Received: 8 May 2020 | Revised: 16 March 2022 | Accepted: 21 March 2022

DOI: 10.1002/jee.20467

RESEARCH ARTICLE

JEE Journal of Engineering Education

Gender, prior knowledge, and the impact of a flipped linear algebra course for engineers over multiple years

Cécile Hardebolle[1,2] | Himanshu Verma[1,3] | Roland Tormey[1,2] | Simone Deparis[1,4]

[1]Center for Learning Sciences (LEARN), Ecole polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland

[2]Teaching Support Center (CAPE), Ecole polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland

[3]Faculty of Industrial Design and Engineering, Department of Sustainable Design Engineering, Delft University of Technology, Delft, The Netherlands

[4]Section of Mathematics, Ecole polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland

Correspondence
Cécile Hardebolle, Ecole polytechnique fédérale de Lausanne (EPFL), Station 1, CH-1015 Lausanne, Switzerland.
Email: cecile.hardebolle@epfl.ch

Abstract

Background: Research shows that active pedagogies could play an important role in achieving more equitable outcomes for diverse groups of students in Science, Technology, Engineering, and Mathematics (STEM). Although flipped classes are a popular active methodology, there is a lack of high-quality studies assessing their impact in ecologically valid settings and exploring how outcomes are related to gender and to prior education.

Purpose: This paper presents two modified replications of an experimental study investigating the impact of the flipped class approach on students' achievement in a large, first-year class in an engineering bachelor's degree.

Methodology: We added a new strand, progressively flipped over 3 years, to eight parallel strands of a high-stakes mandatory linear algebra course for engineers. The study followed a replicated-between-subjects design, with students in the flipped strand learning the same material as those in the other strands and taking the same final exam.

Results: Our results demonstrate that the flipped format did not have any significant impact on students' achievement compared to traditional lecturing. However, both replications in the flipped condition show a reduced attainment gap for women and students with less prior knowledge in mathematics.

Conclusion: While the flipped class seems to have weaker effects on learning than other active methodologies, the evidence in this study indicates that it may have an impact on reducing the attainment gap between different groups of students. It may therefore be particularly interesting to consider in efforts to achieve more equitable outcomes for women and where students have heterogeneous educational backgrounds.

KEYWORDS

active learning, engineering education, experimental research, flipped class, gender, linear algebra

1 | INTRODUCTION

Science, technology, engineering, and mathematics (STEM) fields have had a lingering problem with diversity. Studies have shown that despite policymakers' efforts in the past 50 years to increase the representation of women, in particular, the progress has been slow (Lichtenstein et al., 2015). This trend has held true for more or less all STEM fields, while in engineering education, "minimal progress has been made in recruiting and retaining students, and especially women and minorities" (Lichtenstein et al., 2015, p. 314), and some research suggests that "engineering education is not simply numerically male dominated, it is also culturally associated with masculinity" (Aeby et al., 2019, p. 756). Nosek and his colleagues have found that this implicit bias, which associates STEM fields with men, correlates with differences in the academic performance of men and women in mathematics (Nosek et al., 2009; Nosek & Smyth, 2011). Others have shown that women's perceived inclusion within their respective engineering programs undergoes a gradual decline over time (Marra et al., 2009). Beasley and Fischer (2012) explain that this phenomenon influences women's long-term persistence in the engineering domain.

Seymour and Hewitt (1997) suggest that the culture of educational practices in engineering education, including teaching styles, influences the low retention rates for women. They argue that attracting and (more importantly) retaining women would entail profound changes in existing classroom instruction methodologies. A recent meta-analysis by Theobald et al. (2020) identifies that active and interactive teaching approaches promote inclusivity within STEM education, produce more equitable educational outcomes, and in particular, reduce the achievement gap among different groups of students. This remains, however, an under-researched topic, and Theobald et al. (2020) found that there were no sufficient studies yet to include gender in their analysis of active learning and underrepresented students in undergraduate STEM education. They also identified the need for further research to distinguish which approaches to interactive teaching were most likely to have an impact.

