Response-shift bias in student self-efficacy during an actively taught physics course

Kelly Miller,[1] Tobias Espinosa,[2] Ives Araujo,[3] and Isaura Gallegos[4]
[1]*Department of Physics and Division of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*Mathematics Statistics and Physics Institute, Universidade Federal do Rio Grande, Santo Antônio da Patrulha-RS, 3005 Cel. Francisco Borges de Lima St., 95500-000, Brazil*
[3]*Physics Institute, Universidade Federal do Rio Grande do Sul, 9500 Bento Gonçalves Avenue, Porto Alegre-RS, 91501-970, Brazil*
[4]*Graduate School of Education, Harvard University, Cambridge, Massachusetts 02138, USA*

(Received 5 August 2022; accepted 1 December 2023; published 19 December 2023)

Self-efficacy is an important measure in science education as it is predictive of persistence and success in science, technology, engineering, and mathematics (STEM) courses and is an influential factor in students' decisions to major in STEM fields. It is unclear what effect active teaching strategies have on students' self-efficacy, which is typically measured with a pretest at the beginning of the semester and a post-test at the end of the semester. To better understand what happens to self-efficacy over the course of an actively taught physics class, in addition to the typical pretest and post-test, we used a reflective pretest. At the end of the semester, we asked students to reflect on their abilities at the beginning of the semester and we compared this "reflective" self-efficacy to both their presemester and postsemester self-efficacy. We found that students' reflective self-efficacy was systematically lower than their self-efficacy at the beginning of the semester. Interviews reveal that discrepancies between presemester self-efficacy and reflective self-efficacy are the result of response-shift bias. Because of students' limited experience with active learning environments, response-shift bias makes it difficult to accurately measure students' change in self-efficacy over the semester of an actively taught physics course. We conclude that reflective pretests in combination with interviews can help educators and researchers understand if changes in self-efficacy are being masked by response-shift bias.

DOI: 10.1103/PhysRevPhysEducRes.19.020167

I. INTRODUCTION

It has been projected that, over the next decade, there will be a massive shortage of workers with training in science, technology, engineering, and mathematics (STEM), particularly in the United States [1]. To meet the growing demand for STEM workers, more students need to be interested in pursuing degrees in STEM-related fields.

A. Self-efficacy

Self-efficacy in science has been shown to be an influential factor in students' decisions to major and persist in STEM fields [2–5]. Self-efficacy, which was introduced by Bandura in the late 1970s, refers to one's belief in their ability to complete a specific task (or a set of tasks) in a given dimension [6]. According to Bandura, an individual develops their self-efficacy for a specific task through social and personal experiences which fall into four categories: mastery experiences, vicarious learning experiences, social persuasion experiences, and an individual's physiological and affective state [7,8]. Mastery experiences occur when students have recurring episodes of success or failure by actively engaging. Vicarious experiences occur when students observe others (specifically peers or role models) performing a task. Social persuasion occurs when students receive feedback (both verbal and nonverbal) from others (peers and teachers). Finally, a student's physiological and affective state refers to their mood as it emerges during the performance of a task and is influenced by whether the student experiences stress and anxiety. Students draw on all four of these categories as sources of information in building their self-efficacy [8–11].

In addition to persistence in STEM, science self-efficacy has been shown to be a strong predictor for performance in science courses, resilience, and career choices in STEM [3,12–19]. Self-efficacy has also been shown to be related to student behavior relevant to learning in an active environment, such as perseverance and self-regulation [20].

Self-efficacy has been shown to decrease during traditionally taught (lecture) physics courses [21,22]; however,

in actively taught physics courses, research has shown it to either decrease [21] or stay the same [22]. In our previous work, using the same survey to measure physics self-efficacy (Appendix A) as in this study, *we* have shown that the self-efficacy of female students increases over the semester of an actively taught class while the self-efficacy of male students does not change [23]. In the literature, it is well established that pretest and post-test measurements are threatened by response-shift bias, especially with self-reported measures like self-efficacy [24,25]. Response-shift bias refers to a type of measurement error that occurs when an individual's perception of a construct changes over time, causing them to report a different score on a given outcome measure even though their actual status may not have changed. This can happen, for example, when a person changes their attitudes, beliefs, or expectations about a certain aspect of their life, leading them to view the same experiences differently after the change. The literature has also shown that asking students to reflect back and evaluate themselves at an earlier point in time can help to minimize this bias [26]. In this study, to better understand how students' self-efficacy changes during a semester of an actively taught physics course, we administer a reflective pretest, in addition to the traditional pretest and post-test. The aim is to propose a possible explanation for the lack of consistent change in self-efficacy between the pretest and post-test in an active learning environment.

B. Reflective pretests and response-shift bias

A reflective pretest is administered at the end of an experience and asks respondents to reflect and rate themselves at the beginning of the experience. Reflective pretests are commonly used to minimize the response-shift bias that occurs when a metric is altered between measurements [26]. Response-shift bias is well cited in the literature and refers to a change in the underlying scale used to measure something, as a result of an experience that occurs between a pretest and the post-test [24–28]. More precisely, "response-shift occurs when a respondent's internal metric or frame of reference is changed during the time between the pretest and the post-test, due to the effects of a training program or other intervention" [24]. Linn and Slinde argued that when a comparison is made between a pretest and a post-test, there is an underlying assumption that the scale is the same at both points in time [29]. However, the purpose of many experiences (like taking a physics course) is to change the participants' awareness or understanding of a given construct (like physics problem solving) [26]. Therefore, during the experience, the underlying metric of how participants assess the construct changes, and they are better prepared to assess themselves relative to what they thought they knew about the construct at the beginning of the experience [26]. To keep the underlying metric the same, a reflective pretest can be used as the comparison benchmark for the post-test [26].

Cantrell advanced the use of reflective pretests to account for response-shift bias in science teaching self-efficacy [24]. Cantrell conducted a study on the change in science teaching self-efficacy of preservice teachers during a science methods and practicum course [24]. The difference in scores produced by a traditional pretest and post-test were compared to those produced by a reflective-pretest on the science teaching efficacy belief instrument. Results showed that the traditional pretests and post-tests made it seem as though participants' change in self-efficacy was smaller than when the post test was compared to the reflective pretest. This difference was due to a shift in participants' conception of teaching during the science methods course. Follow-up interviews were conducted, which provided evidence of internal validity for the reflective pretest (compared to the traditional pretest). Interviews revealed that the greater difference (between the post-test and the reflective pretest) was, at least in part, due to the fact that participants rated themselves higher on the traditional pretest (on the first day of class) than they did on the reflective pretest (on the last day of class), "because they did not know what they did not know" [24] (p. 181). All the interviewees commented on the fact that the methods course changed their thinking and attitudes towards teaching and that they used a different internal rating metric at the end of the course than what they had used at the beginning. The participants also stated that the reflective pretest provided a more valid metric than the traditional pretest, because at the end of the course they had a new frame of reference with more information.

Hechter used a similar research design to study the changes in perceptions of science teaching self-efficacy through pre, post, and reflective administrations of the science teaching expectancy belief instrument among preservice elementary teachers. Hechter was interested in understanding how taking a science teaching methods course impacted the science teaching self-efficacy of preservice teachers. Hechter found that the preservice teachers' understanding of their roles in science teaching changed during the science methods course. The findings revealed that the preservice teachers demonstrated a response-shift bias at the end of the course. That is, when they started the methods course, the preservice teachers had a statistically significantly inflated perception of their science teaching self-efficacy compared to when asked to reflect on their pretest self-efficacy at the end of the course. Hechter found the preservice teachers self-reported a significantly lower level of science teaching self-efficacy on the reflective pretest than on the pretest. Hechter cited the purpose of using the reflective pretest was to minimize the response-shift bias that occurred during the experience of the methods course [26].

