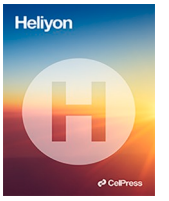
Heliyon 9 (2023) e18997

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Effective teaching in computational thinking: A bias-free alternative to the exclusive use of students' evaluations of teaching (SETs)

Noemi V. Mendoza Diaz [a,*], Trinidad Sotomayor [b]

[a] College of Engineering and School of Education, Texas A&M University, College Station, TX, USA
[b] College of Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile

ARTICLE INFO

Keywords:
Measures of teaching effectiveness
Computational thinking
Students' evaluations of teaching (SETs)
Measures of learning
Biased faculty assessment

ABSTRACT

The tenure system in the United States places significant importance on teaching effectiveness. To date, students' evaluations of teaching (SETs) have been the reigning mechanism for assessing effective teaching. However, prior work has shown that SETs are often biased against underrepresented groups and minorities. The present study analyzes options for effective teaching assessments, which include evaluating final grades and measuring the differences between students' pre- and post-tests (normalized gain) using standard instruments. The content area and the instrument used in this study originated in the computational thinking field, which has a widespread presence in engineering, where minorities are at a disadvantage. This study obtained a total of 88 student participants from four sections of an introductory engineering course at a Southwestern institution. The study utilized a computational thinking diagnostic (CTD) to inform the course teaching approach (the intervention). Results show that (a) normalized learning gains correlated moderately with SETs, (b) final grades correlated strongly with SETs, (c) final grades correlated strongly with normalized learning gains, (d) the educational intervention based on the CTD significantly affected student learning, and (e) SET comments affect evaluations. The implications include the notion that standardized instrument-driven instruction and evaluations can increase the success of minorities on both sides of the classroom. The purpose of this manuscript is to invite the Heliyon readership to get involved in the development of related instruments and to incorporate these measures of learning into their instruction so biases are avoided or minimized.

1. Introduction

Teaching is an important component of academic life, and most individuals who choose to become professors are aware that developing proficiency in teaching is an important aspect of success in academia. It is expected that individuals who choose to become professors enjoy teaching and aspire to make a positive impact on their students. The tenure system in the United States aims to reward good teaching, but the reigning mechanism for evaluating teachers' effectiveness focuses on students' perceptions, or what is called, students' evaluation of teaching (SET). Prior work has shown that SETs tend to produce results that are particularly biased against women and minorities in the professoriate.

* Corresponding author.
E-mail address: nmendoza@tamu.edu (N.V. Mendoza Diaz).

https://doi.org/10.1016/j.heliyon.2023.e18997
Received 3 February 2023; Received in revised form 29 July 2023; Accepted 4 August 2023
Available online 10 August 2023

On the other hand, the development of computational thinking skills is an important part of success in science and engineering. However, underrepresented groups in science and engineering have traditionally been marginalized in computational thinking and other STEM opportunities due to multiple factors that include a lack of mentors and guides from their own underrepresented groups. This issue of marginalization is often exacerbated when people from underrepresented communities become STEM faculty and receive biased evaluations of their teaching [1–5]. Given the significance of this issue, STEM journals have been recently publishing research that addresses such bias in teaching evaluations, including recent studies in Heliyon [6,7].

Motivation: A team of faculty from underrepresented minority backgrounds, challenged by their continuous lower evaluations in comparison with their white counterparts at a predominantly white institution, started conversations about SETs and dived into the literature in the topic. This study is the result of their conversations and constitutes their first attempt at a more effective evaluation of their teaching.

This article aims to achieve three objectives: (1) To compare different approaches to teaching evaluations, (2) to replicate and propose the development and adoption of standardized instruments such as the computational thinking diagnostic (CTD) for student learning in engineering and computing, and ultimately, (3) to inform and empower the engineering and computing community, specifically women and underrepresented minorities in academia, with information about different approaches for assessing effective teaching beyond students' evaluations. Universities could use the mechanisms and results in this study instead of SETs or in conjunction with SETs to measure student learning and evaluate teacher effectiveness in an objective manner that is more supportive of minorities. These alternative measurements are particularly applicable for measuring computational thinking skills, which are universal to STEM disciplines. The results discussed in this paper are also transferable to other STEM content areas.

2. Background and literature review

For decades, researchers have analyzed multiple approaches for assessing teaching effectiveness. There have been eight meta-analyses and major reviews published on the topic [8–15]. Some recent studies have compared SETs with so-called learning measures and have tried to establish correlations to test the validity of SETs. A comparison between early and recent meta-analytic approaches shows tension between two perspectives; the one proposing SETs and the other opposing SETs as measures of teaching effectiveness.

In one of the most recent meta-analysis, Uttl et al. [15] criticized studies proposing SETs, specifically Cohen's multi-cited dissertation study [9], Felman's meta-analysis [13], and Clayson's meta-analysis [8]. Uttl et al. described these studies as non-replicable, unverifiable, and flawed and strongly emphasized that previous findings from cognitive sciences show that it is unrealistic to measure professors' teaching effectiveness simply by asking students to answer questions about their perceptions of their courses, the instructors' knowledge, and how much they learned. Such measurements are imprecise given students' individual differences including their prior knowledge, their motivations, and their interest in the subject [15]. Uttl et al. noted, "Two key findings emerged, (1) the findings reported in previous meta-analyses are an artifact of poor meta-analytic methods, and (2) students do not learn more from professors with higher SETs" [15, p. 40]. Alarmingly, in a later study, Uttl et al. reported conflicts of interest associated with large positive correlations between SET and learning in those earlier meta-analyses [16]. These conflicts of interest included corporate, administrative, evaluation unit, SET author, and funder interests.

