PAPER • OPEN ACCESS

Do evidence-based active-engagement courses reduce the gender gap in introductory physics?

View the article online for updates and enhancements.

You may also like

Do evidence-based active-engagement courses reduce the gender gap in introductory physics?

Nafis I Karim[1], **Alexandru Maries**[2] and
Chandralekha Singh[1]

[1] Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA
15260, United States of America
[2] Department of Physics, University of Cincinnati, Cincinnati, OH 45221, United States
of America

E-mail: mariesau@ucmail.uc.edu

CrossMark

Abstract

Prior research suggests that using evidence-based pedagogies can not only improve learning for all students, it can also reduce the gender gap. We describe the impact of physics education research-based pedagogical techniques in flipped and active-engagement non-flipped courses on the gender gap observed with validated conceptual surveys. We compare male and female students' performance in courses which make significant use of evidence-based active-engagement (EBAE) strategies with courses that primarily use lecture-based (LB) instruction. All courses had large enrolment and often had more than 100 students. The analysis of data for validated conceptual surveys presented here includes data from two-semester sequences of algebra-based and calculus-based introductory physics courses. The conceptual surveys used to assess student learning in the first and second semester courses were the force concept inventory and the conceptual survey of electricity and magnetism, respectively. In the research discussed here, the performance of male and female students in EBAE courses at a particular level is compared with LB courses in two situations: (I) the same instructor taught two courses, one of which was an EBAE course and the other an LB course, while the homework, recitations and final exams were kept the same; (II) student performance in all of the EBAE courses taught by different instructors was averaged and

compared with LB courses of the same type also averaged over different instructors. In all cases, on conceptual surveys we find that students in courses which make significant use of active-engagement strategies, on average, out-performed students in courses of the same type using primarily lecture-based instruction even though there was no statistically significant difference on the pre-test before instruction. However, the gender gap persisted even in courses using EBAE methods. We also discuss correlations between the performance of male and female students on the validated conceptual surveys and the final exam, which had a heavy weight on quantitative problem solving.

(Some figures may appear in colour only in the online journal)

1. Introduction

1.1. Physics education research-based active-engagement methods

In the past few decades, physics education research has identified challenges that students encounter in learning physics at all levels of instruction [1–7]. Building on these investigations, researchers are developing, implementing and evaluating evidence-based curricula and pedagogies to reduce these challenges to help students develop a coherent understanding of physics concepts and enhance their problem solving, reasoning and metacognitive skills [8–18]. In evidence-based curricula and pedagogies, the learning goals and objectives, instructional design, and assessment of learning are aligned with each other and there is focus on evaluating whether the pedagogical approaches employed have been successful in meeting the goals and enhancing student learning.

One highly successful model of learning is the field-tested cognitive apprenticeship model [19]. According to this model, students can learn effectively if the instructional design involves three essential components: 'modelling', 'coaching and scaffolding', and 'weaning'. In this approach, 'modelling' means that the instructional approaches demonstrate and exemplify the criteria for good performance and the skills that students should learn (e.g. how to solve physics problems systematically). 'Coaching and scaffolding' means that students receive guidance and support as they actively engage in learning the content and skills necessary for good performance. 'Weaning' means gradually reducing the support and feedback to help students develop self-reliance [19]. In traditional physics instruction, especially at the college level, there is often a lack of coaching and scaffolding: students come to class where the instructor lectures and does some example problems, then students are left on their own to work through homework with little or no feedback. This lack of prompt feedback and scaffolding can be detrimental to learning.

Some of the commonly used evidence-based active-engagement (EBAE) approaches implemented in physics include peer instruction with clickers popularized by Eric Mazur from Harvard University [20–22], tutorial-based instruction in introductory and advanced courses [23–25] and collaborative group problem solving [26–29], e.g. using context-rich problems [4, 5]. In all of these evidence-based approaches, formative assessment plays a critical role in student learning [30]. Formative assessment tasks are frequent, low-stakes

assessment activities which give feedback to students and instructors about what students have learned at a given point. Using frequent formative assessments helps make the learning goals of the course concrete to students, and provides them with a way to track their progress in the course with respect to these learning goals. When formative assessment tasks such as concept-tests, tutorials and collaborative group problem solving are interspersed throughout the course, learning is enhanced [30, 31].

Moreover, technology is increasingly being exploited for pedagogical purposes to improve student learning. For example, just-in-time teaching (JiTT) is an instructional approach in which instructors receive feedback from students before class and use that feedback to tailor in-class instruction [32, 33]. Typically, students complete an electronic pre-lecture assignment in which they give feedback to the instructor regarding any difficulties they have had with the assigned reading material, lecture videos, and/or other self-paced instructional tools. The instructor then reviews student feedback before class and makes adjustments to the in-class activities. For example, during class, the instructor can focus on student difficulties found via electronic feedback. Students may engage in discussions with the instructor and with their classmates, and the instructor may then adjust the next pre-lecture assignment based on the progress made during class. When JiTT was first conceived and implemented in the late 1990s in physics classes, the required internet technology for electronic feedback was still evolving; developments in digital technology since then have continued to make electronic feedback from students and the JiTT approach easier to implement in classes. For example, Eric Mazur's Perusall system [34] allows students to read the textbook and ask questions electronically and the system uses their questions to draft a 'confusion report' which distills their questions to three most common difficulties, which can be addressed in class. It has been hypothesised that JiTT may help students learn better because out-of-class activities cause students to engage with and reflect on the parts of the instructional material they find challenging [32, 33]. In particular, when the instructor focuses on student difficulties in lecture which were found via electronic feedback before class, it may create a 'time for telling' [35] especially because students may be 'primed to learn' better when they come to class if they have struggled with the material during pre-lecture activities. The JiTT approach is often used in combination with peer discussion and/or collaborative group problem solving interspersed with lectures in the classroom.

In addition, in the last decade, the JiTT pedagogy has been extended a step further with the maturing of technology [36–41] and 'flipped' [42, 43] classes with limited in-class lectures have become common with instructors asking students to engage with short lecture videos (or read certain section of the textbook) and concept questions associated with each video outside of the class and using most of the class time for active-engagement. The effectiveness of flipped classes in enhancing student learning can depend on many factors including the degree to which evidence-based pedagogies that build on students' prior knowledge and actively engage them in the learning process are used, whether there is sufficient buy-in from students, and the incentives that are used to get students engaged with the learning tools both inside and outside the classroom.

Moreover, research suggests that effective use of peer collaboration can enhance student learning in many instructional settings in physics classes, including in JiTT and flipped environments, and with various types and levels of student populations. Although the details of implementation vary, students can learn from each other in many different environments.

For example, in Mazur's peer instruction approach [44], the instructor poses concrete conceptual problems in the form of conceptual multiple-choice clicker questions to students throughout the lecture and students discuss their responses with their peers. Heller *et al* have shown that collaborative problem solving with peers in the context of quantitative 'context-rich' problems [4, 5] can be valuable both for learning physics and for developing effective problem solving strategies.

In evidence-based 'active-engagement non-flipped' courses [45], lecture and interactive activities are combined during the prescribed class time to enhance student learning and students' out-of-class homework assignments are often similar to those assigned in traditionally taught classes. On the other hand, in flipped courses, there is very limited direct instruction (lecture) and the majority of in-class time is used to actively engage students in learning. The effectiveness of flipped classes depends on how the course is designed and incentivized and how out-of-class activities build on in-class activities. In addition, whether instructors create a low or high anxiety active-learning environment can play a critical role in student engagement. It can particularly impact learning for women and students from other underrepresented groups, whose sense of belonging and self-efficacy can either be enhanced or exacerbated depending upon the design of the active-learning environment. More information about these types of issues is discussed in the next section.

Also, the lecture videos that students often watch outside of the class in a flipped class are self-paced, which has both advantages and disadvantages. While pedagogically developed, implemented and incentivized self-paced videos can provide a variety of students with an opportunity to learn at a pace that is commensurate with their prior knowledge, without appropriate pedagogy in the development, implementation, and incentives to learn from these tools, students may not engage with them as intended, especially if they do not have good time-management and self-regulation skills. For example, research on Massive Open Online Courses [46] suggests that a majority of those who complete the entire online course already have a bachelor's degree. Moreover, a student who does not keep up with out-of-class activities such as watching videos and answering the concept questions associated with them before coming to class is unlikely to take full advantage of the interactive in-class activities in a flipped class. Thus, while a well-designed and implemented flipped course has the potential to help a variety of students learn to think like a physicist and can scaffold their learning of physics, many students may not engage and learn from the out-of-class videos if they are not intrinsically motivated and if the videos are not effective [36] or are not implemented and incentivized appropriately. Despite these caveats, well-designed and well-implemented interactive videos [47] and associated questions designed carefully can be beneficial as they can help a variety of students with different prior preparations and allow them to learn at their own pace. Moreover, if the videos are part of an adaptive video-suite for students with different prior knowledge and skills (for example, after a student views a video, he/she can be asked several questions, and if he/she struggles to answer those questions, he/she can be directed to another explanation video that other students who answer those questions correctly can skip). In particular, the videos can provide more scaffolding support as needed to a student who is struggling. Then, after taking full advantage of these out-of-class activities, the EBAE activities can help all students.

1.2. Gender gap in introductory physics courses

Prior research has found that male students outperform female students on standardised conceptual assessments such as the force concept inventory (FCI) [48] or the conceptual survey of electricity and magnetism (CSEM) [49]. The discrepancy between male and female students' performance is typically referred to as a 'gender gap' [50–52]. While sometimes the gender gap can be accounted for at least in part due to different prior preparation or coursework of male and female students, it has also been found even after controlling for these factors [50]. Prior research has also found that using evidence-based pedagogies can reduce the gender gap [53, 54], but the extent to which this occurs varies. Others have found that the gender gap is not reduced despite significant use of evidence-based pedagogies [55]. Prior research has also found a gender gap on other assessments such as a conceptual assessment for introductory laboratories [56] and physics exams [50, 51]. Yet others have found no differences in performance between male and female students on exams [52, 57, 58].