Cartwright and Atwood also studied the response-shift bias of preservice elementary school teachers when measuring self-efficacy and attitudes toward science during a

methods course [28]. They found that the teachers experienced a significant response-shift bias in the constructs related to self-efficacy, confidence, and attitudes toward science but not in some of the other constructs they measured (expectancy, value, and relevancy of science). They concluded that response-shift bias occurs when respondents' initial constructs (such as self-efficacy in teaching science), are incomplete because they do not fully understand something they have yet to experience. The authors also concluded that program evaluators who rely on a pretest and post-test assessment to measure the effectiveness of a course on the self-efficacy of preservice teachers are prone to think the course ineffective when really, during the course, preservice teachers develop a more critical self-analysis (also called response-shift bias) and this masks the change in self-efficacy as a result of the course. Similar to the conclusion made by Cartrell and Hechter, Cartwright and Atwood advocate for the administration of a reflective pretest in cases where participants do not have a complete understanding of the construct being measured [28].

The problem of response-shift bias is also well established in the measurement literature. Howard and Dailey discussed the issue with self-reported measures and suggested that moving the pretest closer to the post-test helps minimize response-shift bias [30]. Others have promoted the use of a reflective pretest to reduce response-shift bias in self-reported measures, especially in the field of education [30–32].

To summarize the literature on response-shift bias and reflective pretests, in a typical self-reporting pretest or post-test design, changes in self-efficacy are assumed to be due to real changes in the latent construct. Yet, a shift in a person's conceptualization of a construct (in our case, academic physics activities) during the post measurement may result in a misleading comparison with the pretest self-efficacy measures. Measurements of self-efficacy conducted with pretests and post tests may fail to reveal if the source of any change in a self-efficacy score is due to a real change in self-efficacy or a shift in a person's conception of a specific aspect of physics (like problem solving) [25–28]. Reflective pretests in combination with prompts that encourage participants to explain why they chose one self-efficacy rating over another can help educators and researchers understand if changes in self-efficacy are being masked by response-shift bias. Considering response-shift bias is critical in any educational evaluation that uses self-reported participants survey data. Response-shift bias could mask the effectiveness of pedagogical interventions, leading to erroneous conclusions about the usefulness of pedagogical strategies.

C. Other kinds of bias

While reflective pretests can certainly address concerns about response-shift bias, educational researchers argue that reflective pretests might introduce additional sources of bias when compared to traditional pre- and post-tests [25]. Cartwright and Atwood identify two types of bias that can be introduced when using reflective pretests [28]. Effort justification bias and self-enhancement bias are both related to a person's desire to demonstrate improvement on a particular construct. Effort justification bias happens when people who have expended effort to improve in a dimension, have a desire to provide evidence of improvement to justify their work. Self-enhancement bias emerges when a person exaggerates the gain or change being measured, to present their learning in a more positive light [28].

D. Purpose of this paper

The purpose of this paper is to investigate if there is evidence of response-shift bias in the self-efficacy data collected during an actively taught physics class towards the goal of better understanding the change in students' self-efficacy in this learning environment. We measured students' physics self-efficacy in an actively taught course using a traditional pretest and post-test with an additional "reflective" pretest. We also interviewed students to understand the reasoning of their self-reported self-efficacy scores and to assess if there was any evidence of response-shift bias (or other types of biases discussed in the literature). The reflective pretest adds a third data point on student self-efficacy and, in combination with the interviews, provides additional insight into how students' beliefs about their own abilities develop and change over the semester of an actively taught physics course.

II. METHODS

We measured students' physics self-efficacy at both the beginning and the end of the semester in an introductory physics course at Harvard University. The course, Applied Physics 50 (AP50A), uses interactive teaching strategies and is both team and project based. We collected data over two consecutive years of this course (Fall 2016; $N = 65$, 37 female, 28 male; Fall 2017; $N = 39$, 25 female, 14 male). The population was 48%–50% premedical students and 50%–52% engineering students, and there was an even distribution of students in their sophomore, junior, and senior years. Students completed the physics self-efficacy survey (PSES) during the first week of the class (as a requirement for enrolling in the class) and again at the end of the semester, after all the course requirements were completed but before students received their final course grade. Students were incentivized to complete the survey at the end of the semester with participation points and all enrolled students completed both the pre- and post-tests. The post-test required students to complete the survey twice, once evaluating their self-efficacy at the time they were completing the survey (i.e., at the end of the semester) and a second time, evaluating their self-efficacy looking back at themselves at the beginning of the semester (i.e., a reflective measure). At the end of the semester, we also

conducted semistructured interviews for a subset of 9 students with the goal of understanding students' thinking when they completed the reflective self-efficacy survey.

A. Pedagogy

The pedagogy in AP50A combines features from both project-based learning [33] and team-based learning [34]. All the learning goals for the course are addressed through three, month-long projects that students work on in teams of 4–5. Projects are inquiry driven and inspired by a real-world problem which incorporates the physics concepts that students are learning in class. By researching and problem solving, students work towards mastery of the knowledge and skills in specific content areas. The projects require students to build a machine which meets specific design constraints (like a Van der Graaf generator, or Rube Goldberg Machine) and then study the underlying physics of the machine. For each of the three projects, students are assigned to a team. Students work on a different team for each project and teams are formed to ensure that students do not work together twice in the same semester. Teams are also constructed to ensure that they are well balanced and diverse with respect to several student characteristics (incoming physics knowledge, gender, college major, year in college, previous experience with building). Students work in these project teams during all in-class activities, including assessments, which have both an individual and team component. In-class, AP50A consists of a blend of four different types of activities, each of which provides students with scaffolding to help them learn content and acquire skills necessary to be successful in the projects. These activities are described in more detail in the next section. There are no lectures in class. The content delivery aspect of the course takes the form of pr class reading assignments posted online on a social annotation reading platform, called *Perusall*.[1] Before each class students are required to log onto *Perusall* and complete the reading assignment. *Perusall* requires that students not only read the assigned text but also annotate it as well as engage with classmates by asking and answering each other's questions about the reading.

B. In-class activities

AP50A meets twice a week with each class lasting 3 h. During each class, the instructor leads the students through 1–3 activities with the more structured activities (for example, Peer Instruction) at the beginning of the class. The following is a description of the four different types of in-class activities in AP50A presented in order of how structured the activity is (from most structured to least structured).

[1]www.perusall.com.

Peer Instruction.—This activity is done at the beginning of each new topic as it allows the instructor to probe students' understanding of the preclass reading and resolve difficult concepts. Peer Instruction has been shown to be an effective strategy for helping students resolve conceptual difficulties by allowing them the opportunity to discuss the concepts with peers who are in a similar zone of proximal development [35]. During the semester, the instructor will conduct approximately eight Peer Instruction sessions (approximately one every second class). Each session lasts between 1 and 2 h. During each session, students answer 8–12 ConcepTests which are short, conceptual questions that focus on a single topic [36]. Initially, students answer each ConcepTest individually and then answer a second time after discussing the question with their team.

Tutorials.—During this activity students spend 1 to 1.5 h working together in their teams on worksheets designed to address common misconceptions about the course content. Tutorials are adapted from those in the book "*Tutorials in Introductory Physics*" [37] developed by the Physics Education Group at the University of Washington.

Estimation activity.—Students are provided with five quantities related to the content of the class. During this activity, students are given 30 min to work with their teams to estimate each of the quantities to the nearest order of magnitude. Students are expected to come up with the estimates based on things that they already know and are instructed not to "Google" any information.

Problem set reflection.—Students are given a week to solve 4–5 physics problems at home. During this activity students work with their teams to discuss and improve their solutions, resolve conceptual difficulties, and reflect on areas that need to be reviewed. After the team discussion, students are provided with "official" solutions to the problems which they compare to those agreed upon by the team. At the end of this activity, students submit their revised solutions with a written reflection highlighting the aspects of the problem set they struggled with and describing how they resolved misunderstandings.

C. Assessment

Assessment in AP50A is continuous, low stakes, and formative. The assessment philosophy is to provide students with regular feedback which they can use to revise and resubmit their work. Instead of traditional exams, AP50A uses two-part collaborative exams (called *readiness assessment activities)*. Projects are evaluated by external judges in a science-fair environment during which feedback is provided for improving their designs before submitting a final report.

