Evaluating students' self-efficacy with respect to computational methods in the undergraduate physics curriculum

Greg Gallagher[1], Danka Maric[2] Gautam Vemuri[1], Andrew Gavrin[1]
[1]*Department of Physics, Indiana University Indianapolis, 402 N Blackford St, Indianapolis, Indiana, 46202 and*
[2]*STEM Education Innovation & Research Institute, Indiana University Indianapolis, 755 W. Michigan Street, Indianapolis, Indiana, 46202*

Computational methods of problem-solving have begun to be a focal point of many physics departments' curricula. However, there are few published tools available to support evaluation of these initiatives. We present preliminary results of our efforts to evaluate a department-wide initiative to incorporate computational methods across the undergraduate physics curriculum. In particular, we focus on changes in students' self-efficacy with respect to computational techniques as they progress through our program. By using a survey instrument that asks students to rate both their *present* and *initial* abilities, we attempt to distinguish changes based on increases in their self-efficacy from changes in their understanding of the complexity and potential of these techniques. We find that as students progress, they consistently decrease their estimates of their initial abilities. We also define several potential measures of "growth" in self-efficacy and explore the differences in these measures.

2024 PERC Proceedings edited by Ryan, Pawl, and Zwolak; Peer-reviewed, doi.org/10.1119/perc.2024.pr.Gallagher

I. INTRODUCTION

Introducing computational methods in undergraduate physics has become a focus in recent years [1], [2]. This effort gained importance after a recent report by the American Association of Physics Teachers (AAPT) suggesting methods of integrating computational methods into undergraduate physics programs [3]. These suggestions were soon reinforced by an APS-AAPT joint task force which emphasized the importance of computational skills for physics majors as they look toward future careers [4]. The creation of the PICUP project [5] also helped promote these efforts.

While most faculty agree with the importance of physics students gaining computational skills through the undergraduate curriculum, there are relatively few reports of department-wide efforts to accomplish this described in the physics education literature. Some notable exceptions to this are the physics programs at Oregon State University [6] and Lawrence University [7], which have both sought to incorporate computation holistically across the undergraduate curriculum. Compounding the issue is a lack of well-established instruments to evaluate the effectiveness of efforts to integrate computational methods into the undergraduate curriculum. To date, no such instruments are available through the PICUP project site [5] or PhysPort [8].

A. Prior work

Given the lack of a suitable instrument in the literature, we developed the survey used in this paper to evaluate our initiative to "make computation normal" across the undergraduate physics curriculum. The overall goal is for graduating students to consider computational approaches to be a "normal" part of physicists' work, and to be comfortable using a variety of computational methods. To accomplish this, the Physics Department at Indiana University Indianapolis set a goal that $\geq 25\%$ of all undergraduate work be computational. We have previously reported details of the development process and preliminary efforts to establish its validity and reliability[9]. This preliminary work established content validity by involving experts through several rounds of a review process. Internal reliability was established via split-half correlation coefficients for each construct, all with values > 0.8 indicating reasonable internal consistency. A factor analysis, further establishing the validity of our instrument, is detailed in another paper submitted to PERC 2024 [10].

The survey has two parts. Part 1 is intended to measure students' agreement with general statements related to the role of computational methods in physics, e.g., "Computational, experimental, and analytical ("pencil and paper math") methods are all necessary in the field of physics" (five-point Likert scale). Part 2 is intended to measure students' self-efficacy with respect to specific computational tools and methods. Because we were concerned that students "don't know what they don't know," this second part consists of two sub parts. Part 2a asks students to rate their *present* skills on, e.g., numerical integration, on a 1-10 scale. Part 2b specifies the same topics, but asks students to rate their *initial* skills defined as "at the time they began the program."

B. Present goals

This paper seeks to expand on our previous work by focusing on analysis of students' changing self-efficacy, as measured by part 2 of the survey, where we assess self-efficacy based on students' self-reported computational abilities. In particular, we investigate the way students' assessments of their "initial abilities" change over time, and the implications for characterizing the results of their experiences in the program. We center our work on two research questions:

RQ1 Do students reevaluate their prior abilities as they progress through the program?

RQ2 If so, can this reevaluation be taken into account in characterizing the change in their self-efficacy over time?

C. Institutional context

IU Indianapolis (formerly IUPUI) is an urban, public university enrolling approximately 25,000 students. Its Carnegie Classification is currently "R2: High Research Activity." [11]. Approximately 30% of undergraduate students are first generation college students. Also, approximately 30% are people of color. The Physics department currently has 13 full-time faculty and offers B.S., M.S., and Ph.D. degrees. The undergraduate curriculum follows a traditional model, including a two-course introductory sequence, two upper-level labs, six required advanced courses, and a research capstone.