The origins of gender gap on the FCI both at the beginning and end of a physics course have been a subject of debate with some researchers arguing that the test itself is gender-biased [59]. Some of the origins of the gender gap are related to societal gender stereotypes [60–63] that keep accumulating from an early age. For example, research suggests that even six year old boys and girls have gendered views about smartness in favour of boys [63]. Such stereotypes can impact female students' self-efficacy [64, 65], their beliefs about their ability to perform well in disciplines such as physics in which they are underrepresented and which have been associated with 'brilliance'. They can also impact their intelligence mindset [66], which is related to beliefs about whether intelligence is innate or whether it is something that can be developed and cultivated via focus and persistence in problem solving in a discipline such as physics. Thus, it may not be surprising that prior research has found that activation of a stereotype, i.e. stereotype threat (ST) about a particular group in a test-taking situation can alter the performance of that group in a way consistent with the stereotype [60–63]. In fact, some researchers have argued [60] that female students, when working on a physics assessment, undergo an implicit ST due to the prevalent societal stereotypes. In particular, Marchand and Taasoobshirazi [60] conducted a study in which high school students were randomly divided into three groups and all students received the following instructions before taking a physics test: 'You will be given four physics problems to solve. These problems are based on physics material that you have already covered'. In the implicit ST condition, these were the only instructions, while in the explicit ST condition, students were also told: 'This test has shown gender differences with males outperforming females on the problems' and in the nullified condition, students were told: 'No gender differences in performance have been found on the test'. They found no statistically significant difference on the physics test between female students' performance in the explicit ST condition and the implicit ST condition but female students in both these conditions performed significantly worse than male students. In contrast, the nullified condition in which female students were instead told that the test they are about to take is gender neutral erased the gender gap (no difference in performance between male and female students). The researchers hypothesised [60] that simply administering a physics test to female students creates an implicit ST (which is partly due to societal gender bias and related issue of anxiety and self-efficacy, which refers to the fact that many female students start doubting their own ability to perform well in a physics test).

1.3. Focus of our research

In this study, we used the FCI [48] in the first semester introductory physics courses and the CSEM [49] in the second semester courses to assess student learning. We also investigated any possible gender gap at the beginning of the course as well as the extent to which evidence-based pedagogies can help reduce it. The FCI, CSEM and other standardised physics surveys [67–72] have been used to assess introductory students' understanding of physics concepts by a variety of educators and physics education researchers. One reason for their extensive use is that many of the items on the surveys have strong distractor choices which correspond to students' common difficulties so students are unlikely to answer the survey questions correctly without having good conceptual understanding. Our research focuses on the following research questions for both algebra-based and calculus-based introductory physics courses:

RQ1. What is the gender gap on the FCI/CSEM pre-test and post-test in LB and EBAE courses? By how much do both male and female students improve from pre-test to post-test in LB and EBAE courses?

RQ2. How does the performance on the FCI/CSEM of male and female students in LB courses compare to EBAE courses in both the pre-test and the post-test?

RQ3. To what extent do male and female students with high or low pre-test scores perform differently in EBAE courses compared to LB courses when the comparison is made for one instructor who teaches both an EBAE and an LB course at the same time?

RQ4. To what extent do male and female students with high or low pre-test scores perform differently in EBAE courses compared to LB courses when the comparison is made between EBAE and LB courses taught by different instructors?

RQ5. Is there any correlation between post-test and final exam scores for male and female students?

Thus, in our research, the performances of male and female students in EBAE courses in a particular type of course (algebra-based or calculus-based physics I or II) are compared with male and female students of LB courses in two situations: (I) the same instructor taught two courses, one of which was an EBAE course and the other an LB course with common homework and final exams; (II) student performances in all of the EBAE courses taught by different instructors were averaged and compared with LB courses of the same type, also averaged over different instructors.

Also, the students were divided into three subgroups based upon their pre-test scores: top 1/3rd, middle 1/3rd and bottom 1/3rd. We investigated whether there was a statistically significant difference between male and female students' average scores on the pre-test, post-test or final exam in two cases: (i) male students were divided into three subgroups according to the pre-test scores of males only and female students were also divided into three subgroups according to the pre-test scores of females only, and then the male and female students' average scores in each subgroup were compared and (ii) all students were divided into the three subgroups according to their pre-test scores *regardless* of their gender and then male and female students in each of the three subgroups were separated and compared. This type of analysis based upon gender was carried out for the male and female students taught by the same instructor (teaching either LB or EBAE course) and also for different instructors teaching LB or EBAE courses of the same type combined. Whenever differences between these two groups were observed (e.g. with male or female students in

the EBAE courses on average performing better than the corresponding students in the LB courses), we investigated which subgroup was benefiting most from the EBAE courses, e.g. those who performed well or poorly on the pre-test given at the beginning of the course. Finally, we investigated the typical correlation between the performance of male and female students' post-test performance on the validated conceptual surveys and their performance on the instructor-developed final exam (which typically places a heavy weight on quantitative physics problems).

2. Methodology

2.1. Courses and participants

The participants in this study were students in 16 different algebra-based and calculus-based introductory physics courses. Out of all introductory physics courses (algebra-based or calculus-based physics I or II) included in this study, there were four EBAE courses: two completely flipped courses in algebra-based introductory physics I and one completely flipped and one interactive active-engagement course in calculus-based introductory physics II. These courses include approximately 700 male and 750 female students in first semester courses and approximately 650 male and 500 female students in second semester courses at a typical large research university in the US (University of Pittsburgh). The details of the courses that fall into three categories are as follows:

(1) A lecture-based (or LB) course is one in which the primary mode of instruction was via lecture. In addition to the three or four weekly hours for lectures, students attended an hour long recitation section taught by a graduate TA. During recitation, the TA typically answered student questions (mainly about their homework problems which were mostly textbook style quantitative problems), solved problems on the board and gave students a quiz in the last 10–20 min.

(2) A flipped course is one in which the class was broken up into two almost equal size groups with each group meeting with the instructor for half the regular class time. For example, for a 200 student class scheduled to meet for four hours each week (on two different days), the instructor met with half the class (100 students) on the first day and the other half on the second day. This was possible in the flipped classes since the total contact hours for each instructor each week with the students was the same as in the corresponding LB courses. Students watched the lecture videos before coming to class and answered some conceptual questions which were based upon the lecture video content. They uploaded the answers to those conceptual questions before class onto the course website and were scored for a small percentage of their grade (typically 4%–8%). Although students had to watch several videos outside of class in preparation for each class, each video was typically 5–10 min long, followed by concept questions. On average, students in a flipped class had to watch recorded videos which took a little less than half the allotted weekly time for class (e.g. for the courses scheduled for four hours each week, students watched on average 1.5 h of videos each week, and in the courses scheduled for three hours each week, students watched around one hour of videos). These video times do not include the time that students would take to rewind the video, stop and think about the concepts and answer the concept questions placed after the videos that counted towards their course grade. In the spirit of JiTT, the instructors of the

flipped courses adjusted the in-class activities based upon student responses to online concept questions which were supposed to be submitted the night before the class. About 90% of the students submitted their answers to the concept questions that followed the videos to the course website before coming to the class. The web-platforms used for managing, hosting and sharing these videos and for having online discussions with students about them asynchronously (in which students and the instructor participated) were Classroom Salon or Panopto. In-class time was used for clicker questions involving peer discussion and then a whole class discussion of the clicker questions, collaborative group problem solving involving quantitative problems in which 2–3 students worked in a group (followed by a clicker question about the order of magnitude of the answer), and lecture demonstrations with preceding clicker questions on the same concepts. In addition to the regular class times, students attended an hour long recitation section which was taught the same way as for students in the LB courses.

It is important to note that the instructors who taught the flipped courses also taught LB courses at the same time (usually teaching two courses in a particular semester: one flipped and one LB). Students in both flipped and LB courses completed the same homework and took the same final exam. For the calculus-based flipped courses, the students also took the same midterm exams. This was not possible for the algebra-based courses because the exams were scheduled at different times. However, in the algebra-based courses they took the same final exam and had the same homework.

(3) In an EBAE interactive non-flipped course, the instructor combined lectures with research-based pedagogies including clicker questions involving peer discussion, conceptual tutorials, collaborative group problem solving, and lecture demonstrations with preceding clicker questions on the same concepts, similar to the flipped courses. In addition, students attended a reformed recitation which primarily used context-rich problems to get students to engage in group problem solving or worked on research-based tutorials while being guided by a TA. The instructor ensured that the problems students solved each week in the recitation activities were closely related to what happened in class. Students also worked on some research-based tutorials during class in small groups, but if they did not finish them in the allotted time, they were asked to complete them at home and submit as homework.

From now on, we refer to the flipped and interactive non-flipped courses as EBAE courses except when relevant. We also note that the number of female students in algebra-based courses is larger than that of male students. Most of the algebra-based students have biological science or related majors like Biology, Psychology, Exercise Science, Neurology/Neuromedicine, Environmental Science, etc. In calculus-based courses, on the other hand, there are more male than female students. Most calculus-based students are in their first year in college, and have physical science related majors such as chemistry, mathematics, engineering (electrical, mechanical, chemical, civil etc), and physics (typically only 5–10 physics majors out of several hundred students). The algebra-based or calculus-based physics courses are mandatory for these students. We do not have information about the background of the students, such as their prior experiences in physics or mathematics before college or whether they took any physics or math courses in high school (although a majority of these students have typically taken at least one high school physics course and the typical percentage of female students in calculus-based 'advanced placement C' high school courses in the US is less than one third).

Table 1. Intra-group FCI pre-/post-test averages (mean) and standard deviations (SD) for first semester introductory male and female students in calculus-based LB courses, and algebra-based EBAE and LB courses. The number of students in each group, *N*, is shown. For each group, a *p*-value obtained using a *t*-test shows that the difference between the pre-/post-test is statistically significant and the difference between the male and female students is also statistically significant. The normalised gain (Norm *g*) from pre-test to post-test and the effect size (eff. size) shows how much male and female students learned from what they did not already know based on the pre-test.