Readiness assurance activities.—These assessments occur at the end of each of the learning units and are designed to help ensure students master the relevant unit content. During the activity, students first work individually to solve a series of complex physics problems. They are

free to consult any resources (textbook, notes, internet) but are not allowed to discuss the problems with others. Students submit their individual responses via an online system and then work together to solve the same set of problems, about which they must come to a consensus before submitting their solutions as a team. As the team submits responses to the problems, the system provides immediate feedback on whether the responses are correct or not. If the response is incorrect, teams resubmit a response for reduced credit up to three times before the system reveals the correct answer with a detailed explanation. Students' overall score is the average of their individual score and their team score. These assessment activities provide a low stakes testing environment during which students learn together and receive immediate feedback.

Project fairs.—At the end of each project cycle, teams present their projects to judges in a science-fair-like environment. On the day of the fair, each team is provided with a "booth" or table which judges circulate around, interviewing students and having them demonstrate their projects. Judges are typically physics faculty members from outside the course and sometimes from neighboring universities. Judges serve as external evaluators and are provided a scoring rubric to assess the projects, the students' ability to explain the underlying physics, and answer questions. Students are scored as a team, and it is made clear to them that it is each team's responsibility to make sure that all its members are prepared to answer the judges' questions. During the interview process judges provide teams with feedback on both their project designs and their explanations of the physics so that students can use this feedback when they write their final project reports.

D. Self-efficacy surveys (including validation)

For this study, we measured physics self-efficacy with the physics self-efficacy survey. This survey can be found, in its entirety, in Appendix A. We developed the PSES by adapting the source of self-efficacy in science courses (SOSESC) survey [38,39] as part of our previous study on self-efficacy [23]. Complete details on the development and validation of the PSES can be found in this earlier work [23]. The PSES measures students' self-efficacy across four different dimensions of academic activities carried out in a collaborative introductory physics course: conceptual physics understanding (CPU), problem solving (PS), collaborative work (CW), and lab or hands-on activities (LHA). The survey consists of 20 items total, five items for each of the four dimensions. Appendix A indicates which PSES question pertains to each of the four dimensions of physics self-efficacy. It is important to note that most other science self-efficacy instruments (the SOSESC, for example) do not measure students' self-efficacy in the collaborative work and lab and hands-on dimensions and instead focus on self-efficacy in problem solving and conceptual understanding. Given the emphasis of AP50A on both teamwork and hands-on activities, we were particularly interested in measuring students' self-efficacy in these two dimensions.

E. Interviews

At the end of the semester, we conducted a semistructured interview with the purpose of understanding the students' reasoning for the difference in their self-efficacy between the reflective pretest and the pretest. An overview of the interview questions (and rationale for each question) is outlined in Table I. We want to emphasize that the

TABLE I. Interview protocol (questions and rational).

Question	Rationale
1. Talk about your experience in AP50, generally speaking.	Provide an opportunity for students to speak freely about their experiences before we start asking more detailed questions.
2. Compared to the beginning of the course, how do you feel about your ability to understand conceptual physics, work collaboratively, solve physics problems, and perform lab and hands-on activities?	Probe students' self-efficacy across the four different dimensions at the beginning of the course compared to at the end of the course.
3. Considering your reported change in your ability to (understand conceptual physics, work collaboratively, solve physics problems, and perform lab and hands-on activities), cite possible factors and/or situations that contributed to this change	This question will allow us to uncover the "sources" of self-efficacy, i.e., the specific components and events in the course that led to the students change in self-efficacy
4. In the reflective self-efficacy survey, you rated your ability to understand conceptual physics, work collaboratively, solve physics problems, and perform lab and hands-on activities differently than at the beginning of the semester (on the pretest). Could you explain this change? (Here we will provide students with the numbers that they chose when rating themselves in the four dimensions both during the pretest and the reflective test)	Students will explain the reasons that led them to readjust their self-efficacy beliefs from the beginning of the semester to the end of the semester (in the reflective test)

interview protocol provided was designed to provide guidance for interviewers about the domains of interest during the interview. The interview protocol, however, was not prescriptive; the interviewers asked clarifying questions as necessary and were free to probe salient comments by students in ways that may not have been specifically outlined in the protocol [40].

We used purposeful sampling to recruit ten students (5 males and 5 females) to be interviewed and nine of these students (5 males and 4 females) volunteered to participate in the interviews. Students were selected based on their self-efficacy scores in the pretest compared to the reflective pretest across the four different physics dimensions. Specifically, we were interested in interviewing students with PSES scores that indicated large differences in self-efficacy between the pretest and the reflective pretest in each of the four dimensions as well as overall (on average, across all 20 questions). We selected eight students (one male and one female) with large differences from each of the four dimensions and two students with a large difference in the overall self-efficacy score.

In the analysis of the interviews, we followed the procedure outlined by Yin [41]. Yin proposes a qualitative procedure for analyzing interviews which can be divided into five phases: compiling, disassembling, reassembling, interpreting, and concluding [41].

The transcripts of the interview recordings were compiled and reviewed for emergent themes. This was done by disassembling the text into smaller fragments of text. Then the fragments of text were reassembled within emergent categories, related to the research question. Originally, we were interested in identifying pedagogical sources of self-efficacy for students in the course. We note that this broader question is not the topic of this paper. Briefly, themes that emerged in the interpreting and concluding steps of the analysis were collaborative teamwork, peers as a reference for social comparison, the influence of hands-on activities, and the role of previous physics academic experiences emerged as sources of self-efficacy. However, in the process of interpreting the interview data in a narrative way we found an additional theme that we could not relate to a source of physics self-efficacy. This theme was related to response-shift bias and is the focus of this paper. A summary of the findings from these nine interviews, relating to response-shift bias specifically, is compiled in Table V (Appendix B).

III. RESULTS

Figure 1 shows students' average self-efficacy scores at the beginning of the semester (pre), at the end of the semester (post) and at the end of the semester reflecting back on their abilities at the beginning of the semester (reflective) for 2016 and 2017 combined. For all three measures, the average self-efficacy score is calculated by averaging students' responses across the 20 items on the PSES survey. We performed two-tailed t tests to compare pre, post, and reflective self-efficacy.

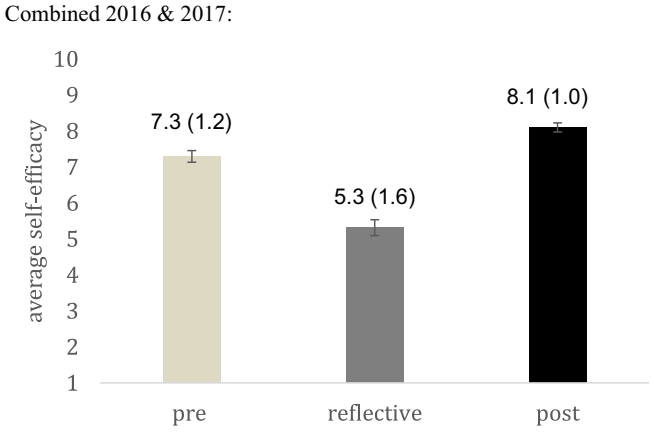

FIG. 1. Combined average student self-efficacy at the beginning of the semester (pretest), at the end of the semester (post-test), and at the end of the semester reflecting back on their abilities at the beginning of the semester (reflective test). $N = 104$ combined for 2016 and 2017. Mean (and standard deviation) are indicated for each test.

Results indicate average post-self-efficacy ($M = 8.10$, $SD = 0.99$) was significantly higher than average pre-self-efficacy ($M = 7.30$, $SD = 1.18$), $t(103) = -6.06$, $p < 0.001$. The effect size (Cohen's d value) for this difference is 0.68, which is considered to be a medium effect size [42].

Average reflective-self-efficacy ($M = 5.30$, $SD = 1.55$) was significantly lower than pre-self-efficacy $t(103) = -14.67$, $p < 0.001$ and significantly lower than post-self-efficacy $t(103) = -19.98$, $p < 0.001$. The effect size for each of these differences are 1.69 and 1.81, respectively.

