Preliminary efforts to evaluate an initiative introducing computation across the undergraduate physics curriculum

A. Gavrin and Gautam Vemuri
Department of Physics, Indiana University Purdue University Indianapolis,
402 N. Blackford St. LD154, Indianapolis, IN, 46202

Danka Maric
STEM Education and Innovation Research Institute, Indiana University Purdue University Indianapolis,
755 W. Michigan St. UL1123, Indianapolis, IN, 46202

We report our preliminary efforts to evaluate a departmental project: the inclusion of computational methods across our undergraduate curriculum. Our overarching goal is for students to consider computational approaches as a “normal” way to solve physics problems, on par with analytical approaches. In this paper, we focus on our efforts to evaluate the development of our students’ attitudes and self-efficacy with respect to key computational methods. We describe our efforts to develop and deploy a survey instrument students complete each semester. This allows us to study, e.g., the points in the curriculum at which students gain confidence with particular methods, or adopt more expert-like attitudes regarding computation in general. We investigated the reliability of our instrument using a split-half process and found the Spearman-Brown coefficients for unequal length were $r = 0.818$, $r = 0.895$, and $r = 0.917$ for the three constructs in our survey. We also provide preliminary data from the early use of the survey and outline next steps for the project.

2021 PERC Proceedings edited by Bennett, Frank, and Vieyra; Peer-reviewed, doi.org/10.1119/perc.2021.pr.Gavrin

I. INTRODUCTION

Over the last 20 years, much has been written about the need to incorporate computational methods in the undergraduate physics curriculum [1–7]. More recently, the importance of inculcating computational methods has been recognized at the national level. In 2016, the American Association of Physics Teachers (AAPT) released *Recommendations for Computational Physics in the Undergraduate Physics Curriculum* [8] and the APS-AAPT Joint Task Force on Undergraduate Physics Programs released its final report, emphasizing the importance of computation as a career skill for physics majors [9].

Despite this attention, only a few physics departments have embraced computational methods at the level we envision; Oregon State University [10] and Lawrence University [11] are notable examples. We were also surprised by the apparent lack of published instruments suited to evaluating the inclusion of computation in the curriculum. The PICUP project site [12] does not include any such tools, nor does PhysPort [13]. Indeed, Caballero notes that there is an opportunity for physics education researchers to support computational instruction through the development of computational assessments [7]. It is this lack of assessment instruments we are beginning to address. Our plan is to develop two instruments. The first is focused on students' attitudes and self-efficacy regarding computational methods. That instrument is the subject of this paper. The second will focus on students' abilities to use computational methods. Those interested in the first instrument may contact the authors for access to the current version.

To help readers understand the scope and constraints on our efforts, we briefly describe our institutional context and the broad outlines of our computational project in Section II. The remainder of the paper will focus on our preliminary efforts to evaluate our results. Section III will outline the methods we used to develop the first of two planned instruments, including efforts to establish validity and reliability. In sections IV and V, we will present and discuss preliminary data gained from this instrument, and Section VI will provide a summary and outline our plans for the future.

II. CONTEXT

A. Institutional context

This work was completed at Indiana University Purdue University Indianapolis (IUPUI), an urban, public university located in downtown Indianapolis, IN. The department is of moderate size: we have 10 tenured or tenure-track faculty members, and three full-time lecturers. We offer B.S., M.S., and Ph.D. degrees.

Our undergraduate curriculum follows a traditional model, including a two-course introductory sequence, a "modern physics" course, two upper-level labs, and single semester treatments of intermediate mechanics, electrodynamics, physical optics, quantum mechanics, and statistical physics. We also require a single semester of faculty-mentored undergraduate research as a capstone.

B. The computational initiative

Beginning in 2016, the faculty began discussing efforts to incorporate computational methods in the curriculum. From the outset, **our overarching goal has been that our graduates will consider computational approaches to be a "normal" way to do physics**. They should not consider the use of computation to be an unusual technique set aside for "special" problems, e.g. many-body physics. The initiative arose as a "grassroots" effort, but was supported from the outset by the department chair (one of the co-authors of this paper) and the college administration.