Type of class	FCI	Female	Gender comparison	Male
	Pre-test	*N*: 146 Mean: 43% SD: 20%	*p*-value: <0.001 ← **gender gap: 13%** → Eff. size: 0.635	*N*: 283 Mean: 55% SD: 20%
Calc LB	Pre-test versus post-test comparison	↑ *p*-value: <0.001 Eff. size: 0.468 Norm *g*: 16% ↓		↑ *p*-value: <0.001 Eff. size: 0.686 Norm *g*: 30% ↓
	Post-test	*N*: 114 Mean: 52% SD: 20%	*p*-value: <0.001 ← **gender gap: 17%** → Eff. size: 0.868	*N*: 200 Mean: 68% SD: 19%
	Pre-test	N: 153 Mean: 30% SD: 15%	*p*-value: <0.001 ← **gender gap: 13%** → Eff. size: 0.749	*N*: 105 Mean: 43% SD: 20%
Alg EBAE	Pre-test versus post-test comparison	↑ *p*-value: <0.001 Eff. size: 1.176 Norm *g*: 28% ↓		↑ *p*-value: <0.001 Eff. size: 0.955 Norm *g*: 32% ↓
	Post-test	*N*: 149 Mean: 49% SD: 18%	*p*-value: <0.001 ← **gender gap: 12%** → Eff. size: 0.653	*N*: 106 Mean: 61% SD: 19%
	Pre-test	*N*: 456 Mean: 30% SD: 14%	*p*-value: <0.001 ← **gender gap: 11%** → Eff. size: 0.691	*N*: 318 Mean: 41% SD: 18%
Alg LB	Pre-test versus post-test comparison	↑ *p*-value: <0.001 Eff. size: 0.930 Norm *g*: 21% ↓		↑ *p*-value: <0.001 Eff. size: 0.863 Norm *g*: 28% ↓
	Post-test	*N*: 383 Mean: 44% SD: 18%	*p*-value: <0.001 ← **gender gap: 13%** → Eff. size: 0.686	*N*: 255 Mean: 57% SD: 20%

We also note that none of the instructors teaching the EBAE courses focused explicitly on whether the active-learning classroom environment helped foster a sense of belonging or focused on improving self-efficacy and instilling a growth mindset in all students. In particular, the instructors did not *explicitly* focus on whether the active-learning classroom was a low anxiety classroom for all students and whether women and other under-represented students felt supported and had the same level of engagement with the active-learning activities.

2.2. Materials

The materials used in this study are the FCI and CSEM conceptual multiple-choice (five choices for each question) standardised surveys, which were administered in the first week of classes before instruction in relevant concepts (pre-test) and after instruction in relevant concepts (post-test). The FCI was used in the first semester courses and the CSEM was used in the second semester courses. Apart from the data on these surveys that the researchers collected from all of these courses, each instructor administered his/her own final exam, which was mostly quantitative (60%–90% of the questions were quantitative, although some instructors had either the entire final exam or part of it in a multiple-choice format with five options for each question to make grading easier). Ten course instructors (who also provided the FCI or CSEM data from their classes) provided their students' final exam scores and most of them also provided a copy of their final exams.

2.3. Methods

Our main goals in this research were to compare the average performances of male and female students in introductory physics courses in different types of classes (e.g. Algebra-based or Calculus-based, EBAE or LB) and to compare male and female students' performances between courses that used EBAE pedagogies with the performances of students in LB courses by using standardised conceptual surveys, the FCI (for physics I) and CSEM (for physics II) as pre-/post-tests. We not only calculated the average gain (post-test–pre-test scores) for each group for males and females but also calculated the average normalised gain, which is commonly used to determine how much students learned from pre-test to post-test taking into account their initial scores on the pre-test, to find out whether the gender gap increased, decreased or remained the same. The normalised gain is defined as $\langle g \rangle = \frac{\% \langle S_f \rangle - \% \langle S_i \rangle}{100 - \% \langle S_i \rangle}$, in which $\langle S_f \rangle$ and $\langle S_i \rangle$ are the final (post) and initial (pre) class averages, respectively. Then, Norm $g = 100 \langle g \rangle$ in percent [16]. This normalised gain provides valuable information about how much students have learned by taking into account what they already know based on the pre-test. We wanted to investigate whether the normalised gain is higher in one course compared to another, and whether it is the same or different for males and females.

In order to compare EBAE courses with LB courses, we performed *t*-tests [73] on FCI or CSEM pre- and post-test data for males and females. We also calculated the effect size in the form of Cohen's *d* defined as $(\mu_1 - \mu_2)/\sigma_{\text{pooled}}$, where μ_1 and μ_2 are the averages of the two groups being compared (e.g. EBAE versus LB or male versus female) and $\sigma_{\text{pooled}} = \sqrt{(\sigma_1{}^2 + \sigma_2{}^2)/2}$ (here σ_1 and σ_2 are the standard deviations of the two groups being compared). We considered: $d < 0.5$ as small effect size, $0.5 \leqslant d < 0.8$ as medium effect size and $d \geqslant 0.8$ as large effect size, as described in [74].

Table 2. Intra-group CSEM pre-/post-test averages (mean) and standard deviations (SD) for second semester introductory male and female students in calculus-based LB and EBAE courses and algebra-based LB courses. The total number of students in each group, N, is shown. For each group, a p-value obtained using a t-test shows that the difference between the pre-/post-test is statistically significant and the difference between the male and female students is also statistically significant. The normalised gain (Norm g) from pre-test to post-test and the effect size (eff. size) shows how much male and female students learned from what they did not already know based on the pre-test.

Type of class	CSEM	Female	Gender comparison	Male
Calc LB	Pre-test	N: 84 Mean: 34% SD: 13%	p-value: 0.007 ← **gender gap: 4%** → Eff. size: 0.349	N: 234 Mean: 38% SD: 13%
	Pre-test versus post-test comparison	↑ p-value: <0.001 Eff. size: 0.821 Norm g: 18% ↓		↑ p-value: <0.001 Eff. size: 0.894 Norm g: 22% ↓
	Post-test	N: 78 Mean: 45% SD: 16%	p-value: 0.003 ← **gender gap: 6%** → Eff. size: 0.381	N: 248 Mean: 51% SD: 17%
Calc EBAE	Pre-test	N: 112 Mean: 35% SD: 14%	p-value: 0.017 ← **gender gap: 4%** → Eff. size: 0.272	N: 220 Mean: 39% SD: 16%
	Pre-test versus post-test comparison	↑ p-value: <0.001 Eff. size: 1.143 Norm g: 28% ↓		↑ p-value: <0.001 Eff. size: 1.384 Norm g: 39% ↓
	Post-test	N: 98 Mean: 53% SD: 18%	p-value: <0.001 ← **gender gap: 10%** → Eff. size: 0.538	N: 193 Mean: 63% SD: 19%
Alg LB	Pre-test	N: 301 Mean: 22% SD: 8%	p-value: <0.001 ← **gender gap: 6%** → Eff. size: 0.452	N: 201 Mean: 27% SD: 13%
	Pre-test versus post-test comparison	↑ p-value: <0.001 Eff. size: 1.450 Norm g: 23% ↓		↑ p-value: <0.001 Eff. size: 1.325 Norm g: 29% ↓
	Post-test	N: 266 Mean: 40% SD: 16%	p-value: <0.001 ← **gender gap: 8%** → Eff. size: 0.451	N: 172 Mean: 48% SD: 18%

Moreover, although we did not have control over the type of final exam each instructor used in his/her courses, we wanted to look for correlations between the FCI/CSEM post-test performance and the final exam performance for different instructors in the algebra-based and calculus-based EBAE or LB courses for male and female students separately. Including both the algebra-based and calculus-based courses, 10 instructors provided the final exam scores for their classes. We used these data to obtain linear regression plots between the post-test and the final exam performance for males, females and all students (combined) for each instructor and computed the correlation coefficient between the performance of students (male/female/all) on the validated conceptual surveys and their performance on the final exam for different instructors. These correlation coefficients between the conceptual surveys and the final exam (with strong focus on quantitative problem solving) can provide an indication of the strength of the correlation between conceptual and quantitative problem solving of male and female students in these courses.

3. Results

3.1. RQ1. Comparison of the gender gap on the FCI/CSEM pre-test and post-test in LB and EBAE courses

3.1.1. Physics I. In table 1, we present intra-group pre-/post-test data (pooled data for the same type of courses) of male and female students on the FCI for the calculus-based and algebra-based introductory physics I courses. For the algebra-based courses, some were EBAE courses while others were LB courses, whereas all the calculus-based courses were LB. We found statistically significant improvements from the pre-test to the post-test for each group (for both male and female students) in both LB and EBAE courses. However, both female and male students exhibited larger normalised gains in the EBAE courses. In the calculus-based LB course, the gender gap increased slightly from 13% to 17%, whereas in the algebra-based courses, the gender gap stayed roughly the same (varied between 11% and 13%) both in LB and EBAE courses. In both the pre-test and the post-test in calculus-based and algebra-based courses, the difference in performance between male and female students was statistically significant and the effect sizes were typically in the medium range. Thus, it appears that in algebra-based courses, using evidence-based pedagogies helped both female and male students learn more, but did not result in a reduction of the gender gap.

3.1.2. Physics II. In table 2, we present intra-group (pooled data for the same type of courses) pre-/post-test data for male and female students on the CSEM survey for algebra-based and calculus-based introductory physics II courses. Similar to the data shown in table 1, we found statistically significant improvements on the CSEM for female and male students both in LB and EBAE courses, however, the learning gains for both female and male students were larger in EBAE courses. With regards to the gender gap, we found that in LB courses it stayed roughly the same (4% on pre-test and 6% on post-test for calculus-based courses, 6% on the pre-test and 8% on post-test for algebra-based courses). However, in the EBAE calculus-based course, the gender gap increased slightly from 4% to 10%, and it appears that male students may have benefited more from evidence-based pedagogies than female students (normalised gain for male students was 39% in EBAE courses compared to 29% for female students).

Table 3. Between-course comparison of the average FCI pre-/post-test scores of algebra-based male and female students in LB courses with EBAE courses when (i) both courses are taught by the same instructor and (ii) different instructors using similar instructional methods are combined. The *p*-values and effect sizes are obtained for male and female students separately when comparing the LB and EBAE courses in terms of students' FCI scores.