2.1. Measures of learning

Teaching and learning are intuitively related, but they are not the same. For many universities and higher-level institutions in the United States, the SET is the standard metric for teaching effectiveness. However, as mentioned, several studies have criticized this measurement, citing gender and ethnic biases, grade inflation, neglect, or lack of interest as validity issues. Some more recent studies have shown that even if SETs are unbiased, they only capture a portion of teaching effectiveness (e.g., student satisfaction), so it is critical to tap into other assessment forms, such as those related to learning [8,15,17–21].

2.1.1. Grades as a measure of learning

There are multiple fields and publications that discuss grades as a measure of learning and explore students' grades in relation to SETs, but given their additional perspectives on typical issues found with the use of SETs, only three are particularly relevant to this study. Clayson's [8] meta-analysis analyzed studies that focused on the relationship between grades and SETs. He indicated that **leniency and reciprocity** were two prevalent findings from these studies: "Students give lenient-grading instructors higher evaluations," and "Students who receive better grades give better evaluations, irrespective of any leniency tendency of the instructor" [8, p.18].

Johnson wrote a book on grade inflation and summarized correlational studies in groups according to whether data were recorded at the individual student level or aggregated at the classroom level [18]. He stated, "Data aggregated by class are usually considered best for assessing the effects of instructor leniency on teaching evaluations, since such data can be used to compare average class grades and average course evaluations" [18, p. 51]. This is of relevance to this study given the aggregated nature of the SETs used for analysis.

Stehle et al. conducted a correlational study and looked into SETs collected after a surgical medical training. The SETs evaluated three factors: overall instructor quality, overall course quality, and students' subjective learning. Researchers compared results from SETs against students' final test scores and practical examinations. They found that the SET scales correlated with each other and with the practical examination but did not correlate with the final test. There was also no significant correlation between the two examination types. The significance of Stehle et al. (2012) [21] work relies upon the difficulty of capturing effective teaching metrics

utilizing different assessment instruments [21].

2.1.2. Pre-tests and post-tests as measures of learning

Some studies have included pre-tests and post-tests in an effort to gauge teaching effectiveness. Lee et al., Marks et al., and Stark-Wroblewski et al. all found that the differences between pre- and post-results came the closest to indicating "learning" and perhaps the closest to revealing "effective teaching." [20,22,23].

Marks et al. analyzed data from 1106 students enrolled in a remedial class in the academic years of 2007, 2008, and 2009 [23]. The students took a pre-test placement exam designed to assess freshmen at state universities, and then they took a post-test at the end of the course. They called this process "an objective measure of student learning" and hypothesized that student course evaluations only partially reflect the learning that takes place in class. The course evaluation had 14 total questions that students rated on a Likert-based scale. The evaluations contained eight items related to the instructor and six items related to the course. The data analyses included curve fitting-regression analyses that considered student evaluations, post-test Z scores, student demographics (i.e., age, gender, ethnicity), pre-test scores, grades in subsequent classes, final grades in the class, GPA, SAT, or ACT test scores, and other information. Marks et al. found a mild relationship between course evaluation and student learning, specifically between student learning and the course evaluation questions that focused on the class value as a whole [23].

In an attempt to map teaching effectiveness to SET scores, Stark-Wroblewski et al. examined learning outcomes through a 30-item test that measured student learning in a general psychology class. After the semester ended, Stark-Wroblewski et al. compared pre- and post-test scores to assess the learning of 165 students in the course. In addition, they investigated student course grades and tested four hypotheses: (1) Expected small to moderate positive correlation between SET scores and grades, (2) Expected a small to medium positive correlation between learning (the difference between pre- and post-tests) and grades, (3) Expected a small to medium positive correlation between learning and SET scores, (4) Expected the relationship between learning and an item on the SET about learning to be moderate to large. The SET included 16 items; the 16th item asked students to reflect on the statement, "I learned in this course …" This question served as the basis for verifying the fourth hypothesis. Stark-Wroblewski et al.'s study found strong correlations between learning scores and grades, weak correlations between student grades and SET scores, and weak correlations between learning and SET scores. They stated that these results indicated that the SETs "may capture important aspects of teaching effectiveness, such as satisfaction with the course, but instructors should consider supplementing SETs with other measures of teaching effectiveness" [20, p. 411].

In physics education, Lee et al. compared SETs to conceptual learning gains [22]. Lee et al. defined conceptual gains through normalized gains (1), which is the proportion of improvement on an instrument from pre-to post-instruction compared to the maximum possible improvement. The following equation portrays this gain, where the brackets denote the class average values:

$$g = \frac{< post\% > - < pre\% >}{100\% - < pre\% >} \tag{1}$$

Lee et al. looked at pre- and post-tests for standardized conceptual inventories in introductory mechanics and introductory electricity and magnetism classes, and they also evaluated students' final course grades. Through the American Association of Physics Teachers, the American Physical Society, and the American Astronomical Society conferences, they recruited 24 new physics faculty members responsible for teaching 37 classes. These professors provided the researchers with their SETs, final grades, and pre- and post-test results. Lee et al. correlated final grades with SETs and normalized gains with SETs and found no correlation between students' ratings of instruction quality and conceptual learning gains. This study is of major relevance to the analysis of this manuscript since learning gains were used.