We gained initial support in the form of an internal grant from a campus center, IUPUI's STEM Education and Innovation Research Institute (SEIRI), supplemented by departmental funds. SEIRI also provided the support of a postdoc with evaluation experience to help us begin the effort reported here. The grant was supplemented by department funds used to support faculty travel to workshops hosted by the Partnership for the Integration of Computation into Undergraduate Physics (PICUP) [12], and to invite colloquium speakers who had experience teaching computational physics in a variety of contexts.

The first year of the project was devoted to expanding our overarching goal into specific student learning outcomes (SLOs), and to further establishing priorities, methods, and responsibilities. We discussed issues such as whether a specialized computational physics course (or sequence) should be developed, what skills and attitudes students should gain, what, if any, computational platform should be preferred, and how to assess the results. Some key conclusions were

1. We would incorporate computation in *all* courses.
2. Our primary focus must stay on physics, not coding.
3. Five SLOs describing skills and attitudes, e.g., students should not be satisfied with working code, but should use that code to "explore the physics."
4. A list of topics with which students should gain some fluency, e.g., numerical integration, data analysis, and using common tools such as Excel and MATLAB
5. That we would need to develop at least two instruments to evaluate the progress of the initiative.

C. Evaluation goals

The balance of this paper focuses on the first of two instruments conceived in item 5, above. **Our approach to evaluating our "normalizing" goal is to understand the path students take in gaining confidence and skill with computational approaches.** Discussions among the faculty led to a plan to use a repeated survey technique that allows us to understand that path, both at the individual student level and in aggregate.

The instrument was developed to address three primary constructs:

- Affect regarding the value of computational methods
- Self-efficacy regarding 10 computational methods
- Students' estimates of their *initial ability* on these same 10 methods

The survey was given to all students in physics majors' courses at the end of the semester from Fall 2018 through Spring 2021. The instructions specify that students should complete the survey after each semester during which they take one or more physics courses. Using this instrument, we hope to be able to answer research questions such as "Do our graduating students have expert-like attitudes regarding the use of computational methods?"; "At what points in the curriculum do their attitudes shift from naive towards expert?"; and similar questions focused on self-efficacy regarding particular skills.

III. METHODS

A. Instrument development

The development, review, and refinement of the present instrument took place in three stages. First, initial items were developed by project leaders and further discussed and adjusted by the full group of faculty members in physics. We worked until a consensus was reached that the instrument could be used as an effective evaluation tool. The instrument begins with demographic questions (names, student IDs, physics courses taken that semester). These are followed by items asking students to rate their agreement with statements related to computational physics, e.g., "Using computational methods helps me understand physics topics" (five-point Likert scale). Students are then asked to rate both their present and initial abilities on ten computational skills, e.g., numerical integration, on a 1-10 scale (initial is defined as "at the time they began the program"). During the second stage of development, the instrument was reviewed by an evaluation expert with instrument development experience at SEIRI, which resulted in a few items being reworded for better clarity. This form of the instrument was used for 4 semesters of data collection during the period of internal funding.

B. Validity and reliability

We sought preliminary evidence of content validity after receiving further funding from NSF. Content validity evidence involves examining the relationship between the instrument content (e.g., themes, wording, item format, tasks, or questions) and the construct it is intending to measure through evaluations from expert judges, among other methods [14]. We asked five content area experts from PICUP to provide feedback on the overall structure of the instrument, as well its clarity and completeness. There was agreement among the experts that the items are examining what they are intended to examine—students' attitudes and self-efficacy concerning computational methods. Some clarifications and additional items were also suggested. For instance, several of the content experts agreed that the term "analytical methods" might be confusing, particularly to beginning students. The phrase "pencil and paper math" was added as a parenthetical explanation. One item was dropped and four new items were added, bringing the total number of items in this section from six to nine. Finally, the order of the items asking students to rate their skills was changed such that present skills were rated before initial skills. This updated version was used in the two most recent semesters of data collection.

We did not have data from the same participants on multiple occasions and thus were unable to examine test-retest reliability. However, we were able to estimate internal consistency by obtaining split-half correlation coefficients. This method examines the agreement between different parts of a measure by splitting scores into two halves and examining the corrected correlation between the two halves, which serves as an estimate of the reliability of the full-length measure [14]. Split-half correlations were respectively examined for the affective, present skills, and initial skills item clusters. We used an odd-even split, in which odd-numbered items are in one subset and even-numbered items are in the other. This type of split avoids any factors related to item order (e.g., participant fatigue) from having an extraneous effect on the coefficient by ensuring items from each portion of the measure are represented in each subset [15].