One educational innovation explored in STEM and engineering education settings is the use of flipped class approaches. Perhaps the simplest definition of the flipped class is that in flipped classes, less cognitively active events which have traditionally taken place inside the class (like sitting in lectures) happen outside, while more cognitively active events which have traditionally taken place outside class (like completing exercises or working on applying the ideas introduced in lectures), now happen in class (Lage et al., 2000). More than a simple reordering of learning activities, flipped classes take advantage of two recent developments in approaches to teaching and learning. First, flipped classes typically aim to increase the use of active learning techniques in class with the teacher present. Since there is evidence that active learning approaches such as peer instruction can play a role in reducing gender-based disparities (Lorenzo et al., 2006), this seemed worth exploring. Second, flipped classes benefit from the increased availability of online educational content to enable the asynchronous presentation of course material, thereby freeing up the teacher's time for more active engagement with students (O'Flaherty & Phillips, 2015).

In this article, we examine the impact on the achievement of a heterogeneous population of students of flipping a large, mandatory, and high-stakes linear algebra course in the undergraduate engineering program of the Ecole polytechnique fédérale de Lausanne (EPFL), a science and engineering university in Switzerland. Our analysis considers two modified replications of flipping the same course across subsequent academic years, where the students' exposure to the flipped class format was changed incrementally from 1 year to another. The experiment presented in this article looked at the influence of the flipped format in an ecologically valid setting and addressed the following research questions:

1. What is the impact of the flipped class format on students' achievement compared to the traditional (lecture-based) format?
2. Is there a differential effect of the flipped format on different student groups (specifically, gender, high school background, and prior level in mathematics)?

In the following sections, we first present a literature review of past research on the flipped format and illustrate the research gaps which we bridge through this article. Then we present our study context, design, participants, and results. Finally, we conclude with a discussion of our results, including the limitations of our research and the implications of our findings for the engineering education context.

2 | RELATED WORK

A shortage of well-skilled engineers and the lack of diversity in the profession are important problems confronting the STEM fields with far-reaching socioeconomic ramifications, such as income inequality and decreased workplace diversity (McKenna et al., 2014; Theobald et al., 2020). Lichtenstein et al. (2015) provided a comprehensive account of the United States policymakers' efforts in the last 50 years to increase diversity in STEM fields. Similar measures to attract and sustain more women in STEM fields were also taken and documented within the United Kingdom and other European countries (Barnard et al., 2012; Powell et al., 2012). However, regardless of these efforts and numerous programs for making engineering fields more inclusive, progress has been slow and disheartening (Aeby et al., 2019; Lichtenstein et al., 2015; Seymour & Hewitt, 1997). A number of researchers attributed the low retention rates of women in engineering—particularly in undergraduate education—to students' negative perceptions and attitudes toward the prevalent culture of educational practices (Lorenzo et al., 2006; Nosek et al., 2009; Seymour & Hewitt, 1997). Secules (2017) argues that this slow progress in diversifying engineering education results from a misplaced focus, which has been "more on the overlooked assets of minority groups than on the acts of overlooking, more on the experiences of marginalized groups than on the mechanisms of marginalization by dominant groups, more on supporting and increasing minority student retention than on critiquing and remediating the systems which lead minority students to leave engineering" (Abstract). In order to address these fundamental and lingering problems in the engineering education context, researchers have argued in favor of an increased focus on the engagement and belonging of diverse students by proactively engaging them through collaborative or active learning approaches (e.g., Atadero et al., 2018; Lorenzo et al., 2006; Minin et al., 2016; Theobald et al., 2020).

In their meta-analysis of over 220 STEM studies, Freeman et al. (2014) found that active learning positively influenced learners' academic achievement and reduced learners' chances of failure as compared to traditional lecturing. Cooperative learning and feedback are also known to have an important positive impact on learning, with effect sizes of respectively $d = 0.59$ (when compared to individual learning) and $d = 0.73$ in the meta-analysis by Hattie (2009). Since lack of time for active learning is a recurrent problem for intensive and time-constrained programs such as those in engineering education, the flipped class seems worthy of investigation; moreover, given that the flipped class represents an active learning methodology incorporating elements of cooperative learning and feedback to students (Cheng et al., 2019; DeLozier & Rhodes, 2017; Lo et al., 2017; Lo & Hew, 2019; O'Flaherty & Phillips, 2015) it could well be expected to have an important positive impact on learning.