	FCI	Female pre-test	Female post-test	Male pre-test	Male post-test
	LB	*N*: 260 Mean: 30% SD: 13%	*N*: 246 Mean: 43% SD: 18%	*N*: 166 Mean: 42% SD: 19%	*N*: 154 Mean: 56% SD: 21%
(i) FCI Alg: LB versus EBAE (same instructor)	LB versus EBAE Comparison	*p*-value: 0.846 Eff. size: 0.020	*p*-value: 0.002 Eff. size: 0.324	*p*-value: 0.786 Eff. size: 0.034	*p*-value: 0.041 Eff. size: 0.258
	EBAE	*N*: 153 Mean: 30% SD: 15%	*N*: 149 Mean: 49% SD: 18%	*N*: 105 Mean: 43% SD: 20%	*N*: 106 Mean: 61% SD: 19%
	LB	*N*: 456 Mean: 30% SD: 14%	*N*: 383 Mean: 44% SD: 18%	*N*: 318 Mean: 41% SD: 18%	*N*: 255 Mean: 57% SD: 20%
(ii) FCI Alg: LB versus EBAE (different instructors combined)	LB versus EBAE Comparison	*p*-value: 0.861 Eff. size: 0.017	*p*-value: 0.009 Eff. size: 0.255	*p*-value: 0.432 Eff. size: 0.091	*p*-value: 0.088 Eff. size: 0.196
	EBAE	*N*: 153 Mean: 30% SD: 15%	*N*: 149 Mean: 49% SD: 18%	*N*: 105 Mean: 43% SD: 20%	*N*: 106 Mean: 61% SD: 19%

3.2. RQ2. Comparison of the performance on the FCI/CSEM of male and female students in LB and EBAE courses in both the pre-test and the post-test

3.2.1. Physics I.   Table 3 shows the between-course male and female student FCI pre-/post-test score comparison between algebra-based LB and EBAE courses, first holding the instructor fixed (same instructor taught both the LB and EBAE courses, used the same homework and final exams) and second, combining all instructors who used similar methods in the same group (only one instructor used EBAE methods, but several who taught LB courses were combined). Table 3 shows that on the pre-test, the performance of male and female students in the LB courses was similar to the EBAE courses. However, on the post-test both male (female) students in the EBAE courses outperformed male (female) students in the LB courses (effect sizes ranging from 0.196 to 0.324). Coupled with the gender gaps shown in table 1, these data suggest that while both female and male students learned more in EBAE courses, the gender gap remained roughly the same in EBAE courses as in LB courses.

3.2.2. Physics II.   Table 4 shows the between-course CSEM pre-/post-test score comparison between calculus-based LB and EBAE courses, first holding the instructor fixed (same instructor taught both the LB and EBAE courses and used the same homework and final exams) and second, combining all instructors who taught using similar methods into the same group. Table 4 shows that on the pre-test, the performance of male and female students in the LB courses was similar to the EBAE courses. However, on the post-test, both male (female) students in the EBAE courses outperformed male (female) students in the LB courses (effect sizes ranging from 0.299 to 0.623). Interestingly, the effect sizes for male students were slightly higher than the effect sizes for female students, suggesting that male students may have benefited more from evidence-based pedagogies. The gender gap data shown in table 2 can be interpreted in a similar manner. Thus, our data suggest that in calculus-based physics II, while both female and male students learned more in EBAE courses, male students may have benefited more than female students, resulting in a slight increase in the gender gap from pre-test to post-test in EBAE courses.

3.3. RQ3. Comparison between EBAE and LB courses taught by the same instructor in terms of the performance of male and female students (divided according to the pre-test scores)

Tables 5(a) and (b) show the average algebra-based FCI and calculus-based CSEM pre-test, post-test, gain, normalised gain and final exam scores for male and female students, along with the *p*-values between each subgroup of male and female students for pre-test, post-test and the final exam in the EBAE and LB courses taught by the same instructor (with the same homework and final exam) with students divided into three subgroups based on their pre-test scores. The male students were divided into three subgroups according to the pre-test scores of male students only and female students were also divided into three subgroups according to the pre-test scores of female students only. (Tables A1 and A2 in the appendix show similar type of information as tables 5(a) and (b) except that the total number of students was divided into three subgroups according to their pre-test scores regardless of their gender and then male and female students were separated for comparison.)

Table 4. Between-course comparison of the average CSEM pre-/post-test scores of calculus-based male and female students in LB courses with EBAE courses when (i) both courses are taught by the same instructor and (ii) different instructors using similar instructional methods are combined. The *p*-values and effect sizes are obtained for male and female students separately when comparing the LB and EBAE courses in terms of students' CSEM scores.

	CSEM	Female pre-test	Female post-test	Male pre-test	Male post-test
(i) CSEM Calc: LB versus EBAE (same instructor)	LB	*N*: 51 Mean: 35% SD: 12%	*N*: 44 Mean: 44% SD: 15%	*N*: 126 Mean: 42% SD: 12%	*N*: 110 Mean: 50% SD: 16%
	LB versus EBAE comparison	*p*-value: 0.590	*p*-value: 0.119	*p*-value: 0.845	*p*-value: 0.001
		Eff. size: 0.097	Eff. size: 0.299	Eff. size: 0.024	Eff. size: 0.455
	EBAE	*N*: 75 Mean: 36% SD: 14%	*N*: 68 Mean: 48% SD: 17%	*N*: 133 Mean: 42% SD: 15%	*N*: 113 Mean: 58% SD: 20%
(ii) CSEM Calc LB versus EBAE (different instructors combined)	LB	*N*: 84 Mean: 34% SD: 13%	*N*: 78 Mean: 45% SD: 16%	*N*: 234 Mean: 38% SD: 13%	*N*: 248 Mean: 51% SD: 17%
	LB versus EBAE comparison	*p*-value: 0.595	*p*-value: 0.003	*p*-value: 0.679	*p*-value: <0.001
		Eff. size: 0.077	Eff. size: 0.448	Eff. size: 0.039	Eff. size: 0.623
	EBAE	*N*: 112 Mean: 35% SD: 14%	*N*: 98 Mean: 53% SD: 18%	*N*: 220 Mean: 39% SD: 16%	*N*: 193 Mean: 63% SD: 19%

Table 5. Average FCI (a) and CSEM (b) pre-test scores (Pre-test), post-test scores (Post-test), gain (Gain), normalised gain (Norm *g*) and final exam scores (Final) for male and female students in the EBAE and LB courses taught by the same instructor (with same homework and final exam). Male students were divided into three groups based upon their pre-test scores and female students were also divided into three groups based upon their pre-test scores separately. For each division (subgroup), a *p*-value was obtained using a *t*-test that shows whether there is statistically significant difference between male and female students on pre-test, post-test and final exam.

	Pre-test split		Pre-test	Post-test	Gain	Norm *g*	Final
(a)	Bottom 1/3	Mean female score	17	35	18	21	47
		p-value	0.010	0.078			0.815
		Mean male score	23	42	19	25	48
FCI Alg LB	Middle 1/3	Mean female score	28	42	14	19	53
		p-value	<0.001	0.012			0.078
		Mean male score	39	55	16	26	60
	Top 1/3	Mean female score	44	57	13	23	60
		p-value	<0.001	0.002			0.422
		Mean male score	65	74	8	24	64
	Bottom 1/3	Mean female score	15	39	23	27	54
		p-value	<0.001	0.164			0.318
		Mean male score	21	43	22	28	50
FCI Alg EBAE	Middle 1/3	Mean female score	28	46	17	24	52
		p-value	<0.001	<0.001			0.008
		Mean male score	38	58	21	33	61
	Top 1/3	Mean female score	49	63	14	28	60
		p-value	<0.001	<0.001			0.011
		Mean male score	63	78	15	41	68
(b)	Bottom 1/3	Mean female score	24	36	12	16	42
		p-value	0.006	0.738			0.209
		Mean male score	28	38	10	13	47
CSEM Calc LB	Middle 1/3	Mean female score	34	44	10	16	52
		p-value	<0.001	0.661			0.770
		Mean male score	40	46	6	10	50
	Top 1/3	Mean female score	49	56	8	15	53
		p-value	0.060	0.513			0.018
		Mean male score	54	59	5	12	62
	Bottom 1/3	Mean female score	22	36	14	18	48
		p-value	0.002	0.050			0.196
		Mean male score	27	44	18	24	53
CSEM Calc EBAE	Middle 1/3	Mean female score	34	48	14	22	52
		p-value	<0.001	0.455			0.104
		Mean male score	41	51	10	18	58
	Top 1/3	Mean female score	56	63	7	16	68
		p-value	0.382	0.003			0.621
		Mean male score	58	74	16	38	70

Table 6. Average FCI (a) and CSEM (b) pre-test scores (Pre-test), post-test scores (Post-test), gain (Gain) and normalised gain (Norm *g*) for male and female students in the EBAE and LB algebra-based and calculus-based courses. All courses in the same group were combined. Male students were divided into three groups based upon their pre-test scores and female students were also divided into three groups based upon their pre-test scores separately. For each division (subgroup), a *p*-value was obtained using a *t*-test that shows whether there is statistically significant difference between male and female students on pre-test and post-test. *Note that FCI data for calculus-based EBAE classes are not available.*

	Pre-test split		Pre-test	Post-test	Gain	Norm *g*
(a)	Bottom 1/3	Mean female score	24	39	15	19
		p-value	<0.001	<0.001		
		Mean male score	35	52	17	26
FCI Calc LB	Middle 1/3	Mean female score	43	52	9	16
		p-value	<0.001	<0.001		
		Mean male score	60	75	14	36
	Top 1/3	Mean female score	67	72	6	17
		p-value	<0.001	0.023		
		Mean male score	85	82	−3	−18
	Bottom 1/3	Mean female score	15	39	23	27
		p-value	<0.001	0.164		
		Mean male score	21	43	22	28
FCI Alg EBAE	Middle 1/3	Mean female score	28	46	17	24
		p-value	<0.001	<0.001		
		Mean male score	38	58	21	33
	Top 1/3	Mean female score	49	63	14	28
		p-value	<0.001	<0.001		
		Mean male score	63	78	15	41
	Bottom 1/3	Mean female score	17	33	15	18
		p-value	<0.001	<0.001		
		Mean male score	23	41	18	23
FCI Alg LB	Middle 1/3	Mean female score	29	44	15	22
		p-value	<0.001	<0.001		
		Mean male score	40	55	15	24
	Top 1/3	Mean female score	44	57	13	23
		p-value	<0.001	<0.001		
		Mean male score	62	76	14	36
(b)	Bottom 1/3	Mean female score	21	43	22	28
		p-value	0.044	<0.001		
		Mean male score	23	57	35	45
CSEM Calc EBAE	Middle 1/3	Mean female score	32	52	20	29
		p-value	<0.001	0.403		
		Mean male score	37	55	18	29
	Top 1/3	Mean female score	51	67	15	31
		p-value	0.039	0.017		
		Mean male score	56	74	18	41

Table 6. (Continued.)