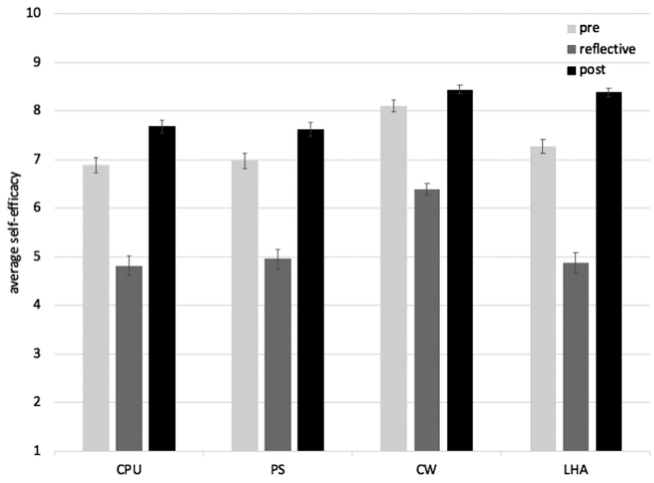

FIG. 2. Combined average student self-efficacy for the pretest, post-test, and reflective pretest across the four dimensions of physics self-efficacy; conceptual physics understanding, problem solving, collaborative work, lab and hands-on activities. Error bars represent the standard error of the mean.

TABLE II. Summary of differences between pre and post-self-efficacy across all four dimensions.

Dimension	Pre-SE $M(SD)$	Post-SE $M(SD)$	$t(103)$	p	Effect size (Cohen's d)
Conceptual physics understanding	6.88(1.51)	7.68(1.30)	5.41	0.0001	0.57
Problem solving	6.97(1.55)	7.61(0.13)	4.06	0.0001	0.44
Collaborative work	8.10(1.23)	8.43(0.93)	2.70	0.008	0.30
Lab and hands-on understanding	7.27(1.44)	8.38(0.98)	7.76	0.0001	0.90

Both of these effect sizes are considered to be very large [43]. We find the trend of student self-efficacy to be nearly identical for 2016 and 2017 and so we combined the data from both semesters for all analyses. It is also worth noting that all data for the nonreflective pretest and post-test have already been published [23]. These results expand on those published earlier with the addition of the reflective pretest and the semistructured interviews.

Figure 2 shows the combined (2016 and 2017) average self-efficacy for the pretest, post-test, and reflective pretest for each of the four dimensions of physics self-efficacy (conceptual physics understanding, problem solving, collaborative work, and lab and hands-on activities). The trend of pre-, post-, and reflective self-efficacy is consistent across the four measured dimensions. Average post-self-efficacy is significantly higher than average pre-self-efficacy in all four dimensions.

Table II summarizes the differences between the average pre-self-efficacy (SE) and post-self-efficacy across the four dimensions.

At the end of the semester, students rated their abilities in all four dimensions as significantly lower when reflecting back to the beginning of the semester compared to when they rated their abilities at the beginning of the semester (reflective test versus pretest). The biggest readjustment of reflective self-efficacy (compared to the pretest) occurred in the lab and hands-on activities and collaborative work dimensions. The dimension in which there was the smallest readjustment was problem solving.

Table III summarizes the differences between the average pre-self-efficacy and reflective-self-efficacy across the four dimensions.

Given students' tendency to readjust their presemester self-efficacy to substantially lower values when reflecting back on this time at the end of the semester, we see significantly larger changes in self-efficacy across all four dimensions between the post-test and the reflective test compared to the differences between the post-test and the pretest. Students' reflective pretest is significantly lower than their pretest in all four dimensions. Additionally, students' postsemester self-efficacy is significantly larger than their reflective self-efficacy in all four dimensions. The dimension with the largest average difference between post-test and reflective test was again lab and hands-on activities. We see the same trend in the other three dimensions where the difference between the post-test and the reflective test is significantly larger than the difference between the post-test and the pretest. Figure 2 represents a summary of the average self-efficacy across all four dimensions for the pre, post, and reflective tests.

Figure 3 shows the combined (2016 and 2017) average self-efficacy scores across the 20 individual survey questions for the pretest, post-test, and reflective test. Questions 1–5 measure conceptual physics understanding self-efficacy, questions 6–10 measure problem solving self-efficacy, questions 11–15 measure collaborative work self-efficacy, and questions 16–20 measure lab and hands-on activities self-efficacy. Notably, the overall trend for the pre, post, and reflective tests is fairly constant over all 20 items. The trend across the 20 items is virtually the same, just shifted up and down on the y axis with the post-test self-efficacy being (marginally) higher than pretest and both being substantially higher than the reflective measure. Although the difference between the post-test and pretest is small, it is statistically significant for most questions. The only questions for which there is no significant difference between the pre- and post-tests are questions 16, 17, and 18 which are all measures of student self-efficacy in the lab and hands-on activity dimension.

TABLE III. Summary of differences between pre and reflective self-efficacy across all four dimensions.

Dimension	Pre-SE $M(SD)$	Reflective-SE $M(SD)$	$t(103)$	p	Effect size (Cohen's d)
Conceptual physics understanding	6.88(1.51)	4.82(1.95)	12.55	<0.0001	1.18
Problem solving	6.97(1.55)	4.95(1.98)	12.39	<0.0001	1.13
Collaborative work	8.10(1.23)	6.38(1.29)	12.80	<0.0001	1.36
Lab and hands-on understanding	7.27(1.44)	4.87(2.03)	11.67	<0.0001	1.36

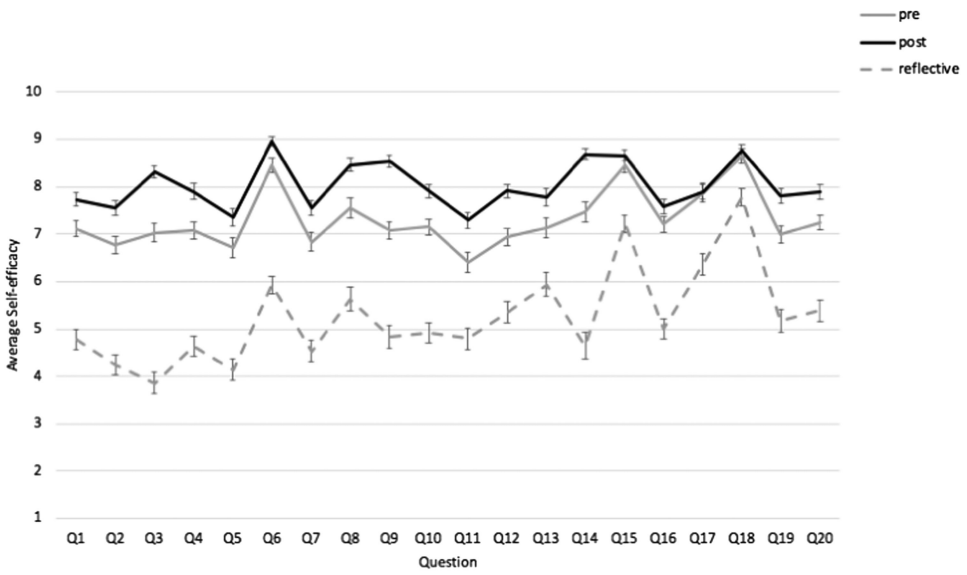

FIG. 3. Combined average student self-efficacy across the 20 survey questions for the pretest, post-test, and reflective test. Questions 1–5 measure conceptual physics understanding self-efficacy, questions 6–10 measure problem-solving self-efficacy, questions 11–15 measure collaborative work self-efficacy, questions 16–20 measure lab and hands-on activities self-efficacy. Error bars represent the standard error of the mean for each question.