We have observed a significant growth in the number of studies looking at the impact of flipped classes on learning in recent years, and these, in turn, have been gathered in a number of recent meta-analyses in the field. While some of these have included a range of disciplines (Cheng et al., 2019), others have looked at the impact of the flipped format specifically in engineering education (e.g., Lo & Hew, 2019) and in math disciplines (e.g., Lo et al., 2017). These meta-analyses suggest a high degree of variability in both results and types of flipped courses. Lo and Hew (2019) analyzed 29 studies within the context of engineering education, published between 2008 and 2017, and show that while the flipped format had a positive—and significant—influence on students' achievement, the effect size was rather small ($g = 0.29$). Lo et al. (2017) conducted a meta-analysis, which solely considered studies on the flipped format in the domain of mathematics education. Their analysis of 21 studies also revealed that the flipped format is moderately effective, with a significant and positive influence on learning as compared to the traditional format. However, the effect size was again modest ($g = 0.30$). This suggests that the flipped format does improve learning, but not as radically and profoundly as anticipated. In their meta-analysis, including 55 publications with 115 effect sizes, Cheng et al. (2019) observed that studies in engineering disciplines showed no statistically significant impact of flipping a class, which indeed had a negative (if nonsignificant) effect. This result also led the authors to take a rather despairing position about the potential of the flipped format in engineering disciplines: "engineering appears to not be a suitable candidate for the flipped class method when compared to other disciplines" (Cheng et al., 2019, p. 810). While a review by Kerr (2015) found that students' grades improved in flipped classes and that students reported a higher satisfaction with the flipped format, the majority of studies in Kerr's meta-analysis examine the impact of the flipped format in classrooms of small sizes; indeed, it is common to look at flipped approaches in classes of 20–50 students (e.g., Mason et al., 2013; Schiltz et al., 2019). In the 2019 meta-analysis by Lo and Hew (2019), only 6 out of the 29 studies included concerned classes with more than 100 students.

The studies we have reviewed so far do not demonstrate a large effect size of flipped classes on student learning, particularly in the specific context of engineering education. A general tendency in the reviewed studies is the high variability in the results, without clear and consistent moderating factors except, perhaps related to the existence of transition activities at the start of the flipped class, identified as quite important in two studies by Lo et al. (i.e., Lo & Hew, 2019; Lo et al., 2017). While some individual studies do show that interactive teaching can have an impact on reducing gender differences in performance (e.g., Lorenzo et al., 2006), none of the meta-analyses cited above have explored the effects of the educational background or gender of students on their academic achievement under the flipped format. In their recent meta-analysis of interactive teaching in STEM education, Theobald et al. (2020) focused on studies that decompose the impact of interactive teaching in such a way that it is possible to look at the performance of students from different ethnic or socioeconomic groups. They do find that active learning narrows achievement gaps with respect to these students; however, they also find notable limitations in the existing data. First, there are relatively few studies that report disaggregated data, as a result of which they were unable to include gender as a variable in their analysis, and second, the poor quality of the descriptions of classroom practices means they are unable to distinguish between different types of interactive teaching.

If there are few studies that have disaggregated data for interactive teaching in general, there are even fewer for flipped classes. In one such study, Gross et al. (2015) report on a repeated investigation of students' academic achievement in a semester-long physical chemistry course for life science majors. The same instructor taught two iterations of the course in the flipped format, and the differences in students' scores were examined across different iterations of the course. The authors identified two separate but associated effects. First, the results showed that, although men performed significantly better than women in the traditional format (first three iterations), the gender difference was no longer statistically significant when the course was taught in the flipped format. Second, there was also a positive impact on the attainment of students with lower prior performance. Overall, the authors concluded that "the positive effects of the flipped class are most pronounced for students with lower grade point averages and for female students" (Gross et al., 2015, p. 1). In a smaller study, Chiquito et al. (2020) found that women achieved higher grades than men in a class taught using the flipped format. Another controlled study by Ryan and Reid (2016) showed that students with lower prior attainment levels gained significantly in the flipped condition, and as a consequence, the difference between previously higher and lower performing students was reduced in the flipped condition.