	Pre-test split		Pre-test	Post-test	Gain	Norm *g*
CSEM Calc LB	Bottom 1/3	Mean female score	21	37	16	20
		p-value	0.001	0.429		
		Mean male score	25	40	15	20
	Middle 1/3	Mean female score	32	43	12	17
		p-value	<0.001	0.030		
		Mean male score	37	49	12	20
	Top 1/3	Mean female score	47	60	13	24
		p-value	0.025	0.744		
		Mean male score	52	57	6	12
CSEM Alg LB	Bottom 1/3	Mean female score	15	34	20	23
		p-value	0.017	0.100		
		Mean male score	16	39	23	27
	Middle 1/3	Mean female score	22	40	19	24
		p-value	<0.001	0.009		
		Mean male score	25	47	22	30
	Top 1/3	Mean female score	31	47	16	23
		p-value	<0.001	<0.001		
		Mean male score	40	59	19	31

3.3.1. Algebra-based Physics I. The data in table 5(a) show that in both the LB and EBAE algebra-based courses, on the pre-test, there was a gender gap on the FCI between male and female students in each group (bottom 1/3, middle 1/3, top 1/3) and since female and male students had comparable gains, the gender gap was maintained in most cases on the post-test. This is consistent with the data shown in table 1 which indicate that in both LB and EBAE courses, when including all students, the gender gap on the FCI stayed roughly the same from pre-test to post-test. When comparing the LB with the EBAE course, we see that both female and male students seemed to benefit equally (gains and normalised gains on the FCI were higher in the EBAE course compared to the LB course), with the exception of the top 1/3 of the students. The top 1/3 of the female students had similar FCI gains in the LB and EBAE course (13% and 14%, respectively), whereas the top 1/3 of the male students showed larger FCI gains in the EBAE course.

On the final exam, the data in table 5(a) suggest that in the EBAE class, the top and middle 1/3 of the male students performed better than the top and middle 1/3 of the female students. Similar, although not as strong, trends can be seen in the LB course. Since students in these two courses took the same final exams, these data suggest that the top and middle 1/3 of the male students benefited slightly more from EBAE pedagogies than top and middle 1/3 of the female students. On the other hand, the bottom 1/3 of the female students benefited more from EBAE pedagogies than the bottom 1/3 of the male students (performance of bottom 1/3 of female students is 54% in the EBAE course and only 47% in the LB course, whereas the performance of the bottom 1/3 of male students is 50% in the EBAE course and 48% in the LB course).

3.3.2. Calculus-based Physics II. The data in table 5(b) indicate that in the LB course, there was a gender gap on the CSEM in the pre-test, but on the post-test, this gender gap decreased and was no longer statistically significant. This result appears to be inconsistent with the data shown in table 2 which indicate that in the LB courses, the CSEM gender gap remained roughly the same (or slightly increased). However, the data in table 2 includes all LB courses, whereas the data in table 5(b) includes only one course which was taught by the same instructor who taught the EBAE course shown in table 5(b). So in this particular calculus-based LB course, the gender gap on the CSEM decreased slightly, but if we include all calculus-based LB courses, the gender gap was roughly the same (or increased slightly).

For the calculus-based EBAE course, the data in table 5(b) suggest that the gender gap increased. The gender gaps for bottom 1/3, middle 1/3 and top 1/3 of the students are 5%, 7% and 2% in the pre-test but 8%, 3% and 11% in the post-test, respectively. Thus, with the exception of the middle 1/3 of the students, the gender gap on the CSEM increased. This suggests that the bottom 1/3 and top 1/3 of the male students may have benefited more from EBAE pedagogies compared to the respective female students. It appears that this was indeed the case when we compare the normalised gains in the LB and EBAE courses: for the bottom 1/3 and top 1/3 of the male students, their CSEM normalised gains were 13% and 12% in the LB course, but 24% and 38% in the EBAE course, respectively. For the bottom 1/3 and top 1/3 of the female students, their CSEM normalised gains were 16% and 15% in the LB course, and 18% and 16% in the EBAE course. On the final exam, in both the LB and EBAE course, it appears that male students performed slightly better than the female students. However, only the 9% gender gap between the top 1/3 of the male and female students in the LB course is statistically significant.

3.4. RQ4. Comparison between EBAE and LB courses taught by different instructors in terms of the performance of male and female students *(divided according to the pre-test scores)*

Tables 6(a) and (b) show the average algebra-based and calculus-based FCI and CSEM pre-test score, post-test score, gain and normalised gain for male and female students, along with the p-values between each subgroup of male and female students for the pre-test and post-test in the EBAE and LB courses. All equivalent (algebra-based or calculus-based physics I or II) courses which used the same instructional strategy (EBAE or LB) were combined and students were divided into three groups based upon their pre-test scores. Male students were divided into three subgroups according to the pre-test scores of male students only and female students were also divided into three subgroups according to the pre-test scores of female students only, and their scores were compared (cases in which male and female scores are significantly different have been highlighted). We note that tables A3 and A4 in the appendix show the same data except that the students were divided into three subgroups according to their pre-test scores regardless of their gender. Then, male and female students were separated for comparison and the cases in which male and female scores are significantly different from each other have been highlighted. In these tables 6(a), (b), A3 and A4, the average final exam performance is not listed because different instructors used different exams which varied in difficulty.

3.4.1. Physics I. For the calculus-based LB course, the data in table 6(a) suggest that the gender gap on the FCI increased from pre-test to post-test for each ability level. The gender

gap for the bottom, middle, and top $1/3$ of the students was 11%, 17%, 18% in the pre-test, but in the post-test, it was 13%, 23%, 10%, respectively. This is consistent with the data in table 1 which indicates that, including all students, the FCI gender gap increased slightly from the pre-test to the post-test except the top $1/3$ group.

For the algebra-based LB and EBAE courses, the data in table 6(a) suggest that the gender gap on the FCI was present at each ability level in the pre-test and it remained roughly the same in the post-test (consistent with the data in table 1). For the LB courses, the gender gap on the FCI for bottom, middle, top $1/3$ of the students was 6%, 11%, 18% on the pre-test and 7%, 11%, 19% on the post-test. For the EBAE courses, the gender gap for bottom, middle, top $1/3$ of the students was 6%, 10%, 14% on the pre-test and 4%, 12%, 15% on the post-test. Interestingly, in both type of courses, it appears that the gender gap on the FCI was more pronounced at higher ability levels (based on FCI pre-test scores). This was especially true in the LB course where the performance of the top $1/3$ of the male students was on the average 18% (19%) higher compared to the top $1/3$ of the female students on the pre-test (post-test). The data in table 6(a) suggest that both female and male students learned more in the EBAE course compared to the LB, but the learning gains were not much larger in the EBAE course compared to the LB course.

3.4.2. Physics II. The data in table 6(b) suggest that in the calculus-based EBAE courses, the gender gap on the CSEM increased, but only for the bottom $1/3$ of the students. On the pre-test, the gender gap between the bottom $1/3$ of the male and female students was 2%, whereas in the post-test, the gender gap was 14%. This suggests that the bottom $1/3$ male students benefited much more from EBAE pedagogies than the bottom $1/3$ of the female students. For the middle and top $1/3$ of the students, the gender gap remained roughly the same. By comparison, in the calculus-based LB courses, the gender gap stayed roughly the same. The gender gap between bottom $1/3$, middle $1/3$, top $1/3$ of the students was 4%, 5%, 5% in the pre-test and 3%, 6%, -3% in the post-test. These findings are consistent with the data shown in table 2, which indicate that the gender gap on the CSEM stayed roughly the same in the LB courses, but increased slightly in the EBAE courses.

Comparing the LB with the EBAE courses in terms of normalised gain, the data in table 6(b) suggest that students at all levels benefit from EBAE pedagogies. However, it appears that the bottom $1/3$ and top $1/3$ of the male students benefited more from EBAE pedagogies compared to the corresponding female students. The normalised CSEM gains for the bottom $1/3$ and top $1/3$ of the male students were 20% and 12% in the LB course but the EBAE course they were much larger at 45% and 41%. For the bottom $1/3$ and top $1/3$ of the female students on the other hand, the normalised CSEM gains were 20% and 24% in the LB course but only slightly larger at 28% and 31% in the EBAE course. A very similar trend was observed in table 5(b). Thus, it appears that for calculus-based physics II, the bottom $1/3$ and top $1/3$ of the male students may have benefited more from EBAE pedagogies compared to the bottom $1/3$ and top $1/3$ of the female students.