Figure 4 shows the combined average self-efficacy for the pretest, post-test, and reflective tests broken down by gender. Male students' average pretest self-efficacy ($M = 7.90$, $SD = 1.30$), is significantly higher than female students' pretest self-efficacy ($M = 6.90$, $SD = 1.10$), $t(102) = 4.27$, $p < 0.001$. On the reflective test, male students' self-efficacy ($M = 5.70$, $SD = 1.60$), is also significantly higher than female students' pretest self-efficacy ($M = 4.90$, $SD = 1.50$), $t(102) = 2.24$, $p = 0.03$. However, there is no statistically significant difference between male and female students' self-efficacy at the end of the semester on the post-test. When evaluating their self-efficacy at the beginning of the semester (both during the pretest and reflecting back on that time at the end of the semester through the reflective test) male students have higher self-efficacy than female students. When students are evaluating their self-efficacy at the end of the semester (during the post-test), however, the gender gap in self-efficacy disappears.

Tables IV and V (Appendix B) provide a summary of the data collected from the semistructured interviews. Table IV identifies the gender of the interviewee and summarizes their pre, post, and, reflective pretest scores. Table IV also summarizes the type of bias demonstrated by each of the interviewed students. Table V provides excerpts from the transcripts which illustrates the biases demonstrated by the students. Additionally, Table V summarizes the components of the course interviewees indicated led to a

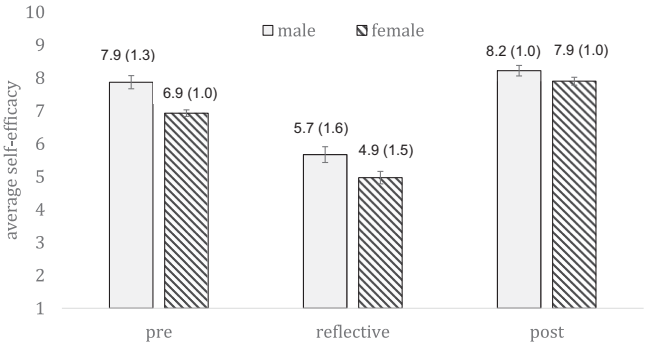

FIG. 4. Combined average self-efficacy for the pretest, post-test, and reflective test for male students versus female students. Mean (and standard deviation) are indicated for each test and each group.

TABLE IV. Summary of semistructured interviews.

ID	Gender	Pre (avg)	Post (avg)	Reflective (avg)	Type of bias
1	F	6.8	5.8	3.8	Response shift
2	F	6.8	7.9	4.0	Response shift
3	M	7	8	2.6	Response shift
4	M	7.2	9.4	2.2	Response shift
5	M	8.1	8.35	5.3	Response shift
6	F	7	7.8	3	Response shift
7	M	7.3	7.6	3.4	Response shift
8	M	4.9	7.2	3.7	Effort justification and response shift
9	F	7.9	8.1	6.6	Response-shift bias

shift in their self-efficacy (sources of change) as well as supporting quotes. These quotes are from students' responses to the last two questions from the interview protocol (found in Table I). The column labeled "source of change quote" contains excerpts of student responses to question 3, which asked students to explain the specific components of the course that led to their change in self-efficacy (between the pretest and the post-test). The column labeled "bias quote" contains excerpts of student responses to question 4, which asked students to explain what led them to readjust their self-efficacy beliefs from the beginning of the semester to the end of the semester (in the reflective test). Quotes were selected to illustrate examples of biases in students' responses to these questions as well as to highlight the individual course components which students highlighted as contributing to a shift in self-efficacy. All nine of the students who were interviewed demonstrated evidence of response-shift bias in their responses. These students argued that their way of thinking about and evaluating physics self-efficacy had changed during the experience of taking AP50A and that this shift in thinking led to a decrease in self-efficacy between the traditional pretest and the reflective pretest.

In the words of one student (ID#8):

> "I don't know how to explain it other than saying it's kind of a correction looking back."

Most of the students interviewed indicated that, at the beginning of the semester, they "did not know what they didn't know" and that, coming out of high school, they thought they understood physics better than they did.

One student (ID#7) said,

> "I realized how little I actually knew once I had gotten into—once I'd finished the course."

The experience of taking AP50 caused them to change their frame of reference for evaluating their own physics efficacy as they realized that they did not understand physics as well as they thought they did going into the course. This alteration of the internal metric or frame of reference because of the experience is the very definition of response-shift bias [24]. In addition to a response-shift bias, one of the students (ID#8) also demonstrated an effort-justification bias, indicating that he felt like he had learned less than expected and that he thought he had lowered his reflective pretest scores to "make me feel like I learned more." This desire to demonstrate improvement on a particular construct to justify invested effort is a classic demonstration of effort justification bias.

IV. DISCUSSION

Similar to the findings of Cantrell [24], Cartwright and Atwood [28], and Hechter [26], we find that students had a statistically significant inflated perception of their physics self-efficacy (both overall and in each of the four dimensions we looked at). In agreement with these three studies, we find substantially larger changes in self-efficacy between the post-test and the reflective test compared to the differences between the post-test and the pretest. As was the case in these aforementioned studies, our results show that if we only consider the traditional pre-post-test, participants' change in self-efficacy is significantly smaller than when we compare the post-test to the reflective pretest.

It is interesting that this readjustment of self-efficacy between the pretest and the reflective pretest appears to be uniform across gender lines. The gender gap that exists in the pretest is almost identical to that in the reflective pretest (see Fig. 4). This is particularly interesting given that this gap goes away in the post-test and this needs further investigation.

Hechter found that the preservice teachers' understanding of their roles in science teaching changed during the science methods course [26]. Similarly, the interviews conducted in our study reveal that students' understanding of the construct changed during the experience of taking AP50A and that this shift in thinking led to a decrease in self-efficacy between the traditional pretest and the reflective pretest. The new experience of taking AP50A caused students to redefine their understanding of conceptual physics understanding, problem solving, collaborative work, and lab or hands-on activities compared to how they thought about these dimensions of physics at the beginning of the course. Before the course, the students did not have enough information to make an accurate judgment of their abilities in these four dimensions. They based their traditional pretest measurements on their previous experiences in physics courses, other science courses, and in life itself. These experiences were different from their experiences in the active learning environment of AP50A. All nine of the students interviewed made this point, either generally or with respect to a specific dimension.

Student ID#3, for example, argued that his experiences in high school made him think he was better able to understand physics (conceptually) than he really was. His average pretest self-efficacy score in the CPU dimension was 7.4 and then, at the end of the semester, he readjusted it to 2.6 on the reflective pretest. HIs CPU self-efficacy on the post-test was 8. As illustrated in Table II, ID#3 realized that the conceptual understanding required at the college level is different from that required in high school:

> "Primarily, the reason that number {reflective self-efficacy} was so low is when coming into college physics, you know, having taken high school physics, you think that you understand physics, you know, on a decent level, and then kind of having a little bit more of a rigorous college physics class, I understood that I did not really understand, these physics concepts on a

> deep level that's required at a college level. I think that was probably the main reason that number was so much lower."

The new experience of taking AP50 changed the metrics this student used to judge his own capabilities in the conceptual physics understanding dimension.

Another student (ID#8) had averages of 6.8 (traditional pretest), 3.2 (reflective pretest), and 7.2 (post-test) in his problem-solving self-efficacy. He explained the difference between the pretest and the reflective pretest by saying that he did not realize the "AP50 definition of problem solving."

The following quote summarizes this realization:

> "I've always been a decently good problem solver. I've always been relatively confident in my problem solving and then—after the course—I realized—in the way the course wants you to solve problems—maybe I'm not too good at problem solving."

Student ID#7 explains how his average (pretest) self-efficacy in collaborative work was based on his vast experience collaborating with others in athletic environments. After taking AP50, he realized that teamwork in an academic environment differs from teamwork in sports and he re-adjusted the parameters with which he used to evaluate himself on the reflective pretest (which was a 7.6 in CW).

This quote illustrates this realization and the subsequent readjustment:

> "For context, I'm on the varsity rowing team here. And from that, I knew that I had a strong base in collaboration and teamwork that I could bring to the class. So that's why I had a high score at the beginning. I think that throughout the course, I started realizing that athletic teamwork and academic teamwork kind of worked in different ways. And people have different commitment levels and levels of excitement, and I realized that maybe I was working with more similar types of people in rowing than I was in AP50. So, I think that's why when I evaluated myself, at the end of the course, I saw that I maybe wasn't as good at working with people who, like, work best between 3 and 6 a.m., because I'm in bed then. I would find myself struggling to maybe empathize or understand how some other students couldn't respect my time. And sometimes that was frustrating. And I evaluated myself a little bit lower, because I thought, well, I could be better at dealing with those situations."