A consistent message emerging from these studies is the need to address the quality of data reported in accounts of interactive teaching in general and of flipped classes in particular. First, many studies do not adequately specify how the flipped format was implemented (including the type of learning activities used and their sequence, as well as the student workload and the number of contact hours), and as a consequence, do not allow the impact of different approaches to flipped classes to be evaluated. Second, the quality of the study designs is highly variable. Many studies do not describe the level of experimental control of the study conditions, including the comparability of teaching (teacher and content in particular), the type of evaluation (student feedback vs. evaluation of learning, including the type of assessment), and the comparability of student groups (especially how students are assigned to groups and the control of their prior attainment). Finally, numerous studies do not provide sufficient data to allow conclusions to be drawn (availability of detailed statistics, in particular disaggregated data, description of the type of learning assessment, etc.). As a consequence, meta-analyses often make design recommendations for studies on the flipped format, such as using comparable student groups with random assignment of participants and control for the previous achievement, collecting objective measures of learning with verified validity and reliability (e.g., Freeman et al., 2014), and experimentally controlling for teachers, content taught and study time (workload).

As a summary, previous research on flipped classes, especially studies relevant to engineering education, exposes several research gaps: (a) the lack of rigorous empirical studies on the flipped format with well-controlled experimental conditions, (b) the scarcity of flipped studies in classes of large size (≥100 students), and (c) the shortage of studies that examine whether the flipped format has a differential impact on different groups of students, in particular in terms of gender, academic background and prior attainment. In this article, we address these gaps by presenting a controlled, replicated study to investigate the impact of flipping a large, mandatory linear algebra course taught to a heterogeneous population of engineering undergraduate students.

3 | RESEARCH DESIGN

3.1 | Context

Secules et al. (2021) have recently highlighted the importance of researchers making clear their own positionality in engineering education. They highlight that the researchers' positionality impacts the questions asked, the perspectives taken regarding theories of knowledge, the relationship between researchers and research subjects/participants, the methodologies used, and the communication style adopted. While such positionality work is common in qualitative research, it is no less relevant in quantitative work like ours. In our case, our study arose out of a juxtaposition between educational research and established teaching practices. While there is considerable research evidence on the potential learning benefits of more interactive teaching, some of the mathematics teaching teams in the university where this study was carried out identified that this was at odds with their pedagogical culture, and some had doubts about its feasibility in practice. Following discussion with the university's educational development team, they decided to develop and rigorously evaluate the impact of an instance of interactive teaching in vivo. Since the mathematics team had already invested in the development of a massive open online course (MOOC), meaning they already had course material in digital format (videos, quizzes, etc.), they were particularly interested in testing a flipped class approach. One senior member of the mathematics teaching team (an experienced mathematician who had little prior experience in educational research) volunteered to develop and teach the flipped class. Although willing to try new pedagogical approaches, he was initially concerned about the impact and value of trying to change his already extremely successful teaching methods. He is one of the authors of this paper.

The mathematics team sought support from the university's pedagogical support team. Two members of that team helped to design the evaluation protocol and supported the teacher in the development of the educational practices and data collection and analysis. Philosophically, the educational development team was committed to an evidence-informed approach to educational development and saw this project as a way to highlight to the school the value of teachers using research approaches to develop their teaching practice. Thus, they adopted this project as strategically important and invested resources in supporting it. The educational development team was also particularly interested in exploring the experience of women in engineering education (and have previously researched this topic). The fourth member of the writing team later joined the project to contribute to data management and analysis.

