11, May, 2015, doi: https://doi.org/10.1103/PhysRevSTPER.11.010112