The Spearman-Brown coefficients for unequal length were $r = 0.814$ for the affect questions, $r = 0.895$ for the present skills items and $r = 0.917$ for initial skills items in the updated instrument. Although there is debate on this topic, high Spearman-Brown coefficients are thought to reflect better reliability, with some experts citing reliability coefficients of 0.7 to 0.8 or above being acceptable for research purposes [16]. Based on these guidelines, the current instrument seems to have reasonable internal consistency.

We should note that the value $r = 0.814$ above was obtained for the most current version of the survey, used for two semesters, with $N = 130$ respondents. The self-efficacy construct was unchanged between versions, and had $N = 323$ respondents. For completeness, we also performed a split-half analysis of the affect questions on the earlier version of the survey, and found results that were consistent $r = 0.866$.

C. Data analysis

The first section of the survey measures students' attitudes towards computational methods. Our approach is to begin with a one-way ANOVA followed by Tukey's HSD (honestly significant difference) test [17]. In each case, we compare all records from students who are completing a 100 level course, 200 level course, etc. If a student takes courses at multiple levels in a given semester, we consider the highest level course taken. In some cases, our results violated the assumption of equal variances between groups. One-way ANOVAs are typically quite robust [17] but since we have unequal group sizes, this violation can be problematic. For the questions that violated this assumption, we applied a Welch correction [18]. For the follow-up test to the Welch corrected ANOVA, we used the Games-Howell test [19], which is designed for assumption violations but functions similarly to Tukey's HSD in that it produces results for all pairwise combinations of treatments (courses levels). Where we report results that are statistically significant ($p < 0.05$), we also report effect sizes using Hedge's g, a measure similar to Cohen's d, but suited to cases with different sample sizes [20].

D. Initial use

Each semester, about 2 weeks before final exams, a link to the survey is sent to all students who are completing the targeted courses. The instructions tell students that they will be asked to complete the survey each semester, but that only one copy is necessary if they are taking more than one physics course. Participation is voluntary, and no incentives were offered for participation.

This design allows us to measure the changes in students' attitudes over time, both in the aggregate and as individuals. Our response rates are reasonable, typically near 20%. We note that selection effects may bias results, and that the sample size in upper level classes is low due to the size of our major. As a result, we cannot yet observe statistically significant results tracking individuals or single courses. Our present data set allows us to find significant results when we aggregate responses over students completing courses at the 100 level, 200 level, etc. (This roughly tracks students' 1st year classes, 2nd year, etc). We report those results, based on 6 semesters: Fall 2018 - 2020, and Spring 2019 - 2021. The results are described in the next section.

IV. RESULTS

A. Attitudes

For this work, we analyzed the five affect questions that were included in both the original and updated surveys. All five produce statistically significant results in the ANOVA, and most produce multiple significant results in the Tukey HSD. As an example, we highlight this item "Computational methods, experiments, and analytical solutions are equally necessary in the field of physics," phrased as 5-point Likert scale. For convenience, we will refer to this item as "item A1" (Affect 1).

TABLE I. Significance and effect size for pairwise comparisons among course levels for item A1, organized by course level.

Course levels compared	p	g
100, 300	0.015	0.62
100, 400	0.040	0.74
200, 300	0.014	0.77
200, 400	0.036	0.91

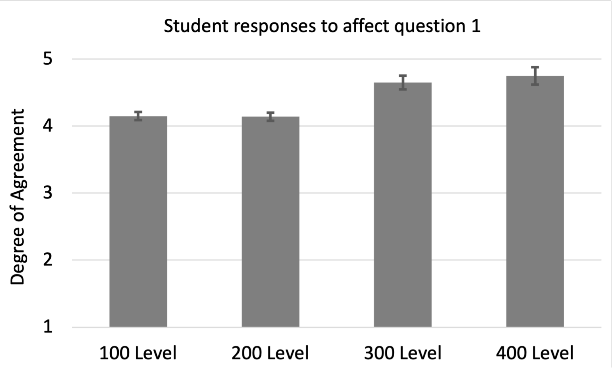

FIG. 1. Students' average responses to item A1 grouped by highest completed course level. Error bars are standard errors.