2.1.3. Standardized testing as measures of learning

In this discussion about the validity of SETs for measuring teaching effectiveness compared with other possible options, arguments around standardized testing are quite compelling. Standardized tests (e.g., SAT, ACT, GRE) are considered the norm by which diverse students are admitted to many higher education programs. In addition to enrollment purposes, some disciplines have adopted standardized testing to assess learning. Hake argued that since most disciplines have failed to develop definitive tests to assess learning, many have opted to use SETs to measure the effectiveness of educational methods because it is an easier approach [17]. He added that, to his knowledge, only Physics and Astronomy have developed such tests for introductory courses. We can consider Lee et al.'s study above as one of these cases. Henderson et al. explicitly stated, "In general, instructors are much more positive about the methods they use to evaluate their teaching than the methods their institutions use to evaluate their teaching. Both instructors and institutions could benefit from broadening the assessment sources they use to evaluate teaching effectiveness through increased use of standardized measures based on student learning and greater reliance on systematic formative assessment" [24, p. 1].

The abundance of literature points the study to the following objectives; (1) to compare different teaching assessments as well as to test the intervention using statistical analysis for the specific case of computational thinking in engineering and computing, (2) to inform and empower the engineering and computing community, especially women and minorities, in alternative ways of assessing teaching effectiveness, and (3)to replicate and propose development and adoption of instruments in engineering and computing. Thus, armed with information about multiple ways to assess teaching effectiveness and measures of learning, this study compares average SET results at a Southwestern institution (SW-SETs), to pre- and post-testing results (normalized learning gains) and to final grades. This study uses a computational thinking diagnostic (CTD), developed by the author, as the standardized instrument in a pre- and post-testing setting.

3. Research questions and design

This study has taken advantage of the expertise in the development of standard tests by other fields mentioned above (e.g., physics) to measure effective teaching and explores those approaches in computational thinking. In this context, the following research questions are investigated.

1. How do different approaches that measure effective teaching compare in an introductory programming course?
a) What is the relationship (correlation) between the Normalized Gains and SW-SETs in the groups?
b) What is the relationship (correlation) between the Final Grades and SW-SETs in the groups?
c) What is the relationship (correlation) between the Final Grades and the Normalized Gains in the groups?
2. What is the overall effect of the computational thinking intervention course on students' learning gains utilizing the CTD instrument?
3. What are the patterns shown in the comments written in the SETs and their potential effect on the evaluation of teaching?

In order to answer these research questions, a sequential mixed-method approach will be utilized [25].The first two questions will be answered through quantitative techniques, involving pre- and post-test time-series design (no control group, just within-group), while the last question will be answered qualitatively. The following sections will seek to answer these questions.

4. Study setting

In the early 2000s, a Southwestern institution launched a technology undergraduate program. Over the past two decades, the curriculum for this program has changed on numerous occasions. The latest change involved tenured and tenure-track professors creating the program's mission with a strong emphasis on computational thinking. The current curriculum includes first-year general university courses, and the following years are a combination of management, human resource development, and technology classes.

The introductory course is designed to provide foundational knowledge about computational thinking and managing information systems. The computational thinking portion occupies half to two-thirds of class time and assignments during the semester. A computational thinking diagnostic is applied at the beginning and end of the course as a mechanism to assess student's knowledge levels upon entry, emphasize aspects that need more attention to orient the class, and assess students' knowledge levels upon exiting to explore learning gains. When developing this diagnostic, it was not contemplated its use as a teaching effectiveness tool. The next section will describe this computational thinking diagnostic in more detail as well as the participants and the methodology used.

5. Method

5.1. Participants

This study consisted of student participants enrolled in four different class sections of an undergraduate introductory technology course. The sessions from Spring 2019 are termed Spring19-Group1 and Spring19-Group2, and the Fall 2019 sections are termed Fall19-Group1 and Fall19-Group2. There were 110 total students enrolled in these four class sections, but this study includes data from only 88 students. One instructor taught three of the course sections and a separate instructor taught the fourth section, both instructors were foreign-born, one female Latina and the other male Asian. The university system that records students' SW-SETs teaching evaluations dissociates responses from student identifiers. The SW-SETs system provides aggregated results in the form of mean and standard deviations to instructors in eight different items and an overall mean (see appendix). It also includes comments made by students.

5.2. Materials

5.2.1. The standard test: computational thinking diagnostic

Pre-college literature and curriculum clearly define and target computational thinking. Initiatives such as robotics clubs, code.org, or advanced placement (AP®)-computer science courses are a testament to the significant stakeholder interest in developing student technological abilities as early as possible. At these levels, it is easy to differentiate computational thinking skills from specific language/program coding skills. For example, frameworks such as the College Boards' AP® or the International Society for Technology in Education (ISTE) often refer to computational skills such as abstraction, data representation, decomposition, and computing impact with the widely acknowledged understanding that these skills can be learned using flowcharts and pseudo-code [26].