Since the corresponding MOOC already existed, the mathematics team chose the linear algebra course, a first-semester bachelor's course required for all engineering programs in the university and taught to approximately 1800 students, in order to assess the impact of a flipped class in an ecologically valid setting. Nine different teachers teach the course with the same content in nine different classes, each of which we refer to as a “strand.” The number of students varies from one strand to another based on the number of student registrations in different engineering programs (mechanical engineering or electrical engineering for example). This course is weighted at six European Credit Transfer System (ECTS) credits, the credit weighting system used throughout the European Universities region, which corresponds to a total of approximately 180 h of work over the whole semester, including both in-class and independent study time. The course is assessed by an end-of-semester exam composed of multiple-choice questions (MCQs) 80% of which are common among the different linear algebra strands (the remaining 20% of questions may differ from one strand to another). This exam accounts for about a quarter of the average mean score used to decide whether students can continue to the second semester of the first year. Therefore, in addition to being a heavy course in terms of workload, the linear algebra course is also considered a high-stakes course.

3.2 | Design of the study

During the study, nine different teachers taught the course concurrently—on the same day and same hour: one teacher taught one strand in the flipped format (experimental “flipped” condition), and the other eight strands were taught in the traditional manner by eight other teachers (control condition). The comparison of the flipped format with students in multiple other strands taught by eight other teachers reduced the possibility that what we were measuring was simply a “teacher” effect. We further verified this by comparing the attainment of students in the experimental teacher's strand with students more generally in the years prior to the beginning of the flipped experiment.

We implemented the flipped class approach in an incremental way, as illustrated in Figure 1 and described below:

Year 0: In the autumn semester of the 2016–2017 academic year, the volunteering teacher taught the course in the flipped manner for 1 week (the last or 14th week of the semester, as shown in Figure 1) to a class of 295 students. We designed this phase as a pilot of the experiment in order to inform the teacher about the design of in-class activities and ways to adapt their pace to that of the students, and to elicit early feedback from students about the teaching methodology, as well as their perceptions regarding such experimentation in subsequent years.
Year 1: In the autumn semester of the 2017–2018 academic year, the teacher taught the course in the flipped manner for 5 weeks (Part B in Figure 1) from the fifth to the ninth week of the semester to a class of 109 students. In the eight other parallel strands, the class sizes varied from 193 to 298 students, with a median of 226 students. He taught the first 4 weeks (Part A in Figure 1) and the final 5 weeks (Part C in Figure 1) in a traditional instructional format similar to the other strands taught by different teachers.

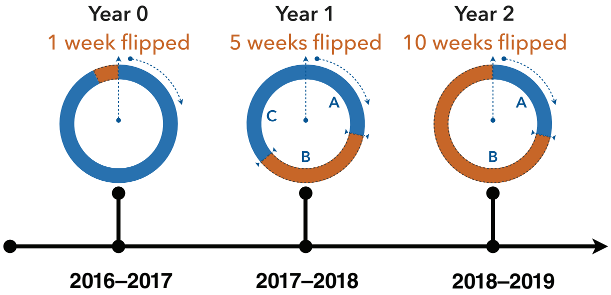


FIGURE 1 Incremental design of our flipped class study: The study officially started in the autumn semester of 2017–2018 (referred to as Year 1), but we flipped 1 week toward the end of the semester of the academic year 2016–2017 (referred to as Year 0). Since Year 1, we changed incrementally the number of weeks taught in the flipped format. In Year 1, only 5 weeks were taught in the flipped format, whereas in Year 2, 10 weeks were flipped.

Year 2: In the autumn semester of the 2018–2019 academic year, the teacher taught the course in the flipped manner for 10 weeks (Part B in Figure 1) from the fifth week to the end of the semester to a class of 202 students. In the eight other parallel strands, the class sizes varied from 180 to 307 students, with a median of 261 students. He taught the first 4 weeks (Part A in Figure 1) in a traditional instructional format similar to the other strands of the linear algebra course, as in the first 4 weeks of Year 1.

The rationale for incrementally increasing the duration of the flipped format (not including Year 0) was to (1) pragmatically assess the impact of the flipped model in ecologically valid settings within what is an extremely high-stakes course for the students taught in large classrooms, while mitigating the potential negative effects of such experimentation on students' learning, and (2) enable the teacher to develop strategies with a smaller group and then explore how these could be scaled up with a larger group while making sure that the course remains aligned in terms of the curriculum with the other parallel strands.