This students' post-test self-efficacy in collaborative work was 9.2 (compared to a pretest score of 9). Using only the traditional pretest, it would appear as though this student's collaborative work self-efficacy did not improve over the course of the semester. However, both the students' responses to the interview questions as well as his reflective pretest score demonstrates that his way of thinking about collaborative work changed during the experience of taking the course and this readjustment made it difficult to measure any shift in self-efficacy with the traditional pre-post-test.

This same student (ID#7) followed a similar pattern for his lab and hands-on activity self-efficacy. This student had average self-efficacy scores of 7.6 (traditional pretest), 2 (reflective pretest), and 10 (post-test) in the lab and hands-on experiences dimension. He had considerable lab experience from high school, which caused him to overestimate his capabilities at the beginning of the semester. The kinds of lab experiences introduced in AP50A were unknown to the student and this caused him to readjust the parameters he used to judge his ability to perform lab or hands-on activities during the reflective pretest:

> "I had a lot of experience in the lab—I thought I would be a little bit better in that part of AP50 but, AP50 lab is not like any other science lab. I didn't have much experience with building or working in a workshop so, I think that was why I rated myself lower after I realized that was what the AP50 lab experience was like."

By evaluating the students' "sources of change" responses, we determined that there was no one course component that stood out above the others as being solely responsible for students' change in self-efficacy over the course of the semester. Most of those interviewed commented on the importance of the projects in leading to a change in their self-efficacy. Many students mentioned the hands-on, building aspect and the importance of the projects in demonstrating physical concepts and improving their conceptual understanding. Four students mentioned the importance of the experience of working collaboratively as a part of a team and the impact that had on their self-efficacy and overall learning. Several students also mentioned the important role of problem solving in the course and how working on the problem sets with their team and having to explain the problems to one another led to an improvement in their understanding.

The fact that female students' self-efficacy at the beginning of the semester (both as measured by the pretest *and* the reflective pretest) is statistically significantly lower than the self-efficacy of male students is very interesting. It seems that female students rate their precourse self-efficacy lower than male students regardless of whether they are evaluating it at the beginning of the semester or at the end of the semester (through the reflective pretest). Female students' postcourse self-efficacy is statistically indistinguishable from that of male students. The interviews did not provide any insight into the difference between male

and female students' self-efficacy on either the pretest or the reflective pretest. The interviews also, unfortunately, did not illuminate the reason for why this gender gap in self-efficacy disappears at the end of the semester beyond the reasons we have already discussed in our previous work on this topic [23]. In future work we will strive to better understand the gender differences in self-efficacy as seen in both the pretest and the reflective pretest.

V. CONCLUSION

Our results indicate that students' lack of experiences in active learning environments (such as AP50) can lead them to overestimate their capabilities in the four measured physics dimensions. At the beginning of the course, students did not know what they did not know. They had no useful, comparable experiences to use to evaluate their own abilities in the context of this actively taught class. Our student interview results demonstrate how response-shift bias can obfuscate improvements in self-efficacy scores when only using a pre- and postdesign. Thus, in agreement with previous studies, considering response-shift bias is critical in evaluating educational experiences that use self-reported survey data.

In addition, our results show that students' response-shift bias is often masking a shift in their physics self-efficacy, especially with respect to specific dimensions. Generally, this masking is problematic when evaluating the effectiveness of pedagogical interventions and can lead to erroneous conclusions about the usefulness of pedagogical strategies especially in cases where the educational experience is completely new to the participants, and they have no useful benchmark for evaluating their own efficacy in the new context. Similar to the conclusion drawn by Cartwright and Atwood [28] we conclude that program evaluators who rely solely on a pre- and post-test assessment to measure the effectiveness of a course on student self-efficacy are prone to incorrectly evaluate the course as ineffective. Because of students' limited experience with active learning environments, response-shift bias makes it difficult to accurately measure students' change in self-efficacy over the semester of an actively taught physics course. Although more work needs to be done on the use of reflective pretests in education to examine their validity as a data collection method. Reflective pretests in combination with interviews which probe participants' reasoning behind their self-efficacy rating can help educators and researchers understand if changes in self-efficacy are being masked by response-shift bias.

APPENDIX A: PSES SURVEY

The appendix contains the physics self-efficacy survey designed to measure students' self-efficacy across four dimensions: conceptual physics understanding, problem solving, collaborative work, and lab and hands-on activities. All PSES items are tagged with their respective dimensions. The tags were not shown to respondents.

Physics Self-efficacy Survey.

For each statement, rate your belief in your ability to do the following tasks by recording a number from 0 to 10.

0 = Highly certain cannot do.

5 = Moderately certain can do.

10 = Highly certain can do.

[Dimension of physics self-efficacy that the question pertains to].

1. Understand physical concepts [Conceptual physics understanding].
2. Relate different physics concepts with each other [Conceptual physics understanding].
3. Design physics experiments using materials in hands-on activities (i.e., in class or in lab) [Lab and hands-on activities].
4. Communicate physics in a way that my classmates understand [Collaborative work].
5. Answer conceptual physics questions in class by myself [Problem solving].
6. Work together with my classmates to complete a complex task (e.g., a physics project) [Collaborative work].
7. Solve qualitative physics problems [Problem solving].
8. Collect data while conducting physics experiments [Lab or hands-on activities].
9. Write reports summarizing physics experiments [Lab or hands-on activities].
10. Relate physics concepts with daily life applications [Conceptual physics understanding].
11. Interpret the physical meaning of an equation [Conceptual physics understanding].
12. Interpret graphs explaining physical phenomenon [Conceptual physics understanding].
13. Handle mathematical calculations while solving physics problems [Problem solving].
14. Use the equipment during hands-on activities (e.g., in class or in lab) [Lab or hands-on activities].
15. Be flexible in the face of conflicts and disagreements in group activities [Collaborative work].
16. Evaluate the plausibility of results of physics problems [Problem solving].
17. In group activities, encourage my classmates to participate in discussions [Collaborative work].
18. In a discussion, listen to the opinion of my classmates, even when I think I am right [Collaborative work].
19. Apply physical equations in order to solve physics problems [Problem solving].
20. Interpret data while conducting physics experiments [Lab or hands-on activities].

APPENDIX B

TABLE V. Summary of semistructured interviews.