Similar to the calculus-based LB courses, for the algebra-based LB courses, the gender gap stayed roughly the same for students in the bottom, middle and top $1/3$ of the class (based on CSEM pre-test scores).
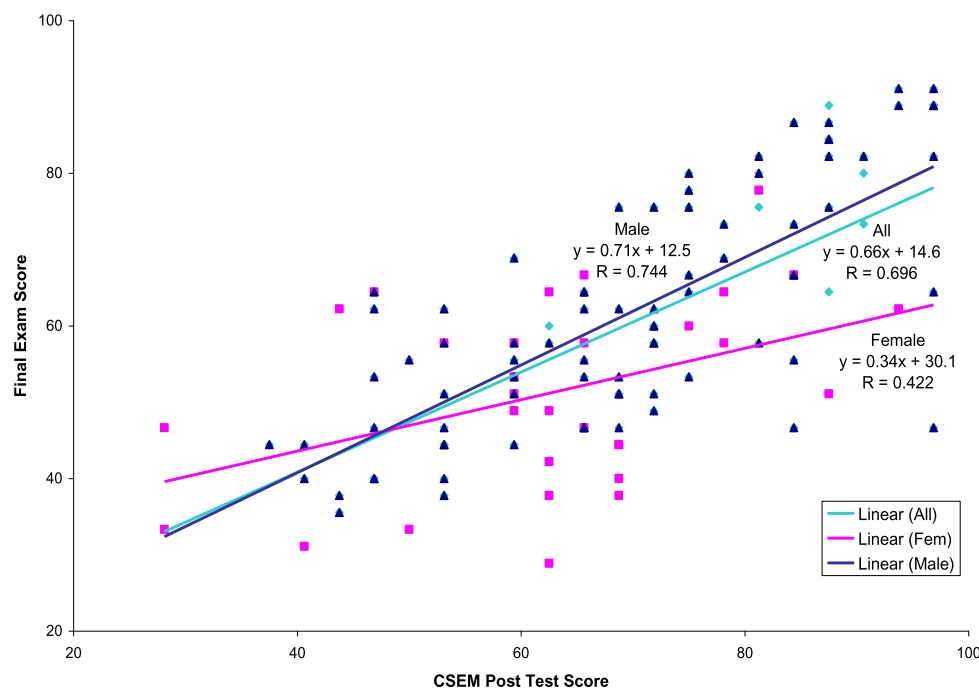
Figure 1. Linear regression and correlation coefficients of the CSEM post-test scores (conceptual) and final exam scores (heavy focus on quantitative problem solving) for male students, female students and all students (males and females combined together) for Inst 1 (EBAE instructor) for calculus-based introductory physics courses. The correlation coefficients for other FCI/CSEM instructors have been summarised in table 7.

3.5. RQ5. Correlation between CSEM post-test and final exam scores for male and female students

Figure 1 plots the CSEM post-test performance along with the final exam performance for male, female and all students in one particular calculus-based EBAE course. Figure 1 shows that the linear regressions [73] and there are moderate to strong correlation between post-test and the final exam scores. We also plotted linear regressions for the other courses and the data look similar to figure 1 but are not included here. Instead, we include all the correlation coefficients (CSEM post-test versus final exam) for all the courses for which we were able to obtain both the post-test and final exam data. Table 7 summarises the correlation coefficients between CSEM/FCI post-test and final exam scores for each instructor who provided final exam data.

Despite the fact that different instructors had different final exams and at least some of the content in the final exam does not match the FCI or CSEM tests (e.g. in physics I, many topics were covered which are not on the FCI, such as momentum and collisions, static equilibrium and rotations, fluid dynamics and others), the correlation coefficients including males and females range from 0.415 to 0.816, which are considered to be moderate to high

Table 7. Correlation coefficients (R) between CSEM/FCI post-test and final exam scores of male and female students for each instructor (Inst) who provided final exam data. The final exam data were not provided by physics II instructors in algebra-based courses.

Physics I (Calc)			Physics I (Alg)			Physics II (Calc)		
Inst and course	Female	Male	Inst and course	Female	Male	Inst and course	Female	Male
Inst 1: LB	0.427	0.525	Inst 1: EBAE	0.568	0.628	Inst 1: EBAE	0.422	0.744
Inst 2: LB	0.674	0.573	Inst 1: LB	0.510	0.554	Inst 2: EBAE	0.571	0.415
Inst 3: LB	0.774	0.816	Inst 2: LB	0.626	0.728	Inst 2: LB	0.553	0.540
						Inst 3: LB	0.230	0.592

correlations. Although there are some differences between the correlation coefficients for male and female students in a given course, there is no clear discernable trend.

4. Discussion and summary

4.1. General findings for EBAE versus LB courses without consideration of gender

In all cases investigated, we find that on average, introductory students in the courses which made significant use of EBAE methods outperformed those in courses primarily taught using LB instruction on standardised conceptual surveys (FCI or CSEM) on the post-test even though there was no statistically significant difference on the pre-test. This was true both in the algebra-based and calculus-based physics I (primarily mechanics) and 2 (primarily E&M) courses. Also, the differences between EBAE and LB courses were observed both among students who performed well on the FCI/CSEM pre-test (given in the first week of classes) and also those who performed poorly, thus indicating that EBAE instructional strategies helped students at all levels. However, the typical effect sizes for the differences between equivalent EBAE and LB courses was between 0.23 and 0.49, which are small. Thus, the benefits of these EBAE approaches were not as large as one might expect to observe. There are many potential challenges to using EBAE instructional strategies effectively, including but not limited to:

- Content coverage. There is often a lot of content covered in introductory physics courses and it is challenging to cover the same amount of content while also including frequent active-learning activities during class and students are expected to take responsibility for learning some those things outside of classes.
- Lack of student buy-in of EBAE pedagogies, which may result in lack of appropriate engagement with self-paced learning tools outside of class. It is therefore important for instructors to frame the course for students and discuss the various instructional approaches that will be used in a course and why they are expected to be beneficial for student learning. Providing data to the students that support the use of evidence-based active-learning strategies [47] can be helpful, and when possible, including explicit

discussions connecting students' and instructors' goals for taking the course can also be beneficial [75].

- Lack of student engagement with in-class active-learning activities (e.g. clicker questions and group problem solving). Students may not recognise on their own that they will learn best if they engage with the in-class activities to the best of their ability. Therefore, ensuring that in-class activities help all students learn is important. Furthermore, since peer collaboration is exploited in many EBAE classes to enhance student learning, ensuring that these activities are designed and incentivized in a manner that not only fosters positive inter-dependence (success of one student is contingent on the success of the group) but also individual accountability (students are expected to show that they learned from working in a group) is essential [4, 5].
- Large class sizes may be an impediment. One approach faculty used in flipped courses was to split the class in two (the instructor met with each group for only half of the time as compared to an LB course, his/her total contact hours with students remained the same), thus forming smaller class sizes. But even if the class size goes from 200 to 100 students by this process of breaking the class into two halves, it may still be challenging to manage the in-class activities effectively. Undergraduate or graduate teaching assistants need to be trained to effectively help in facilitating in-class activities. In group activities, students often work at different rates, so effective approaches need to be adopted to ensure that those who finish early can help others.

4.2. Impact on gender gap

We found that the EBAE courses did not result in reducing the gender gap. For algebra-based courses, students at all levels learned more in the EBAE courses; however, it appears that both female and male students benefited from evidence-based pedagogies equally and the gender gap present in the pre-test was also found on the post-test. For calculus-based courses, our data suggest that male students actually benefited more from evidence-based pedagogies, which resulted in an increase in the gender gap from the pre-test to the post-test. One hypothesis for why the gender gap was steady in algebra-based courses but grew in the calculus-based courses is that in the calculus-based courses there are significantly fewer women which can impact their sense of belonging and self-efficacy. These issues were not investigated in this study.

Previous research has also found that sometimes evidence-based pedagogies result in a reduction of the gender gap [53, 54], while in other cases they do not [55]. The reasons for the gender gap even in the pre-test are complex and some have attributed the persistence of gender gap to issues such as societal gender stereotypes, ST, high anxiety classes, lack of social belonging for women in physics classes, the culture promoting fixed intelligence mindset (with men having the innate ability to excel in subjects such as physics), and low self-efficacy [59–66].

Some have suggested that the gender gap found on conceptual assessments may at least in part be due to ST [57, 60–63], and the extent to which the classroom environment is perceived as threatening by female students which in turn can depend on the instructor and the instructional design. For example, as discussed earlier, research suggests that, for high school female students, taking a physics test can create an 'implicit ST' and can degrade their performance [60]. Such threat may be present even when taking the FCI or CSEM test at the beginning of the semester as a pre-test and lead to a gender gap in performance. Also, female

students may have a lower sense of social belonging and low self-efficacy in a physics class due to societal stereotypes about who belongs in physics and who is capable of doing physics.

Some have suggested that classes which are not only collaborative but also emphasise collaboration and reduce competition (e.g. by not grading on a curve) are likely to be perceived more positively by female students and may partly be responsible for the reduced gender gap in [53]. Even EBAE classrooms can be characterised as high or low anxiety classes depending on the extent to which the instruction was designed to be inclusive and whether it explicitly focused on promoting a sense of belongingness, self-efficacy and growth mindset for all students. The extent to which the instructor plays an encouraging role to promote these positive motivational factors, and emphasises that he/she is there as a guide to help all students succeed and also emphasises that struggling is a stepping stone to success and should be viewed positively may also play a role in dispelling the negative impacts of societal gender stereotypes about physics that accumulate over a female student's lifetime. Since fixed mindset about innate intelligence can be a factor for the poor performance of female students, instructors should take advantage of research findings about the importance of promoting a growth mindset [66] in their physics classes. In particular, research shows that students who believe that the brain is like a muscle, and intelligence is malleable and can increase with effort are more likely to persevere and perform better than those who think that intelligence is fixed [66]. Moreover, research suggests that mindset can be changed with a very short intervention [66]. Since according to the national data [76], fewer female students are likely to have taken challenging high school physics courses (e.g. Advanced Placement) before taking the college-level course, college EBAE courses which do not explicitly take into account these motivational factors may unknowingly create a high anxiety classroom environment for students who have less prior knowledge (who are more likely to be female students). For example, if students work in small groups in an EBAE course and some students in the group 'show off' their knowledge and the instructional design does not promote a growth mindset, or the importance of hard work and persistence in learning physics, students who have taken less challenging physics course may have their self-efficacy issues exacerbated as opposed to reduced. Therefore, these motivational issues should be addressed in all physics classes as part of the instructional design to create inclusive classroom environment, as suggested in [77].

In summary, in order to enhance student learning in EBAE courses, it important not only to develop effective EBAE learning tools and pedagogies commensurate with students' prior knowledge but also to investigate how to implement them appropriately and how to motivate and incentivize their usage to get buy-in from students in order for them to engage with them as intended. Furthermore, reducing the gender gap on conceptual assessments is a challenging endeavor and evidence-based pedagogies may not be sufficient. In order to reduce gender gap, it may be useful to pay attention to other factors, e.g. improving the sense of belonging and self-efficacy of female students, improving their intelligence mindset (so that they do not think of male students as having an innate ability to excel in physics that they do not have and view intelligence as something that is malleable and can be cultivated by focus and effort), and reducing competition and emphasising collaboration.