But what is happening at the university level? At the Southwestern institution, approximately 45% of entry-level engineering students have been exposed to computational thinking. As in all educational aspects, this digital divide is more prevalent for underrepresented groups, including women, ethnic minorities, and low socioeconomic status (SES) students. Motivated by this disparity, a Computation Thinking Diagnostic (CTD) was developed with emphasis on those aspects independent of coding or programming language. The CTD is a 14-item test with an additional question related to prior coding/programming experience through an AP course, competitions, or self-training. Following ISTE's model, this test evaluates the following student skills: 1) decomposition and solution; 2) pattern matching; 3) abstraction; and 4) automation [27]. The one CTD item shown in Fig. 1 illustrates the measurement of student decomposition and solution skills. The Rubik's cube represents the three dimensions that every programmer should consider

when, for example, plotting 3D graphs. It can also represent other orthogonal 3R dimensions. Decomposition skills can be measured by properly calculating the number of iterations when traversing the necessary nested loops associated with finding the coordinates or points in the cube.

For the validity of the instrument, its latest version has been psychometrically validated for one factor: the computational thinking factor [28].

5.3. Data collection

This study obtained data from students' computational thinking diagnostic scores (pre- and post-scores), students' evaluations of teaching through SW-SETs, and students' final class grades. Students who did not answer both the pre- and post-CTD tests or who dropped the course were not included in the study sample. Therefore, although there were 110 students enrolled in the course, this study includes data from only 88 students.

Per University regulations, all students were invited to complete evaluations of the instructor's teaching at the end of the semester in an anonymized format. Students evaluated their instructor by responding to eight questions using the following Likert scale rating: Strongly Agree – 5, Agree – 4, Undecided – 3, Disagree – 2, Strongly Disagree – 1. Only 72 students completed SW-SETs and the overall averages were used in this study. (See Appendix).

At the end of the semester, instructors used the university platform to report students' final letter grades (A, B, C, and D). To analyze overall student learning in this research, the following weights were assigned to each letter grade: A – 4, B – 3, C – 2, and D – 1. These numeric values are standard values used in GPA calculations. Table 1 provides the normalized learning gains and the averages for the metrics utilized in this study. To answer the research questions, the statistical methods are correlations and the Wilcoxon Signed Rank Test.

5.4. Declaration

Ethics statement: The study utilized historical post-fact de-identifiable data from students who had taken the course, so there was no need to obtain informed consent from them. The study was approved by the Human Research Protection Program-Institutional Review Board in the Southwestern Institution under approval #IRB2023-0123.

6. Results

6.1. Quantitative analysis

The response to the first research question, (a) which sought to find the relationship (correlation) between the Normalized Gains and SW-SETs in the groups, resulted in the calculations shown in Fig. 2. The Pearson's coefficient was 0.39; for the behavioral sciences, 0.3–0.50 is considered a "moderate correlation" [29,30].

For the 1-b research question, which sought to find the relationship (correlation) between the Final Grades and SW-SETs in the

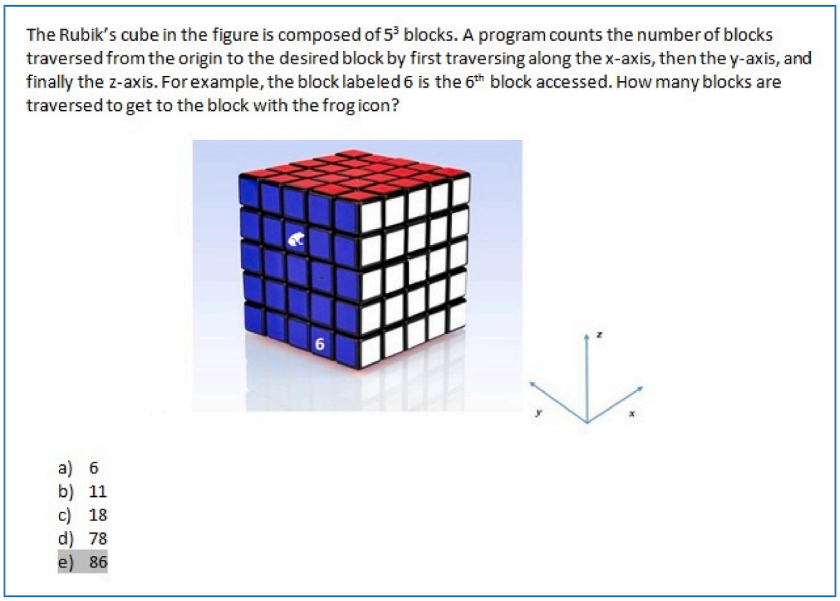


Fig. 1. Decomposition computational thinking item.

Table 1
Measures of learning and SW-SETs averages per group.