II. METHODS

A. Survey method and participants

All data described in this paper were acquired using the survey described above. Near the end of each spring and fall semester, all students completing a physics majors' course are asked to complete the survey. They receive two prompts via email and are asked to complete the survey online. The survey was first implemented at the end of the fall semester of 2018 and has been given each semester since. Some changes to the Likert scale portion were made to enhance clarity after the Spring 2020 semester, while the self-efficacy items discussed in this paper have remained unchanged.

The survey is administered electronically by one of the authors (MD) who is not a member of the physics department. Students are not given an incentive to complete the survey, and are informed that their participation is optional and can be withdrawn at any time. We are aware that this mechanism

can produce some level of self-selection bias, and acknowledge that this should be considered in drawing conclusions from our results.

The wording used in part 2a of the survey is "Please rank your *present* ability on these topics/skills." For part 2b, it is "Please rank your ability on these topics/skills *at the time you started taking physics at IUPUI*." In both cases, the emphasis is as it appears to students. The topics/skills, as they appear in the survey are identified in Table I. The first four items relate to commonly used specific computational languages/packages, while the remaining seven relate to specific computational skills/techniques. For brevity, we will use the question labels in the first column throughout the remainder of this paper.

In our analysis of the results, we exclude incomplete responses. If a student completes the survey more than once in a semester, we retain only the last submission. These requirements resulted in dropping 79 surveys due to incomplete responses and 22 surveys due to repeats. After applying these rules, we have $N = 516$ surveys in our data set. To characterize students' progress through the curriculum, we bin responses based on the highest level course (100 level, 200 level, etc.) taken. These levels correspond roughly to 1^{st} year students, 2^{nd} year students, etc. Students do occasionally take courses out of order, this is a source of noise in our results. In the present data set, we have $N_{100} = 247, N_{200} = 210, N_{300} = 36$, and $N_{400} = 23$. This is approximately a 50% increase over the data set reported in [9].

B. Analysis

We are interested in characterizing the change in students' computational self-efficacy. To this end, we define:

$$G_{X00} = \bar{P}_{X00} - \bar{I}_{X00}, \tag{1}$$

TABLE I. The survey items analyzed and identifying labels.

Label	Survey Item
Q1	Excel (or other spreadsheets)
Q2	LaTeX
Q3	MATLAB
Q4	Other Programming Language
Q5	Rescaling a problem to eliminate units
Q6	Testing code to determine its overall correctness
Q7	Visualization of functions and other results
Q8	Matrix operations
Q9	Solving differential equations
Q10	Numerical integration and differentiation
Q11	Data analysis

where $\bar{P}_{X00}$ and $\bar{I}_{X00}$ are respectively the means of the present (part 2a) and initial (part 2b) responses for each academic level bin, as described above. These values can be interpreted as the growth students at the X00 level feel they have achieved between joining the program and the date of their survey.

Typically, measuring students gains as they progress through the program would result from a comparison of the mean *present* competencies for each survey item across academic level. While this method is common, and has some advantages, it ignores the possibility that students re-evaluate the scale, specifically their *initial* ability, as they gain more knowledge. In particular, beginning students may not realize how complex and powerful computational methods may be, since they have not yet been exposed to them at the university. This could give them a false sense that "they know this." This is related to the well-known Dunning-Kruger effect [12], in which people tend to overestimate their competency when they are unaware of their own competency level.

We now define two alternative measures of student growth over the full course of the program: G_1 is defined as:

$$G_1 = \bar{P}_{400} - \bar{P}_{100}, \tag{2}$$

the difference between *present* competencies as expressed by 400-level students and 100-level students. This is a common mechanism to measure change using survey data, as it relies only on participants reporting present data, and can be calculated based on a single administration of the survey. However, it is highly sensitive to students' overestimates of their abilities during that single administration.

The second method, G_2, includes the possibility that students will re-evaluate their own initial starting point (i.e., they revisit their own definition of the scale) as they gain knowledge. We define

$$G_2 = (\bar{P}_{400} - \bar{I}_{400}) - (\bar{P}_{100} - \bar{I}_{100}) = G_{400} - G_{100} \tag{3}$$

where $\bar{I}_{400}$ and $\bar{I}_{100}$ are the means of the *initial* competencies identified by the 400-level and 100-level students, respectively. This reduces the impact of students' changing estimates of their initial competencies. Regrouping the terms in 3 provides a clearer picture of the effect of introducing the initial terms:

$$G_2 = G_1 + I_{eff}. \tag{4}$$

Here, we have introduced $I_{eff} = (\bar{I}_{100} - \bar{I}_{400})$, the effect of students' re-evaluation of their initial skills. If students did not re-evaluate their own abilities, we would find $\bar{I}_{eff} \approx 0$. We will show in Section III that this assumption is unwarranted. In fact, we find that $\bar{I}_{eff}$ is usually significant and positive. That is, students completing 1^{st} year courses rate their initial skills considerably higher than 4^{th} year students.