Acknowledgments

Appendix

Table A1. Average FCI pre-test scores (Pre-test), post-test scores (Post-test), gain (Gain), normalised gain (Norm g) and final exam scores (Final) for male and female students in the flipped and LB courses taught by the same instructor (with same homework and final exam) with students divided into three groups regardless of their gender based on their pre-test scores. For each division (subgroup), a p-value was obtained using a t-test that shows whether there is statistically significant difference between male and female students on pre-test, post-test or final exam.

	Pre-test split		Pre-test	Post-test	Gain	Norm g	Final
FCI Alg LB	Bottom 1/3	Mean female score	19	35	17	20	48
		p-value	0.192	0.171			0.082
		Mean male score	15	40	25	30	40
	Middle 1/3	Mean female score	32	46	13	20	53
		p-value	0.967	0.955			0.772
		Mean male score	32	45	13	19	54
	Top 1/3	Mean female score	49	62	13	26	64
		p-value	0.046	0.121			0.934
		Mean male score	56	70	13	31	64
FCI Alg EBAE	Bottom 1/3	Mean female score	16	40	24	28	54
		p-value	0.149	0.610			0.438
		Mean male score	18	42	24	29	51
	Middle 1/3	Mean female score	31	45	14	21	52
		p-value	0.038	0.013			0.159
		Mean male score	33	54	21	32	57
	Top 1/3	Mean female score	53	71	18	38	63
		p-value	0.091	0.209			0.212
		Mean male score	58	75	17	40	67

26

Table A2. Average CSEM pre-test scores (Pre-test), post-test scores (Post-test), gain (Gain), normalised gain (Norm g) and final exam scores (Final) for male and female students in the flipped and LB courses taught by the same instructor (with same homework and final exam) with students divided into three groups regardless of their gender based on their pre-test scores. For each division (subgroup), a p-value was obtained using a t-test that shows whether there is statistically significant difference between male and female students on pre-test, post-test and final exam.

	Pre-test split		Pre-test	Post-test	Gain	Norm g	Final
	Bottom 1/3	Mean female score	27	37	11	15	44
		p-value	0.945	0.505			0.921
		Mean male score	26	35	8	11	44
CSEM Calc LB	Middle 1/3	Mean female score	39	49	10	16	52
		p-value	0.701	0.298			0.688
		Mean male score	39	45	7	11	53
	Top 1/3	Mean female score	55	62	7	15	57
		p-value	0.507	0.613			0.578
		Mean male score	53	59	6	13	59
	Bottom 1/3	Mean female score	25	41	16	21	50
		p-value	0.542	0.480			0.436
		Mean male score	24	44	20	26	53
CSEM Calc EBAE	Middle 1/3	Mean female score	38	49	11	18	52
		p-value	0.433	0.909			0.288
		Mean male score	39	49	11	17	57
	Top 1/3	Mean female score	57	63	6	14	69
		p-value	0.949	0.015			0.869
		Mean male score	57	73	16	36	69

Table A3. Average FCI pre-test scores (Pre-test), post-test scores (Post-test), gain (Gain) and normalised gain (Norm g) for male and female students in the flipped and LB algebra-based and calculus-based courses. All courses in the same group were combined with students divided into three groups regardless of their gender based upon their pre-test scores. For each division (subgroup), a p-value was obtained using a t-test that shows whether there is statistically significant difference between male and female students on pre-test and post-test. *Note that FCI data for Calculus-based EBAE classes are not available.*

	Pre-test split		Pre-test	Post-test	Gain	Norm g
FCI Calc LB	Bottom 1/3	Mean female score	30	43	13	19
		p-value	0.083	0.016		
		Mean male score	32	49	17	25
	Middle 1/3	Mean female score	53	60	7	14
		p-value	0.049	<0.001		
		Mean male score	55	71	16	36
	top 1/3	Mean female score	83	84	1	7
		p-value	0.918	0.707		
		Mean male score	82	82	−1	−3
FCI Alg EBAE	Bottom 1/3	Mean female score	16	40	24	28
		p-value	0.149	0.610		
		Mean male score	18	42	24	29
	Middle 1/3	Mean female score	31	45	14	21
		p-value	0.038	0.013		
		Mean male score	33	54	21	32
	Top 1/3	Mean female score	53	71	18	38
		p-value	0.091	0.209		
		Mean male score	58	75	17	40
FCI Alg LB	Bottom 1/3	Mean female score	19	34	15	19
		p-value	0.467	0.427		
		Mean male score	20	36	16	20
	Middle 1/3	Mean female score	32	46	13	20
		p-value	0.617	0.501		
		Mean male score	33	47	15	22
	Top 1/3	Mean female score	50	64	14	27
		p-value	0.002	0.005		
		Mean male score	56	71	15	34

Table A4. Average CSEM pre-test scores (Pre-test), post-test scores (Post-test), gain (Gain) and normalised gain (Norm *g*) for male and female students in the flipped and LB algebra-based and calculus-based courses. All courses in the same group were combined with students divided into three groups regardless of their gender based upon their pre-test scores. For each division (subgroup), a *p*-value was obtained using a *t*-test that shows whether there is statistically significant difference between male and female students on pre-test and post-test. *Note that CSEM data for algebra-based EBAE classes are not available*.

	Pre-test split		Pre-test	Post-test	Gain	Norm *g*
CSEM Calc EBAE	Bottom 1/3	Mean female score	23	44	22	28
		p-value	0.571	0.001		
		Mean male score	22	57	35	45
	Middle 1/3	Mean female score	35	56	22	33
		p-value	0.771	0.891		
		Mean male score	35	56	21	32
	Top 1/3	Mean female score	54	65	10	23
		p-value	0.718	0.033		
		Mean male score	55	72	17	38
CSEM Calc LB	Bottom 1/3	Mean female score	23	39	16	20
		p-value	0.367	0.827		
		Mean male score	24	39	15	20
	Middle 1/3	Mean female score	35	46	11	17
		p-value	0.833	0.535		
		Mean male score	35	48	13	20
	Top 1/3	Mean female score	51	61	10	21
		p-value	0.966	0.460		
		Mean male score	51	58	7	14
CSEM Alg LB	Bottom 1/3	Mean female score	15	36	21	25
		p-value	0.402	0.576		
		Mean male score	15	38	23	27
	Middle 1/3	Mean female score	23	42	20	25
		p-value	0.981	0.192		
		Mean male score	23	46	23	30
	Top 1/3	Mean female score	33	45	13	19
		p-value	0.002	<0.001		
		Mean male score	37	56	19	30

ORCID iDs

Alexandru Maries ⬤ https://orcid.org/0000-0001-5781-7343
Chandralekha Singh ⬤ https://orcid.org/0000-0002-1234-5458

References

[1] McDermott L C 1991 Millikan lecture 1990: what we teach and what is learned-closing the gap *Am. J. Phys.* **59** 301
[2] Fraser J, Timan A, Miller K, Dowd J, Tucker L and Mazur E 2014 Teaching and physics education research: bridging the gap *Rep. Prog. Phys.* **77** 032401
[3] Docktor J and Mestre J 2014 Synthesis of discipline-based education research in physics *Phys. Rev. ST Phys. Educ. Res.* **10** 020119
[4] Heller P, Keith R and Anderson S 1992 Teaching problem solving through cooperative grouping: 1. Group versus individual problem solving *Am. J. Phys.* **60** 627
[5] Heller P and Hollabaugh M 1992 Teaching problem solving through cooperative grouping: 2. Designing problems and structuring groups *Am. J. Phys.* **60** 637
[6] Hammer D 2000 Student resources for learning introductory physics *Am. J. Phys. Phys. Educ. Res. Suppl.* **68** S52
[7] Redish E, Steinberg R and Saul J 1998 Student expectations in introductory physics *Am. J. Phys.* **66** 212
[8] Hake R 1998 Interactive engagement versus traditional methods: a six-thousand student survey of mechanics test data for introductory physics courses *Am. J. Phys.* **66** 64
[9] Singh C 2014 What can we learn from PER: physics education research? *Phys. Teach.* **52** 568
[10] Singh C 2002 When physical intuition fails *Am. J. Phys.* **70** 1103
[11] Reif F 1995 Millikan lecture 1994: understanding and teaching important scientific thought processes *Am. J. Phys.* **63** 17
[12] Lin S Y and Singh C 2015 Effect of scaffolding on helping introductory physics students solve quantitative problems involving strong alternative conceptions *Phys. Rev. ST PER* **11** 020105
[13] Lin S Y and Singh C 2013 Using isomorphic problem pair to learn introductory physics: transferring from a two-step problem to a three-step problem *Phys. Rev. ST PER* **9** 020114
[14] Lin S Y and Singh C 2011 Using isomorphic problems to learn introductory physics *Phys. Rev. ST PER* **7** 020104
[15] Mason A and Singh C 2011 Assessing expertise in introductory physics using categorization task *Phys. Rev. ST Phys. Educ. Res* **7** 020110
  Mason A and Singh C 2016 Using categorization of problems as an instructional tool to help introductory students learn physics *Phys. Educ.* **51** 025009
[16] Singh C 2009 Categorization of problems to assess and improve proficiency as teacher and learner *Am. J. Phys.* **77** 73
[17] Singh C 2008 Assessing student expertise in introductory physics with isomorphic problems: I. Performance on a nonintuitive problem pair from introductory physics *Phys. Rev. ST Phys. Educ. Res.* **4** 010104
[18] Singh C 2008 Assessing student expertise in introductory physics with isomorphic problems: II. Effect of some potential factors on problem solving and transfer *Phys. Rev. ST PER* **4** 010105
[19] Collins A, Brown J and Newman S 1989 Cognitive apprenticeship: teaching the crafts of reading, writing and mathematics *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser* ed L B Resnick (Hillsdale, NJ: Lawrence Erlbaum) p 453
[20] Crouch C H and Mazur E 2001 Peer instruction: ten years of experience and results *Am. J. Phys.* **69** 970
[21] Mason A and Singh C 2010 Helping students learn effective problem solving strategies by reflecting with peers *Am. J. Phys.* **78** 748
[22] Mason A and Singh C 2016 Impact of guided reflection with peers on the development of effective problem solving strategies and physics learning *Phys. Teach.* **54** 295
[23] McDermott L, Shaffer P and (the Physics Education Group, University of Washington) 2002 *Tutorials in Introductory Physics* (NJ: Prentice-Hall)
[24] Singh C 2008 Interactive learning tutorials on quantum mechanics *Am. J. Phys.* **76** 400