	Pre-test	Post-test	Normalized Gain	Final Grade	SW-SETs
Spring19-Group1 (n = 13)	5.08	7.69	0.29	3.31	3.81
Spring19-Group2 (n = 21)	4.62	6.05	0.15	3.05	3.46
Fall19-Group1 (n = 29)	5.55	7.31	0.21	3.29	3.98
Fall19-Group2 (n = 25)	6.84	8.4	0.22	3.12	3.29

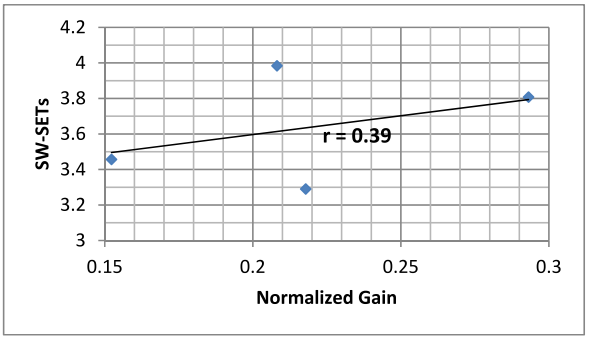


Fig. 2. Moderate correlation between Normalized Gain and SW-SETs.

groups, the correlation calculation resulted in 0.86, which is considered a strong correlation for the behavioral sciences (>0.5). Fig. 3 shows these results.

Investigating the 1-c research question, the relationship (correlation) between the Final Grades and the Normalized Gains in the groups resulted in a Pearson's coefficient of 0.77, which indicates a strong correlation (see Fig. 4).

Treatment intervention: Finding the effect of the course on students' learning gains and answering the second research question involved checking the data for normality using the Shapiro-Wilk test [31]. The Shapiro-Wilk test works best with smaller sample sizes and is more powerful than other similar statistical tests [32]. The Shapiro-Wilk test showed that the data was not normally distributed ($p = .015$), so a Wilcoxon Signed Rank Test was conducted instead of a *t*-test [33].

The Wilcoxon Signed Rank Test is a counterpart of the paired sample *t*-test in nonparametric hypothesis testing; it revealed that the post scores were **significantly higher** than pre scores ($Z = -5.526$, $p < .01$). The Cohen's D effect size on the pre-post change indicated a medium effect of 0.59 [34]. Table 2 shows these results.

6.2. Qualitative analysis

The third research question, related to students' comments and their potential effect on the evaluation of teaching, was answered through a qualitative approach. As the appendix shows, the SW-SETs also had space for comments for each of the eight items to be evaluated. It was not mandatory to provide comments for each of the items, yet a letter score had to be provided. The number of letter grades for each of the items is shown in Table 3. Some of the examples of item comments are as follows:

Item Comments Examples.

[A] "Dr. X always was ready to present every day and always prepared a lesson plan" (Spring 2019, Group 1, Instructor A).

[C] "Since this is the first time this course was being taught in this form, I think it was difficult to know early what the exact objectives would be" (Spring 2019, Group 2, Instructor A).

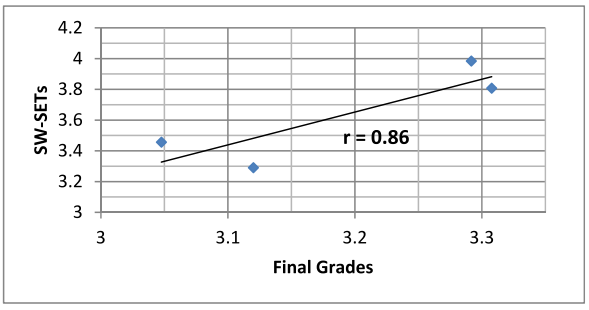


Fig. 3. Strong correlation between SW-SETs and final grades.

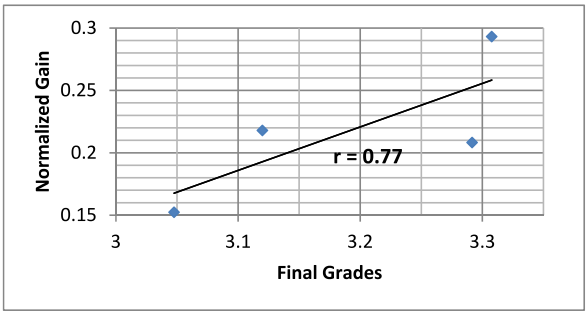


Fig. 4. Strong correlation between normalized gains and final grades.

Table 2
Wilcoxon signed-rank test result.

	N	M	SD	Z	p	Cohen's D
Pre CTD	88	5.62	2.45	−5.53	**<.01**	0.59
Post CTD	88	7.38	2.44			

Table 3
Summary of letter grades for SETs items and qualitative analysis of general comments.

	A	B	C	D	E	General Comments		
						Negative	Neutral	Positive
Spring 2019								
Group 1-Instructor A	9	8	4	7	3	4	1	0
Group 2-Instructor A	1	10	6	7	4	5	0	1
Fall 2019								
Group 1-Instructor A	13	5	2	4	0	5	3	3
Group 1-Instructor B	0	2	4	10	11	5	2	3
TOTALS	**23**	**25**	**16**	**28**	**18**	**19**	**6**	**7**

[D] "His lack of communication tells me he doesn't care" (Fall 2019, Group 2, Instructor B)

At the end of the SW-SETs there was also a textbox for general comments. The general comments were analyzed via content analysis in three categories: negative, neutral and positive. Table 3 shows the analysis of these general comments. Two coders were involved in this analysis reaching an inter-rater reliability coefficient of 0.87 which is considered a very strong agreement [35,36]. Students were also not obligated to provide general comments at the end of the survey. Some examples of these comments are as follows.