ID	Bias quote (Response to Q4)	Source of change	Source of change quote (Response to Q3)
1	"I think AP 50 exposed me to a lot of different formulas and, just like manipulations of variables and things that I wasn't familiar with—things that I didn't know, I didn't know….. So I think I was probably overestimating my background when I gave that (pre) score." "In my high school, I had no academic experience working collaboratively. But like, I'd worked in a ton of labs, doing collaborative research, I'd worked in something called Odyssey of the Mind, which is like extensive collaborative work. And so I wasn't unfamiliar with collaboration. But I also wasn't familiar with the AP50 type of collaborative experience, like, working with a team for a short period of time, the type of collaboration I had done involved working with the same team for an extensive period of time"	Projects	"I think a lot of it was also just being forced to understand the concepts to be able to present to the project fair judges and like, actually creating something requires understanding the basic concepts. But I think just like building something that's presentable for judging and being forced, like holding yourself accountable, like with judging, in order to know the concepts is what, I do feel better on the concepts themselves."
2	"I think I'm kind of biased and where I'm coming from is that I took AP physics in high school and had a pretty terrible teacher and so one of my main problems with going into engineering was that I don't like physics which I realize now was just because I had an awful teacher who made it very stressful and when I came here (to college) I thought I didn't know anything and that was stressful but really is all just very circumstantial to high school. So—I think this class made me feel more comfortable with physics and with the idea that it's not awful and impossible to wrap your head around."	Projects	"I think for me the emphasis on getting to build a lot of projects (I'm a mechanical engineer and I'm also very interested in product design) was really great. To be able to see three very distinct projects and to really build them in a hands-on way—I do think that also like helped with my understanding of the physics but I think secondarily it also was just very nice—that's what I enjoyed doing and it was great to be able to do that in a physics course"
3	"Primarily, the reason that number {reflective self-efficacy} was so low is when coming into college physics, you know, having taken high school physics, you think that you understand physics, you know, on a decent level, and then kind of having a little bit more of a rigorous college physics class, I understood that I did not really understand, these physics concepts on a deep level that's required at a college level. I think that was probably the main reason that number was so much lower."	Projects, Peer Instruction	"Probably through the projects, I improved my understanding the most. Having to explain to the judges what exactly we're doing and why we're doing it, and then being asked questions on the spot and having to answer those questions—I think that's a direct evaluation of *do you understand the material well enough to be able to explain it to somebody*? And I think one of the good things about this class is it, it forces you to tie in what we're learning in a textbook, to real life" "I also think Peer Instruction for me has been one of the most important things because like when we annotate the reading, a lot of the material is challenging and, for me I definitely don't understand every concept and point that's in the reading but Peer Instruction definitely hammers in like what you read and puts it kind of in a perspective of what do you need to know and what should you be thinking about and then being able to discuss the questions {during Peer Instruction}—I feel like that's definitely one of the most important aspects and I think that's a very crucial thing"

(Table continued)

TABLE V. *(Continued)*

ID	Bias quote (Response to Q4)	Source of change	Source of change quote (Response to Q3)
4	"I'm a senior. So, I've done a lot of pre-med classes, like basic classes at this point. So, I guess I was just feeling confident about my ability to take this class {at the beginning of the semester}…But I probably wasn't thinking specifically enough about the physics concepts and my understanding."	Peer Instruction, problem sets, annotating pre-reading assignments	"I think the problem sets were hugely helpful. I think the textbook was also very helpful for me—the process of annotating it, even though I wasn't always doing the practice problems that were scattered throughout the chapter or anything, I was still like subconsciously taking all of that stuff in. And then finally, I was applying it all during Peer Instruction sessions or the problem sets. I think I just felt like I had a solid grasp on how to approach physics problems, far more so than I did at the beginning of the semester."
5	"I think it could be that I was just maybe a bit overconfident starting this semester. And this ap50 has been a really neat experience for me to challenge how I thought I approached problems.……like, okay, maybe there were some things like with the conservation of energy that I didn't actually know, as strong, so maybe that's why I might have dropped (from the pre-test to the reflective pre-test). You know, there was a point when I reevaluated what I thought I knew at the beginning….. The class went from just doing textbook problems to really working things out. So yeah, I think that's probably what made me realize, okay, how I solve physics problems was not adequate before"	Projects, experimental design activities	"Yeah, after the course, I felt that the experimental design activities we did helped me really visually cement some of these concepts. What comes to mind is what we did with the basketball and dropping the balls, and watching the videos and tracker, that all really helped reinforce what I thought I knew and also then brought to life for me."
6	"I thought I knew physics. But then going through the semester, I realized that, like I had learned a lot. Which is why at the beginning of the semester, I went for such a high rating. But it was not accurate."	Working on teams, projects, problem sets	"I think among the activities, I think this class really focuses on real life examples. And that really helps in terms of understanding what's going on. Actually it's usually good to see things and relating things—really helps me feel more confident……You have to solve your own problems, and then tap into all the students. Like, that really helps in terms of seeing different ways of solving problems from other people's perspectives…..Oh, also Peer Instruction in the team round. Okay, understanding how other people think and then maybe applying their problem-solving techniques."

(Table continued)

TABLE V. *(Continued)*

ID	Bias quote (Response to Q4)	Source of change	Source of change quote (Response to Q3)
7	"So, I think that the post beginning 3.4 is lower than the initial, because I realized how little I actually knew, once I had gotten into, once I'd finished the course. I think that my general understanding at the beginning was informed by just my single high school course, back like four years ago, and I'd never taken any AP classes or any physics courses here at Harvard until AP50. So, there was a general kind of feeling that oh, I liked my physics class in comparison to my other science classes in high school. And so, I thought that I knew physics generally better than chemistry or biology, for example. And so that's kind of how I looked at it. I thought I was better at physics at the beginning than I was."	Peer Instruction	"So, I think the biggest thing was hearing Kelly {the instructor} speak. Her teaching style really breaks down the concepts to their bare elements. I think that when you're coming from a not very science engineering background, you're wanting to see the most basic. And then you need, like, you need the puzzle pieces to start putting things together. And I think she's very good at that. And having the chance to discuss the concepts with people during the Peer Instruction activities also really helped."
	"For context, I'm on the varsity rowing team here. And from that, I knew that I had a strong base in collaboration and teamwork that I could bring to the class. So that's why I had a high score at the beginning. I think that throughout the course, I started realizing that athletic teamwork and academic teamwork kind of worked in different ways. And people have different commitment levels and levels of excitement, and I realized that maybe I was working with more similar types of people in rowing than I was in ap50. So I think that's why when I evaluated myself, at the end of the course, I saw that I maybe wasn't as good at working with people who, like, work best between 3 am and 6 am because I'm in bed then. I would find myself struggling to maybe empathize or understand how some other students couldn't respect my time. And sometimes that was frustrating. And I evaluated myself a little bit lower, because I thought, well, I could be better at dealing with those situations."		
	"I had a lot of experience in the lab—I thought I would be a little bit better in that part of AP50 but, AP50 lab is not like any other science lab. I didn't have much experience with building or working in a workshop so, I think that was why I rated myself lower after I realized that was what the ap50 lab experience was like."		

(Table continued)

TABLE V. *(Continued)*

ID	Bias quote (Response to Q4)	Source of change	Source of change quote (Response to Q3)
8	"In general, I felt that I've learned less than I would have expected over the course of a semester from AP50. So, I think then, not obviously, willingly or consciously, I would have maybe lowered that beginning bar threshold to kind of make me feel like I learned more." "Coming into this class I had had a lot of exposure to physics in different classes and in high school and I think this might have convinced me that I had a bit more physics understanding than I really had. I've always been a decently good problem solver. I've always been relatively confident in my problem solving and then—after the course—I realized—in the way the course wants you to solve problems—maybe I'm not too good at problem solving—I didn't know as much about physics as I thought I did. So—I don't know how to explain it other than saying it's kind of a correction, looking back, being more confident in myself, I guess, originally, or not being aware of the kind of problem solving that would be required of me."	Problem sets, working in teams	"I do feel like I've gotten better at working in teams in this class ….it feels strange that the biggest thing I've learned from the physics class is how to work better in teams but I really have taken a lot from that. I mean I've learned how to work with a wide range of people. Your first group—you all get along but your second group like you really can't stand someone but we got to work with them, right? So that's been good for me because that's always been something I've kind of struggled with. I'm not a patient person and so I've had to learn how to be more patient which has been very helpful."
9	"And then I guess, looking back, I still sort of felt like I had covered the material. And so the class was a little bit more like a refresher. So I think that the biggest difference in this class from others I had taken was the collaborative component. And so I think that I had an idea of what it meant to work collaboratively in high school and in previous classes here, but I think that it was not the same. It wasn't the same to be working in teams for every single thing that we did—everything had a collaborative component. We had to discuss questions in class with our teams and then obviously the projects were collaborative. And so I think that I definitely learned a lot about working in a team and, you know, my various strengths and weaknesses working in a team. And I think that I had not, you know, had the opportunity to really explore that as much in an academic setting until this class. And so I think that the pre (test) number was reflective of the fact that I had not been asked to work as collaboratively throughout a class before as I had been during this semester. And so then the post beginning measurement reflected that like, looking back, oh, wait, I didn't know what I was doing."	Working in teams	"I think I learned a lot about how to convey ideas in a strong way and how to communicate with others. I think that there's a balance between being able to communicate in a way where your ideas are being heard and being seriously considered, but there's still just one idea on the table of many possible good ideas. And then, you know, over the semester, it sort of got to the point where it's like, okay, like I need to recognize when I have a good idea, and I need to actually accurately say, if I'm not sure about my idea, I have to say exactly what I'm not sure about. Not like, Oh, this is an overall bad idea. It's like, I think this is a good idea but, this is exactly why I'm hesitating about this. And then I think, like, sort of just being able to take feedback when people say that idea is not gonna work, you know?"