[25] Marshman E and Singh C 2016 Interactive tutorial to improve student understanding of single photon experiments involving a Mach–Zehnder Interferometer *Eur. J. Phys.* **37** 024001

[26] Kalman C, Milner-Bolotin M and Antimirova T 2010 Comparison of the effectiveness of collaborative groups and peer instruction in a large introductory physics course for science majors *Can. J. Phys.* **88** 325

[27] Singh C 2005 Impact of peer interaction on conceptual test performance *Am. J. Phys.* **73** 446

[28] Singh C 2002 Effectiveness of group interaction on conceptual standardized test performance, *Proc. 2002 Physics Education Research Conf. (Boise)* p 67 (http://doi.org/10.1119/perc.2002.pr.017)

[29] Sayer R, Marshman E and Singh C 2016 The impact of peer interaction on the responses to clicker questions in an upper-level quantum mechanics course *Proc. 2016 Physics Education Research Conf. (Sacramento, CA)* p 304 (http://doi.org/10.1103/PhysRevPhysEducRes.12.020133)

[30] Black P and Wiliam D 1998 Assessment and classroom learning *Assess. Educ.* **5** 7

[31] Moll R and Milner-Bolotin M 2009 The effect of interactive lecture experiments on student academic achievement and attitudes towards physics *Can. J. Phys.* **87** 917

[32] Novak G, Patterson E T, Gavrin A and Christian W 1999 *Just-in-Time Teaching: Blending Active Learning with Web Technology* (Upper Saddle River, NJ: Prentice-Hall)

[33] Sayer R, Marshman E and Singh C 2016 A case study evaluating just-in-time teaching and Peer instruction using clickers in a quantum mechanics course *Phys. Rev. PER* **12** 020133

[34] https://perusall.com/

[35] Schwartz D and Bransford J 1998 A time for telling *Cognit. Instruct.* **16** 475

[36] Mayer R 2001 *Multimedia Learning* (Cambridge: Cambridge Press)

[37] Chen Z, Stelzer T and Gladding G 2010 Using multi-media modules to better prepare students for introductory physics lecture *Phys. Rev. Spec. Top. Phys. Educ. Res.* **6** 010108

[38] Reif F and Scott L 1999 Teaching scientific thinking skills: students and computers coaching each other *Am. J. Phys.* **67** 819

[39] Schroader N, Gladding G, Guttman B and Stelzer T 2015 Narrated animated solution videos in a mastery setting *Phys. Rev. ST PER* **11** 010103

[40] Singh C 2004 Interactive video tutorials for enhancing problem-solving, reasoning, and meta-cognitive skills of introductory physics students *AIP Conf. Proc.* **720** 177

[41] Singh C and Haileselassie D 2010 Developing problem solving skills of students taking introductory physics via web-based tutorials *J. Coll. Sci. Teach.* **39** 42

[42] Bishop J L and Verleger M A 2013 The flipped classroom: a survey of the research 2013 *ASEE Annual Conf. and Exposition*

[43] Berrett D 2012 How 'flipping' the classroom can improve the traditional lecture *The Chronicle of Higher Education* February 19

[44] Mazur E 1997 *Peer Instruction: A User's Manual* (Engelwood Cliffs: Prentice-Hall)

[45] Haak D, Hille RisLambers J, Pitre E and Freeman S 2011 Increased structure and active learning reduce the achievement gap in introductory biology *Science* **332** 1213

[46] Breslow L, Pritchard D, DeBoer J, Stump G, Ho A and Seaton V 2013 Studying learning in the worldwide classroom: research into edX's first MOOC *Res. Pract. Assess.* **8** 13

[47] Laws P, Willis M, Jackson D, Koenig K and Teese R 2015 Using research-based interactive video vignettes to enhance out-of-class learning in introductory physics *Phys. Teach* **53** 114

[48] Hestenes D, Wells M and Swackhamer G 1992 Force concept inventory *Phys. Teach* **30** 141

[49] Maloney D, O'Kuma T, Hieggelke C and Van Heuvelen A 2001 Surveying students' conceptual knowledge of electricity and magnetism *Am. J. Phys. Suppl.* **69** s12

[50] Madsen A, McKagan S B and Sayre E C 2013 Gender gap on concept inventories in physics: what is consistent, what is inconsistent, and what factors influence the gap? *Phys. Rev. ST PER* **9** 020121

Traxler A L, Cid X C, Blue J and Barthelemy R 2016 Enriching gender in physics education research: a binary past and a complex future *Phys. Rev. Phys. Educ. Res.* **12** 020114

Adegoke B A 2012 Impact of interactive engagement on reducing the gender gap in quantum physics learning outcomes among senior secondary school students *Phys. Educ.* **47** 462

[51] Docktor J and Heller K 2008 Gender differences in both force concept inventory and introductory physics performance *AIP Conf. Proc.* **1064** 15

[52] Bates S, Donnelly R, MacPhee C, Sands D, Birch M and Walet N R 2013 Gender differences in conceptual understanding of Newtonian mechanics: a UK cross-institution comparison *Eur. J. Phys.* **34** 421

[53] Lorenzo M, Crouch C and Mazur E 2006 Reducing the gender gap in the physics classroom *Am. J. Phys.* **74** 118

[54] Beichner R, Saul J, Abbott D, Morse J, Deardorff D, Allain R, Bonham S, Dancy M and Risley J 2007 The Student-centered activities for large enrollment undergraduate programs (SCALE-UP) project *Research-Based Reform of University Physics* ed E Redish and P Cooney (published http://compadre.org/per/per_reviews/volume1.cfm)

[55] Cahill M, Hynes K, Trousil R, Brooks L, McDaniel M, Repice M, Zhao J and Frey R 2014 Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum *Phys. Rev. ST PER* **10** 2

[56] Day J, Stang J, Holmes N, Kumar D and Bonn D 2016 Gender gaps and gendered action in a first-year physics laboratory *Phys. Rev. ST PER* **12** 2

[57] Maries A and Singh C 2015 Stereotype threat? Effects of inquiring about test takers' gender on conceptual test performance in physics *AIP Conf. Proc.* **1697** 120008

[58] Coletta V P and Phillips J A 2005 Interpreting FCI scores: normalized gain, preinstruction scores, and scientific reasoning ability *Am J. Phys.* **73** 1172
V Coletta V P, Phillips J A and Steinert J 2012 FCI normalized gain, scientific reasoning ability, thinking in physics, and gender effects *AIP Conf. Proc.* **1413** 23

[59] Majors T and Engelhardt P 2014 Gender & LEAP pedagogy: what does the gender force concept inventory have to say? *PERC Proc.* p 167 (http://www.per-central.org/items/detail.cfm?ID=13510)

[60] Marchand G C and Taasoobshirazi G 2013 Stereotype threat and women's performance in physics *Int. J. Sci. Educ.* **35** 3050

[61] Appel M and Kronberger N 2012 Stereotypes and the achievement gap: stereotype threat prior to test taking *Educ. Psychol. Rev.* **24** 609

[62] McKown C and Weinstein R S 2003 The development and consequences of stereotype consciousness in middle childhood *Child Development* **74** 498

[63] Bian L, Leslie S-J and Cimpian A 2017 Gender stereotypes about intellectual ability emerge early and influence children's interests *Science* **355** 6323

[64] Bandura A 1977 Self-efficacy: toward a unifying theory of behavioral change *Psychol. Rev.* **84** 19

[65] Miller K, Schell J, Ho A, Lukoff B and Mazur E 2015 Response switching and self-efficacy in Peer instruction classrooms *Phys. Rev. ST Phys. Educ. Res.* **11** 010104

[66] Dweck C 2006 *Mindset: The New Psychology of Success* (New York: Random House Incorporated)

[67] Singh C and Rosengrant D 2003 Multiple-choice test of energy and momentum concepts *Am. J. Phys.* **71** 607

[68] Rimoldini L and Singh C 2005 Student understanding of rotational and rolling motion concepts *Phys. Rev. ST PER* **1** 010102

[69] Ding L, Chabay R, Sherwood B and Beichner R 2006 Valuating an assessment tool: brief electricity and magnetism assessment *Phys. Rev. ST PER* **1** 10105

[70] Singh C and Rosengrant D 2001 Students' conceptual knowledge of energy and momentum *Proc. Physics Education Research conf. (Rochester)* p 123 (http://doi.org/10.1119/perc.2001.pr.018)

[71] Li J and Singh C 2012 Developing a magnetism conceptual survey and assessing gender differences in student understanding of magnetism *AIP Conf. Proc.* **1413** 43
Li J and Singh C 2017 Developing and validating a conceptual survey to assess introductory physics students' understanding of magnetism *Eur. J. Phys.* **38** 025702

[72] Singh C 2006 Student understanding of symmetry and Gauss's law of electricity *Am. J. Phys.* **74** 923
Singh C 2005 Student understanding of symmetry and Gauss's law *AIP Conf. Proc.* **790** 65–8

[73] Glass G and Hopkins K 1996 *Statistical Methods in Education and Psychology* 3rd edn (London: Pearson)

[74] Cohen J 1988 *Statistical Power Analysis for the Behavioral Sciences* 2nd edn (Oxford: Routledge)

[75] Smith G A 2008 First-day questions for the learner-centered classroom *Natl Teach. Learn. Forum* **17** 1

[76] https://aip.org/statistics/women

[77] Aguilar L, Walton G and Wieman C 2014 Psychological insights for improved physics teaching *Phys. Today* **67** 43