The omnibus one-way ANOVA showed a significant difference between class levels, $F(3, 340) = 4.57, p = 0.001$. The Tukey HSD follow-up showed significant pairwise differences in four of the 6 possible pairwise comparisons. The results are detailed in Table I. The data is also illustrated in Fig. 1. The numbers of respondents are $N_{100} = 183, N_{200} = 132, N_{300} = 23, N_{400} = 12$.

B. Competencies

The second portion of the survey asks students to rate their present ability on a scale of 1 to 10 for ten computational skills. Eight of the ten produce statistically significant results in the ANOVA, and most produce multiple significant results in the TukeyHSD. We highlight two of these competencies here: use of MATLAB, and matrix operations. We respectively designate these items "SE1" and "SE2" (self-efficacy 1 and 2). For item SE1, the omnibus one-way ANOVA results were Welch's $F(3, 40.750) = 22.658, p < 0.01$. For item SE2, we find $F(3, 41.465) = 11.098, p < 0.01$. As above, the results for all statistically significant pairwise comparisons are summarized in Table II. The data is shown in Fig. 2. The numbers of respondents to these questions were $N_{100} = 166, N_{200} = 128, N_{300} = 21, N_{400} = 12$.

TABLE II. Significance and effect size for pairwise comparisons among course levels for items SE1 and SE2

MATLAB		
Course levels compared	p	g
100, 200	< 0.01	0.59
100, 300	< 0.01	1.04
100, 400	< 0.01	1.35
200, 400	< 0.01	0.85

Matrix Operations		
Course levels compared	p	g
100, 300	< 0.01	0.82
100, 400	0.029	0.70
200, 300	< 0.01	0.69

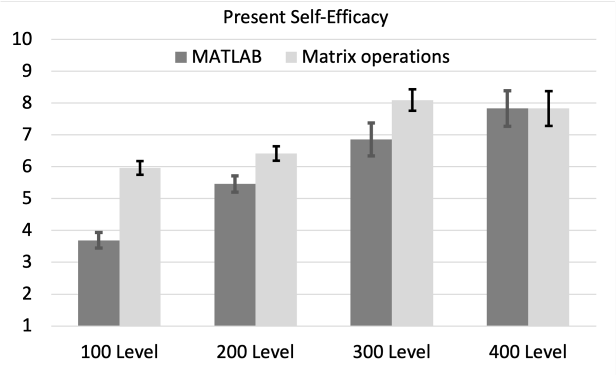

FIG. 2. Students' responses to items SE1 and SE2 grouped by highest completed course level. Error bars are standard errors.

V. DISCUSSION

To date, our results suggest that our instrument is valid and reliable at a basic level. After several stages of review and adjustment, faculty within and beyond the department agree that the items are clear and focused on the desired constructs indicating content validity. The split-half correlation coefficients we obtained for each construct were over 0.8, indicating reasonable internal consistency. We will continue our efforts to establish validity and reliability in coming semesters.

The data presented above enables us to begin to address our stated goal of understanding the path along which our students develop the desired computational skills and attitudes. Interpreting Hedge's g values is similar to Cohen's d, with most guidance characterizing $g = 0.5$ as a medium effect, and $g \geq 0.8$ as a large effect [20]. By this standard, all three examples described here, plus many others in our data, show students making medium or large gains in adopting expert-like attitudes and in increasing self-efficacy.

One notable observation is that some skills and attitudes appear to develop in a stepwise fashion between the 200 and 300 level. Results for item A1 presented in Table I and Fig. 1 develop this way. Likewise, the second table and figure show that students' self-efficacy with respect to matrix operations (item SE2) also takes a substantial jump at this level. In contrast, students' self-efficacy with respect to using MATLAB (item SE1) develops more steadily, as shown in Table II and Fig. 2. We see these trends in other questions not presented here as well.