[1] President's Council of Advisors on Science, and Technology, *Prepare and inspire: K-12 education in science, technology, engineering, and math (STEM) for America's future* (President's Council of Advisors on Science, and Technology, Washington, DC, 2010). Retrieved from https://nsf.gov/attachments/117803/public/2a--Prepare_and_Inspire--PCAST.pdf.

[2] C. Dweck, *Is Math a Gift? Beliefs That Put Females at Risk, Why Aren't More Women In Science? Top Researchers Debate the Evidence*, edited by S. Ceci and W. Williams (APA Press, Washington, DC, 2007), pp. 47–55.

[3] C. Dweck, *Mindsets and Math/Science Achievement, Institute for Advanced Study, Commission on Mathematics and Science Education* (Carnegie Corporation of New York, New York, 2008).

[4] I. Goodman, C. Cunningham, and C. Lachapelle, *The Women's Experience in College Engineering* (Goodman Research Group, Inc., Cambridge, MA, 2002), https://files.eric.ed.gov/fulltext/ED507395.pdf.

[5] R. Lapan, A. Adams, S. Turner, and J. Hinkelman, Seventh graders' vocational interest and efficacy expectation patterns, J. Career Develop. **26**, 215 (2000).

[6] A. Bandura, Self-efficacy: Toward a unifying theory of behavioral change, Psychol. Rev. **84**, 191 (1977).

[7] A. Bandura, in *Encyclopedia of Human Behavior*, edited by V. S. Ramachaudran (Academic Press, New York, 1994), pp. 71–81.

[8] A. Bandura, *Self-Efficacy: The Exercise of Control* (Longman, New York, 1997).

[9] R. W. Lent, F. G. Lopez, S. D. Brown, and P. A. Gore, Latent structure of the sources of mathematics self-efficacy, J. Vocat. Behav. **49**, 292 (1996).

[10] S. L. Britner, Motivation in high school science students: A comparison of gender differences in life, physical, and earth science classes, J. Res. Sci. Teach. **45**, 955 (2008).

[11] S. L. Britner and F. Pajares, Sources of science self-efficacy beliefs of middle school students, J. Res. Sci. Teach. **43**, 485 (2006).

[12] F. Pajares, *Gender differences in mathematics self-efficacy beliefs, Gender Differences in Mathematics: An Integrative Psychology Approach* (Cambridge University Press, New York, 2005), pp. 294–315.

[13] A. Zeldin, S. Britner, and F. Pajares, A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers, J. Res. Sci. Teach. **45**, 1036 (2008).

[14] B. Zimmerman, Self-efficacy: An essential motive to learn, Contemp. Educ. Psychol. **25**, 82 (2000).

[15] S. Andrew, Self-efficacy as a predictor of academic performance in science, J. Adv. Nurs. **27**, 596 (1998).

[16] R. W. Lent, S. D. Brown, and K. C. Larkin, Self-efficacy in the prediction of academic performance and perceived career options, J. Counsel. Psychol. **33**, 265 (1986).

[17] K. D. Multon, S. D. Brown, and R. W. Lent, Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation, J. Counsel. Psychol. **38**, 30 (1991).

[18] J. Pietsch, R. Walker, and E. Chapman, The relationship among self-concept, self-efficacy, and performance in mathematics during secondary school, J. Educ. Psychol. **95**, 589 (2003).

[19] J. Dalgety and R. K. Coll, Exploring first-year science students' chemistry self-efficacy, Int. J. Sci. Math. Educ. **4**, 97 (2006).

[20] G. Trujillo and K. D. Tanner, Considering the role of affect in learning: Monitoring students' self-efficacy, sense of belonging, and science identity, CBE Life Sci. Educ. **13**, 6 (2014).

[21] R. Dou, E. Brewe, J. P. Zwolak, G. Potvin, E. A. Williams, and L. H. Kramer, Beyond performance metrics: Examining a decrease in students' physics self–efficacy through a social networks lens, Phys. Rev. Phys. Educ. Res. **12**, 020124 (2016).

[22] V. Sawtelle, E. Brewe, and L. H. Kramer, Positive impacts of modeling instruction on self-efficacy, AIP Conf. Proc. **1289**, 289 (2010).

[23] T. Espinosa, K. Miller, I. Araujo, and E. Mazur, Reducing the gender gap in students' physics self-efficacy in a team- and project-based introductory physics class, Phys. Rev. Phys. Educ. Res. **15**, 010132 (2019).

[24] P. Cantrell, Traditional vs. reflective pretests for measuring science teaching efficacy beliefs in preservice teachers, School Sci. Math. **103**, 177 (2003).

[25] L. G. Hill and D. L. Betz, Revisiting the reflective pretest, Am. J. Eval. **26**, 501 (2005).

[26] R. P. Hechter, Changes in preservice elementary teachers' personal science teaching efficacy and science teaching outcome expectancies: The influence of context, J. Sci. Teach. Educ. **22**, 187 (2011).

[27] D. Moore and C. A. Tananis, Measuring change in a short-term educational program using a reflective pretest design, Am. J. Eval. **30**, 189 (2009).

[28] T. J. Cartwright and J. Atwood, Elementary pre–service teachers' response–shift bias: Self–efficacy and attitudes toward science, Int. J. Sci. Educ. **36**, 2421 (2014).

[29] R. L. Linn and A. S. Jeffrey, The determination of the significance of change between pre-and posttesting periods, Rev. Educ. Res. **47**, 121 (1977).

[30] G. S. Howard and R. D. Patrick, Response-shift bias: A source of contamination of self-report measures, J. Appl. Psychol. **64**, 144 (1979).

[31] J. H. Bray, E. M. Scott, and S. H. George, Methods of analysis with response-shift bias, Educ. Psychol. Meas. **44**, 781 (1984).

[32] J. H. Bray and G. S. Howard, Methodological considerations in the evaluation of a teacher-training program, J. Educ. Psychol. **72**, 62 (1980).

[33] P. C. Blumenfeld, E. Soloway, R. Marx, J. Krajcik, M. Guzdial, and A. Palincsar, Motivating project-based learning: Sustaining the doing, supporting the learning, Educ. Psychol. **26**, 369 (1991).

[34] L. K. Michaelsen, A. B. Knight, and L. D. Fink, *Team-Based Learning: A Transformative Use of Small Groups* (Greenwood Publishing Group, New York, 2002).

[35] J. Lenaerts, W. Wieme, and E. Van Zele, Peer Instruction: A case study for an introductory magnetism course, Eur. J. Phys. **24**, 7 (2002).

[36] E. Mazur, *Peer Instruction: A User's Manual* (Prentice-Hall, Upper Saddle River, NJ, 1997).

[37] L. C. McDermott and P. S. Shaffer, *Tutorials in Introductory Physics* (Pearson, Boston, 2010).

[38] H. Fencl and K. Scheel, Pedagogical approaches, contextual variables, and the development of student self-efficacy in undergraduate physics courses, AIP Conf. Proc. **720**, 173 (2004).

[39] V. Sawtelle, *A Gender Study Investigating Physics Self-Efficacy* (Florida International University, Miami, FL, 2011).

[40] K. Roulston, *Reflective Interviewing: A Guide to Theory and Practice*, (Sage Publications, Thousand Oaks, CA, 2010), pp. 1–216.

[41] R. K. Yin, *Qualitative Research from Start to Finish* (Guilford Publications, New York, 2015).

[42] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Routledge, London, 1988).

[43] S Sawilowsky, New effect size rules of thumb, J. Mod. Appl. Stat. Methods **8**, 597 (2009).