Note that this definition of I_{eff} allows a direct comparison with G_1 over the same time delay. The surveys are given at the *end* of each semester, so G_1 measures the skill difference

between a 400-level and 100-level student at that time. The prompts for the initial state define "initial" as when the student *first began* the undergraduate program, implying the beginning of a semester. Our definition of G_2 in terms of I_{eff} subtracts the growth made by students in the first semester, equalizing the time lag with G_1.

III. RESULTS

A. Present and initial self-efficacy

Figure 1 shows the mean initial and present self-efficacy in each skill for each undergraduate level. The clusters in panel (a) show I_{X00} for each item. Panel (b) is similar, for the P_{X00} values. We have omitted Item Q4 from this figure and subsequent analysis because the "other programming language" identified by students varied widely, and because it is not directly related to evaluating the impact of our curriculum.

We conducted a one-way omnibus ANOVA for each of the twenty groups [13]. We find statistically significant ($p \leq 0.05$) differences for fourteen of the groups as follows:
Initial: Q3 ($p = 0.019$), Q5 ($p < 0.001$), Q9 ($p = 0.044$), Q10 ($p < 0.001$), and Q11 ($p < 0.001$)
Present: Q2, Q3, Q6, Q7, Q8, and Q9 ($p < 0.001$) Q5 ($p < 0.01$), Q10 ($p = 0.050$), and Q11 ($p = 0.030$)
This analysis identifies those groups in which there is a statistically significant difference among the $X00$ levels for a given survey item.

B. Growth over the full program

We now turn our attention to the growth measures for the full program defined in subsection II B. Figure 2 plots the values of G_{400}, the growth in self-efficacy defined in Eqn. (1) for each computational skill. This represents how much students feel they have grown in computational ability from joining the program to completing one or more 400 level courses. Additionally, a pairwise t-test was conducted comparing the initial and present mean competencies in each skill (I_{400} vs. P_{400}). These results are presented in columns 2 and 3 of Table II.

Figure 3 shows a comparison between methods G_1 and G_2 of measuring growth. We choose not to compare these values directly with those in Fig. 2 due to the difference in time lag noted previously. Again, we have conducted independent t-tests for each method, and the results are summarized in Table II, columns 4-7. Because of the difference in sample size between groups ($N_{400} = 23$ and $N_{100} = 247$), Levene's test for homogeneity of variance was conducted for each method and corrections were made to the p-values accordingly.

IV. DISCUSSION

The results presented above suggest that substantial changes occur in students' computational self-efficacy as they progress through our undergraduate program. Figure 1 (a) shows their self-reported initial competencies in skills for each item, clustered by level. This supports our concern that only measuring students' *present* self-efficacy could miss valuable information. This is further shown by Fig. 3, where a consistent difference is shown between growth measures. It

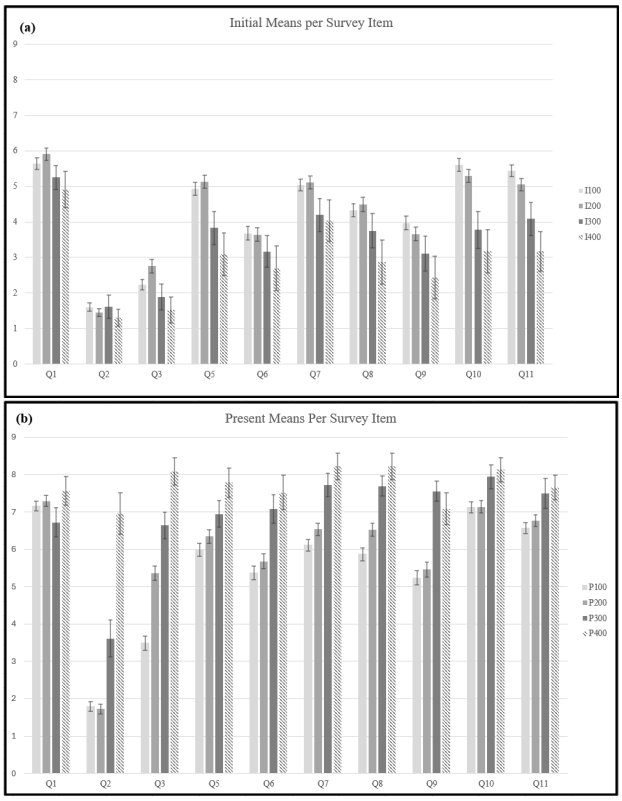

FIG. 1. The initial (a) and present (b) mean responses for each survey item. The light grey bar represents the 100-level students, the next lightest represents 200-level students, the dark grey represents 300-level students, and the striped bar represents the 400-level students. Error bars are the standard errors.