As we gather more data, we expect that this trend will sharpen, and we will be able to investigate which courses at each level are most responsible for these improvements. This observation highlights one of the chief benefits this instrument offers. It allows us to determine where in the curriculum students' gains are occuring, and to compare those gains to the goals of the courses taken. If some courses seem to underperform, corrective action can be considered. Similarly, if certain courses produce large gains, it may be possible to adopt the methods used in those classes to improve others.

VI. CONCLUSION

Our efforts are still at an early stage, but our results thus far point towards some preliminary conclusions. We have developed an instrument intended to measure students' attitudes and self-efficacy towards computation, and we have established content validity by involving experts in several rounds of review of the instrument. Regarding reliability, we have thus far established internal consistency by measuring split-half correlation coefficients. Over several semesters use of the instrument, we find that our students' attitudes become more expert-like as they progress through the curriculum, and that their confidence with using computational methods also grows, with effect sizes that are in many cases substantial. We note that some of the measures we focus on grow steadily, while others seem to take a particularly large step between the 200 and 300 level. We speculate that this is a result of the significant step up in sophistication in our classes and decrease in class size at that level.

As we continue to acquire data, we will soon be able to look more closely at these developmental processes, identifying the particular courses in which students make progress on specific survey items. Additional data will also allow us to more fully establish validity and reliability. Factor analysis will allow us to examine the instrument structure, and correlating results from this instrument with data on student performance in courses will help establish predictive validity.

ACKNOWLEDGMENTS

The authors acknowledge financial support from IUPUI's 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] H. Gould and J. Tobochnik, Integrating computation into the physics curriculum, in *Computational Science — ICCS 2001*, edited by V. N. Alexandrov, et al. (Springer-Verlag, Berlin/Heidelberg, 2001).

[2] R. Landau, Variation of instructor-student interactions in an introductory interactive physics course, Comput. in Sci. and Eng. **8**, 22 (2006).

[3] J. R. Taylor and B. A. King, Using computational methods to reinvigorate an undergraduate physics curriculum, Comput. in Sci. and Eng. **8**, 38 (2006).

[4] R. S. Chabay and B. A. Sherwood, Computational physics in the introductory calculus-based course, Am. J. Phys. **76**, 307 (2008).

[5] N. Chonacky and D. Winch, Integrating computation into the undergraduate curriculum: A vision and guidelines for future developments, Am. J. Phys. **76**, 327 (2008).

[6] R. M. Serbanescu, P. J. Kushner, and S. Stanley, Putting computation on a par with experiments and theory in the undergraduate physics curriculum, Am. J. Phys. **79**, 919 (2011).

[7] M. D. Caballero and L. Merner, Prevalence and nature of computational instruction in undergraduate physics programs across the United States, Phys. Rev. Phys. Educ. Res. **14**, 020129 (2018).

[8] 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.

[9] P. Heron and L. McNeil, *PHYS21: Preparing Physics Students for 21st Century Careers* (American Physical Society, College Park, MD, 2016).

[10] 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).

[11] D. M. Cook, Computation in the Lawrence physics curriculum, https://www2.lawrence.edu/dept/physics/ccli/compatlu.pdf (2006), retrieved 5/16/21.

[12] Partnership for the Integration of Computation into Undergraduate Physics, https://www.compadre.org/picup/, retrieved 5/25/21.

[13] PhysPort, https://www.physport.org/, retrieved 6/1/21.

[14] *Standards for Educational and Psychological Testing* (American Educational Research Association, 2014).

[15] R. F. DeVellis, *Scale Development: Theory and Applications* (Sage Publications, Newbury Park, CA, 2017).

[16] R. M. Furr, Split-half reliability, in *Encyclopedia of Research Design*, edited by Salkind, Neil J. (Sage Publications, Newbury Park, CA, 2012).

[17] D. C. Howell, *Statistical Methods for Psychology* (Cengage Learning, Independence, KY, 2007) pp. 316–318, 370–372.

[18] B. L. Welch, On the comparison of several mean values: An alternative approach, Biometrika **38**, 330 (1951).

[19] P. A. Games and J. F. Howell, Pairwise multiple comparison procedures with unequal n's and/or variances: A monte carlo study, J. of Educ. Stat. **1**, 113 (1976).

[20] 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).