6.2.1. General comments examples

Coded Positive: "I think Dr. X is a great Professor. She really cares for her students and wants to be sure that they are understanding what is taught. She holds office hours and is always available after class to talk about anything. She gets a bit sidetracked sometimes but overall she is a great professor" (Fall 2019, Group1, Instructor A).

Coded Neutral: "great instructor, she is very knowledgeable and willing to help. The course is hard to follow though." (Spring 2019, Group1, Instructor A).

Coded Negative: "Course had no direction and was a mess of overlapping ideas that provided no information. The class needs considerably more structure and a clear direction for the instructor to follow. The instructor appeared to care but if she really cared this class would not have been this useless or have been this irrelevant." (Fall 2019, Group1, Instructor A).

The item letter score provides the quantitative basis for which the SET score is calculated. It also sets the tone of the comment, in the event the student chooses to provide a comment. Letters A and B, leaned towards more positive scores, and they account for 48 entries. Letters D and E, more towards negative scores, account for 46. This shows a balance between negative and positive entries.

The general comments however leaned more toward negative evaluations. This conveys a sense that only the most dissatisfied students are vocal about their dissatisfaction than those more on the positive end of the spectrum. This answers the last research question related to the effect of the comments in the evaluations since it shows that SETs without comments tend to be more moderate than those with comments.

7. Discussion and conclusion

The first part of this manuscript introduced readers to critical perspectives of past and current research related to SETs. It also provided a summary of journal articles' findings about the effectiveness of SETs while presenting a review of three alternative measures of teaching effectiveness – all related to measures of learning: (1) final grades, (2) pre-tests and post-tests (learning gains), and (3) standardized tests. These alternative measures of teaching effectiveness are relevant, because the data collected across two semesters of an introduction to programming course afforded the opportunity to analyze the validity of SETs for the specific case of computational thinking skills at a Southwestern institution.

Results of the study show a moderate correlation between normalized gains (pre- and post-tests) and SW-SETs. While SW-SETs provides an overall mean of students' perceptions and the CTD provides the score in a test, both relate to a certain degree but not entirely. This reinforces the notion that normalized gains and SETs evaluate different constructs.

The results also revealed a strong correlation between final grades and SW-SETs, which can be interpreted as an instance of reciprocity in the participant groups. The phenomenon where students who receive better grades give better evaluations, independent of the instructor, is well documented in the literature as mentioned above.

Finally, similar to Stark-Wroblewski et al.'s findings, the strong relationship between learning gains and final grades in this study can be interpreted as a confirmation that these measures of learning are consistent and may more accurately assess teaching effectiveness [20].

In regard to understanding how the intervention affects the process of learning computational thinking skills and its relationship to effective teaching, results show that the intervention utilized and assessed by the computational thinking diagnostic created a significant effect on students' learning. It is encouraging to find that this intervention had a positive effect on all students' learning, regardless of the instructor. In contrast, the qualitative findings were discouraging because it became evident that comments were usually done by dissatisfied students who set the tone for more negative implications rather than positive. If not for the different mechanisms to assess teaching effectiveness reported here, these SETs could've painted a biased picture towards the instructors of these sections.

The findings open the discussion about what constitutes effective teaching, which might also be related to the course or intervention design, implementation, and assessment, specifically when using standardized methods. One of this article's objectives was to replicate and propose the development of standardized instruments for learning in the STEM fields, which is an effort that evidently the physics community is already engaged in. By developing assessment instruments that guide instruction in a more collegial way, numerous communities, with their shared wisdom and perspectives, can be represented more inclusively. The exercise of developing standard instruments can also serve a mentorship purpose, either formal or informal, since experienced mentors can coach novice instructors in the art and craft of item generation and test development. Lee et al.'s recent study in the physics field served as the basis for this study now focused on computational thinking. Since teaching programming practices are part of the foundational courses in engineering, a more informed and standardized approach to teaching coding or programming could be useful in many aspects, including for assessing teaching effectiveness [22].

As part of the aims of this study and as the results show, it is expected that these results inform and empower women and minorities in engineering and computing by providing evidence that student satisfaction and perceptions are used as a proxy to evaluate teaching. The qualitative findings corroborate the notion that the most dissatisfied students can be more vocal and influence the way faculty teaching is perceived. This constitutes a problem in a society as deeply polarized and where biases have been extensively reported.

This article has implications for minorities in faculty positions at predominately white institutions who often find themselves vulnerable to evaluations solely based on students' perceptions and associated biases. Since this research highlighted debates in current literature related to student evaluations of teaching, the results present a more critical view of what institutions utilize to evaluate fundamental teaching aspects, especially in the tenure system where promotion and tenure is awarded based on evidence of effective teaching.

This paper aimed to offer a better understanding of the different approaches for assessing effective teaching beyond student evaluations, and the results reinforce the notion that for minorities in academia, there are alternative bias-free approaches for evaluating teaching skills. Recognizing the relevance of students' perceptions and satisfaction, these more bias-free alternatives could be used in combination yet providing the appropriate dimension of what teaching effectiveness implies.

Finally, knowing the motivations behind this study and based on the results and discussion, we can see how minority faculty are affected by biased SETs. This study aspires to continue identifying the most effective measures of learning and teaching computational thinking and invites the audience to consider the alternatives presented in this article, including a call to action in the development of standard assessment tools to measure teaching effectiveness.