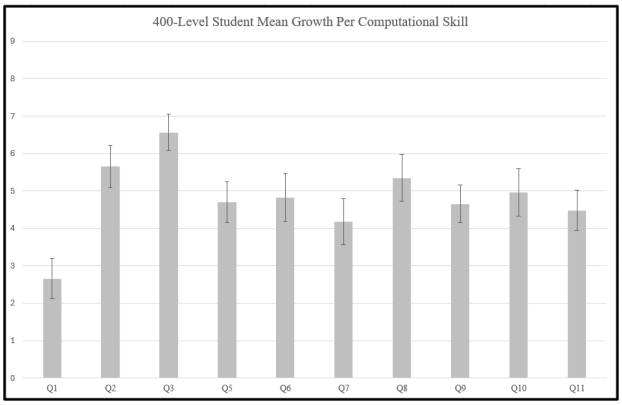

FIG. 2. The growth in self-efficacy for 400-level students (G_{400}) across each computational skill. Error bars are the standard errors.

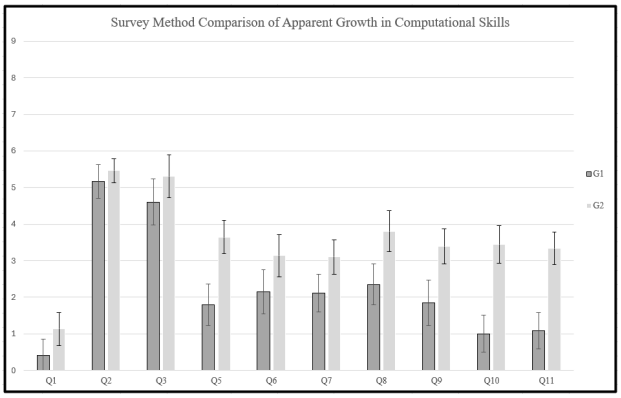

FIG. 3. Comparison of G_1 and G_2 for 100- and 400- level students. The darker bars represent the mean difference between the present skills (G_1). The lighter bars represent the mean difference between the growth measures (G_2). Error bars are the standard errors.

should be noted that I_{eff}, as defined in Equation 4, can be visualized either as the difference between the right-most and left-most bars in each cluster of Fig. 1 (a), or as the difference between the two bars in each cluster of Fig. 3.

Figures 2 and 3 illustrate the changes visually. Figure 2 shows substantial growth (G_{400}, as defined in 1) substantial growth in every category, though the change in $Q1$ (Excel) is noticeably smaller than the others. We believe this is due to Excel being the only item surveyed with which students have significant experience in high school. Figure 3 shows that $G_2 > G_1$ for every surveyed item. Again, the difference is substantial for many items.

Table II shows effect sizes as Hedges g values, similar to Cohen's d, with $g = 0.5$ generally characterized as a moderate effect while $g \geq 0.8$ characterizing a large effect [14].

We also note some interesting trends when comparing the results in panels (a) and (b) of Fig. 1. For example, survey items Q5 through Q8 all show significant decreases in initial means while showing significant increases in present means as students go through the program. Whereas survey items Q10 and Q11 show less difference in present means while the initial means change more dramatically than the first group. This could indicate that students' experience larger gains in understanding the scope of those topics as they progress through the program. For example, lower-level students are less likely to fully understand what it means to use computational methods to perform numerical integration/differentiation, so an interpretation of the prompt could be viewed simply as taking a derivative/integral.

TABLE II. The results of t-test comparisons, including p-values (p) and Hedge's effect sizes (g) where $p < 0.05$. The first comparison (P_{400} vs I_{400}) was conducted using a pairwise t-test with the remaining two conducted via independent samples. The † symbol indicates p-values corrected for failed Levene's tests for homogeneity of variance.

Survey	P_{400} vs I_{400}		P_{400} vs P_{100}		G_{400} vs G_{100}	
Item	p	g	p	g	p	g
Q1	< 0.001	0.990	0.355	–	0.018	0.516
Q2	< 0.001	2.003	$< 0.001^{\dagger}$	2.433	$< 0.001^{\dagger}$	3.618
Q3	< 0.001	2.741	$< 0.001^{\dagger}$	1.521	< 0.001	2.059
Q5	< 0.001	1.731	$< 0.001^{\dagger}$	0.676	< 0.001	1.695
Q6	< 0.001	1.518	$< 0.001^{\dagger}$	0.755	< 0.001	1.207
Q7	< 0.001	1.367	$< 0.001^{\dagger}$	0.888	$< 0.001^{\dagger}$	1.451
Q8	< 0.001	1.733	$< 0.001^{\dagger}$	0.873	< 0.001	1.513
Q9	< 0.001	1.853	$< 0.001^{\dagger}$	0.634	< 0.001	1.683
Q10	< 0.001	1.575	0.044	0.441	$< 0.001^{\dagger}$	1.459
Q11	< 0.001	1.680	$0.005^{\dagger}$	0.465	$< 0.001^{\dagger}$	1.693

V. CONCLUSIONS

We acknowledge several sources of uncertainty, e.g., self-selection bias in our sample and the need to further validate our instrument. However, we argue that **RQ1** can tentatively be answered in the affirmative. Our results show that as students progress through the program their estimates of their "initial ability" decrease consistently over time and over a range of different computational skills. With regard to **RQ2**, we believe that our definition of G_2 does a reasonable job of accounting for this decrease. The trends in G_2 match those observed in G_1, which is agnostic to initial values, but accounts for the changes in skills such as numerical integration/differentiation, data analysis, and rescaling problems, which beginning students are more likely to misunderstand. As shown in Table II, the effect sizes for statistically significant results are larger and, for Q1, the growth is statistically significant only for G_2.

As we acquire more data, particularly at the 300- and 400-level, we anticipate the ability to refine our analysis to the individual course level. Understanding what courses lead to the most significant gains can allow the department to consider applying similar methods to other courses with the intention of further increasing gains in self-efficacy. We also hope to be able to pair data from the survey with direct measures of students' abilities. This would require substantial changes to our human subjects protocol, as it would require at least some knowledge of which students participated in the survey to be available to physics faculty.

ACKNOWLEDGMENTS

The authors acknowledge financial support from IU Indianapolis' STEM Education and Innovation Research Institute (SEIRI) and from National Science Foundation grant DUE-2021209. We also wish to thank the Partnership for Integration of Computation into Undergraduate Physics (PICUP).

[1] T. J. Atherton, Resource Letter CP-3: Computational physics, American Journal of Physics **91**, 7 (2023).
[2] R. H. Landau, Resource Letter CP-2: Computational Physics, American Journal of Physics **76**, 296 (2008).
[3] AAPT recommendations for computational physics in the undergraduate physics curriculum, https://www.aapt.org/Resources/upload/AAPT_UCTF_CompPhysReport_final_B.pdf (2016), retrieved 5/16/21.
[4] P. Heron and L. McNeil, *PHYS21: Preparing Physics Students for 21st Century Careers* (American Physical Society, College Park, MD, 2016).
[5] Partnership for the Integration of Computation into Undergraduate Physics, https://www.compadre.org/picup/, retrieved 5/25/21.
[6] D. H. McIntyre, J. Tate, and C. A. Manogue, Integrating computational activities into the upper-level Paradigms in Physics curriculum at Oregon State University, Am. J. Phys. **76**, 340 (2008).
[7] D. M. Cook, Computation in the Lawrence physics curriculum, https://www2.lawrence.edu/dept/physics/ccli/compatlu.pdf (2006), retrieved 5/16/21.
[8] PhysPort, https://www.physport.org/, retrieved 6/1/21.
[9] A. Gavrin, G. Vemuri, and D. Maric, Preliminary efforts to evaluate an initiative introducing computation across the undergraduate physics curriculum, 2021 PERC Proceedings (2021).
[10] G. Gallagher, G. Vemuri, A. Gavrin, and D. Maric, An exploratory factor analysis of a survey measuring undergraduate student attitudes of computational methods in physics, 2024 PERC Proceedings (submitted) (2024).
[11] Carnegie Classification of Institutions of Higher Education, https://carnegieclassifications.acenet.edu/institution/indiana-university-purdue-university-indianapolis//, retrieved 4/29/24.
[12] J. Kruger and D. Dunning, Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments, Journal of Personality and Social Psychology (1999).
[13] D. C. Howell, *Statistical Methods for Psychology* (Cengage Learning, Independence, KY, 2007) pp. 316–318, 370–372.
[14] C. O. Fritz, P. E. Morris, and J. J. Richler, Effect size estimates: Current use, calculations, and interpretation, J. Exp. Psychol. Gen. **141**, 2 (2012).