7.1. Limitations

At this institution, there is no opportunity to associate SETs with individual students since responses are anonymous and voluntary. Since SW-SETs are limited by this aggregate anonymous data, other students' demographic factors such as gender or ethnicity could not be evaluated.

The small sample size is another limitation of this study. Additional course sections, where introduction to programming is taught, were not available for analysis. As expected, given their higher stakes, SETs are highly guarded by instructors. Since the sample size was small, there is a restriction on transferability of findings and a restriction on covariate analysis. Also, participants in this study may not have completed the pre- or post-test but could have participated in the SET at the end of the semester, which also constitutes a

limitation. It is expected that further studies can utilize larger samples including more detailed participant information. Despite these limitations, the study provides a snapshot of the SETs and other forms of evaluating teaching effectiveness across disciplines, time, and approaches. This study learns from other fields to propose alternative forms of measuring teaching with a more attuned focus on the desired outcome: enhancing students' learning.

Declarations

Author contribution statement

Noemi V. Mendoza Diaz: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Trinidad Sotomayor: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Data availability statement

The data that has been used is confidential.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

APPENDIX. PERSONALIZED INSTRUCTOR/COURSE APPRAISAL SYSTEM (SW-SETs) ITEMS

A = Strongly Agree (5)B = Agree (4)C = Undecided (3)D = Disagree (2)E = Strongly Disagree (1).

ITEMS	A	B	C	D	E	Comments
The instructor clearly communicated the objectives of the course.						
The instructor was well prepared.						
The instructor's workload expectations were realistic.						
The instructor demonstrated interest in students' progress.						
The instructor created a course climate conducive to respecting diverse viewpoints.						
The instructor clearly communicated that he or she was available for academic consultation outside of class.						
Overall, the instructor was a good teacher.						
On the whole, this was a good course.						
Essay Questions on the Appraisal						
Please add your overall thoughts on the course and the instructor						

References

[1] W. Aspray, Women and Underrepresented Minorities in Computing, first ed., Springer International Publishing, Cham, 2016 https://doi.org/10.1007/978-3-319-24811-0.

[2] J. Van Sickle, K.R. Schuler, J.P. Holcomb, S.D. Carver, A. Resnick, C. Quinn, D.K. Jackson, S.F. Duffy, N. Sridhar, Closing the achievement gap for underrepresented minority students in STEM: a deep look at a comprehensive interbention, J. STEM Educ. Innovations Res. 21 (2) (2020). https://www.jstem.org/jstem/index.php/JSTEM/article/view/2452/2160.

[3] A.L. Graves, E. Hoshino-Browne, K.P. Lui, Swimming against the tide: gender bias in the physics classroom, J. Women Minorities Sci. Eng. 23 (1) (2017), https://doi.org/10.48550/arXiv.1705.09636.

[4] L. Price, I. Svensson, J. Borell, J.T.E. Richardson, The Role of gender in students' ratings of teaching quality in Computer Science and Environmental Engineering, Inst. Electrical Electronics Eng. Transactions Educ. 60 (4) (2017) 281–287, https://doi.org/10.1109/TE.2017.2696904.

[5] D. Gupta, P. Garg, Kumar, Analysis of students' ratings of teaching quality to understand the role of gender and socio- economic diversity in higher education, Inst. Electrical Electronics Eng. Transactions Educ. 61 (4) (2018) 9–327, https://doi.org/10.1109/TE.2018.2814599.

[6] K. Arrona-Palacios, C. Okoye, N. Camacho-Zuñiga, E. Hammout, S. Luttmann-Nakamura, J. Hosseini, D. Escamilla, Does professors' gender impact how students evaluate their teaching and the recommendations for the best professor? Heliyon 6 (10) (2020) https://doi.org/10.1016/j.heliyon.2020.e05313.

[7] F. Quansah, Item and rater variabilities in students' evaluation of teaching in a university in Ghana: application of Many-Facet Rasch Model, Heliyon 8 (12) (2022), https://doi.org/10.1016/j.heliyon.2022.e12548.

[8] D.E. Clayson, Student evaluations of teaching: are they related to what Students Learn?: a meta-anaylsis and review of the literature, J. Mark. Educ. 31 (1) (2009) 16–30, https://doi.org/10.1177/0273475308324086.

[9] P.A. Cohen, Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies, Rev. Educ. Res. 51 (3) (1981) 281–309, 1.2307/1170209.

[10] P.A. Cohen, Validity of students ratings in psychology courses: a research synthesis, teach, Psychology 9 (2) (1982) 78–82, 10.1207/s15328023top0902_3.

[11] P.A. Cohen, Comment on A Selective review of the validity of student ratings of teaching, J. Eng. Educ. 54 (4) (1983) 448–458, https://doi.org/10.2307/1981907.

[12] D.A. Dowell, J.A. Neal, A selective review of the validity of student ratings of teachings, J. High. Educ. 53 (1) (1982) 51, https://doi.org/10.2307/1981538.

[13] K.A. Feldman, The association between student ratings of specific instructional dimensions and student achievement: refining and extending the synthesis of data from multisection validity studies, Res. High. Educ. 30 (6) (1989) 583–645, https://doi.org/10.1007/BF00992392.
[14] L.W. McCallum, A meta-analysis of course evaluation data and its use in the tenture decision, Res. High. Educ. 21 (2) (1984) 150–158, https://doi.org/10.1007/BF00975102.
[15] B. Uttl, C.A. White, D.W. Gonzalez, Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related, Stud. Educ. Eval. 54 (2017) 22–42, https://doi.org/10.1016/j.stueduc.2016.08.007.
[16] B. Uttl, K. Cnudde, C.A. White, Conflict of interest explains the size of student evaluation of teaching and learning correlations in multisection studies: a meta-analysis, PeerJ 7 (2019), e7225, https://doi.org/10.7717/peerj.7225.
[17] R.R. Hake, Problems with student evaluations: is assessment the remedy? Indiana University. https://web.physics.indiana.edu/hake/AssessTheRem1.pdf.
[18] V.E. Johnson, Grade Inflation: A Crisis in College Education, Springer, Cham, 2003, https://doi.org/10.1086/421860.
[19] P. Spooren, B. Brockx, D. Mortelmans, On the validity of student evaluation of teaching: the state of the art, Rev. Educ. Res. 83 (4) (2013) 598–642, https://doi.org/10.3102/0034654313496870.
[20] K. Stark-Wroblewski, R.F. Ahlering, F.M. Brill, Toward a more comprehensive approach to evaluating teaching effectiveness: supplementing student evaluations of teaching with pre-post learning measures, Assess Eval. High Educ. 32 (4) (2007) 403–415, https://doi.org/10.1080/02602930600898536.
[21] S. Stehle, B. Spinath, M. Kadmon, Measuring teaching effectiveness: correspondence between students' evaluations of teaching and different measures of student learning, Res. High. Educ. 53 (8) (2012) 888–904, https://doi.org/10.1007/s11162-012-9260-9.
[22] L.J. Lee, M.E. Connolly, M.H. Dancy, C.R. Henderson, W.M. Christensen, A comparison of student evaluations of instruction vs. Students' conceptual learning gains, American J. Phys. 86 (7) (2018) 531–535, https://doi.org/10.1119/1.5039330.
[23] M. Marks, D. Fairris, T. Beleche, Do Course Evaluation Reflect Student Learning? Evidence from a Pre-test/post-test Setting, Riverside, Department of Economics, University of California, Riverside, 2010, https://doi.org/10.1016/j.econedurev.2012.05.001.
[24] C. Henderson, C. Turpen, M. Dancy, T. Chapman, Assessment of teaching effectiveness: lack of alignment between instructors, institutions, and research recommendations, Phys. Rev. Special Top. - Phys. Educ. Res. 10 (1) (2014), 010106, https://doi.org/10.1103/PhysRevSTPER.10.010106.
[25] J.W. Creswell, V.L. Plano Clark, Designing and Conducting Mixed Methods Research, first ed., Sage Publications, Newbury Park, 2017 https://doi.org/10.1111/j.1753-6405.2007.00096.x.
[26] A.P. Central, Computer Science Principles Exam, 2020. https://apcentral.collegeboard.org/courses/ap-computer-science-principles/exam?course=ap-computer-science-principles. (Accessed 24 March 2022). on.
[27] J. Krauss, K. Prottsman, Computational Thinking and Coding for Every Student. The Teacher's Getting-Started Guide, Corwin Press, Thousand Oaks, 2016.
[28] N.V. Mendoza Diaz, D.A. Trytten, R. Meier, S.Y. Yoon, An Engineering Computational Thinking Diagnostic: A Psychometric Analysis, 2021 IEEE Frontiers in Education Conference (FIE), Lincoln, NE, USA, 2021, pp. 1–5, https://doi.org/10.1109/FIE49875.2021.9637142.
[29] J. Cohen, Statistical Power Analysis for the Behaviorial Sciences, second ed., L. Erlbaum Associates, Mahwah, 1988 https://doi.org/10.4324/9780203771587.
[30] J.D. Evans, Straightforward Statistics for the Behavorial Sciences, Thomson Brooks/Cole Publishing Co., Pacific Grove, 1996, https://doi.org/10.2307/2291607.
[31] S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (Complete Samples), Biometrika 52 (3/4) (1965) 591–611, https://doi.org/10.2307/2333709.
[32] N.M. Razali, Y.B. Wah, Power comparisons of shapiro-wilk, Kolmogorov-smirnov, lilliefors and anderson-darling tests, J. Stat. Model. Anal. 2 (1) (2011) 21–33. https://www.researchgate.net/publication/267205556_Power_Comparisons_of_Shapiro-Wilk_Kolmogorov-Smirnov_Lilliefors_and_Anderson-Darling_Tests.
[33] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1 (6) (1945) 80–83, https://doi.org/10.2307/3001968.
[34] L. Cohen, Measurement of Life Events in Life Events and Psychological Functioning: Theorical and Methodological Issues, first ed., SAGE Publications Ltd., Newbury Park, 1988 https://doi.org/10.1016/0022-3999(90)90016-W.
[35] O.R. Holsti, Content Analysis for the Social Sciences and Humanities, Addison-Wesley, Boston, 1969, https://doi.org/10.1177/003803857000400231.
[36] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem. Med. 22 (3) (2012) 276–282, https://doi.org/10.11613/BM.2012.031.