Our experiment, therefore, followed a replicated between-subjects design, where the students in the flipped strand learned the same topics and concepts weekly as their peers in the other strands (control condition). Our experiment can be referred to as a "modified replication" (APA, 2020) since we have repeated the experiment in order to bolster confidence in the results while incorporating slight changes from one iteration to the next.

For the sake of clarity and presentation, in this article, we will present results from Year 1 and Year 2 only, Year 0 being a pilot phase. It is worth noting that since each replication involved different populations of students with different assessments (exams with different sets of questions), Year 1 and Year 2 should be considered as two separate experiments. However, we merged the data for Year 1 and Year 2 for the last part of our analysis presented in Section 6.3, which considers the relationship between achievement and the prior mathematical attainment of students. We based this choice on the fact that the small number of students in some groups limited the possibilities for a meaningful year by year analysis, while at the same time the existence of similar tendencies across the 2 years suggested that merging the data did not distort the analysis.

We obtained approval from the ethics committee of the university to conduct the study. In compliance with the ethical protocol that was approved, students were duly informed, they consented to their data being used in the study when signing up for the course, and they could freely withdraw from the study and/or the course. We used two types of data records in the study: (1) class records (i.e., data collected in the context of the linear algebra course) and (2) university records. The latter contains demographic information about the students, such as gender and high school background, as well as information about their academic achievement in high school, which each student submits upon entering the university.

3.3 | Participants

The flipped course was offered only to volunteering students. In this section, we first describe the characteristics of the volunteers and then present the exclusion criteria we used to filter the data according to our ethics and data analysis protocols before describing the characteristics of the study participants.

3.3.1 | Volunteers

The teacher informed all the newly enrolled students in engineering disciplines about the possibility of participating in a flipped course by email 6 weeks before the start of the semester. He provided explanations on the nature of the class and of the experiment, as well as on the expectations. A total of 519 students (29% of all students taking the linear algebra course) volunteered to participate in the study in Year 1 and 373 (20%) in Year 2. The volunteers represented all five engineering programs of the university, in proportions representative of their respective numbers of registration in the first year (which remained approximately constant over the 2 years of the study). An average of 13% of volunteers were from civil and environmental engineering, 20% from computer and communication sciences, 5% from chemical engineering, 46% from mechanical and micro-engineering, and 17% from life sciences engineering.

Two weeks prior to the start of the semester, the teacher and a pedagogical advisor collectively assigned the volunteers to either the experimental or the control group using a stratified random sampling approach. We defined the strata based on the variables of interest that we introduce below, namely gender and high school background (secondary educational level). As a result, the proportionality of students' gender and prior educational background was preserved within the experimental and the control groups. The number of students who were assigned to the experimental (flipped) condition was 109 in Year 1 and 202 in Year 2 to fit the class sizes we had planned for. As a result, the size of the control group also changed from Year 1, where it included 410 students, to Year 2, where it had 171 students. The effective numbers we can report on are lower due to the filtering we present in the section below.

3.3.2 | Exclusion criteria

We cleaned the initial data of volunteers from both Year 1 and Year 2 and removed some participants before we analyzed the data. The following steps elaborate on how we proceeded:

1. We excluded the volunteering students who were minors (<18 years of age) at the time of data collection from our analyses, in line with our ethics and data management protocols. We identified 27 students as minors.
2. We also removed the volunteering students who were absent in the end-of-semester exam from the initial list of volunteers. A total of 43 students were on the list of absentees.
3. Students could withdraw from the study without giving any reason at any time before the anonymization of data. In addition, they were also free to withdraw by simply deregistering from the flipped strand during the first 2 weeks of the semester. A total of 30 students who were assigned to the flipped condition withdrew from the experiment, and consequently, we filtered out their data.
4. Finally, we filtered out the repeating students (152 students). Because of the fact that the repeating students have already finished their first semester once, their repeated exposure to the subject material may add a bias to our findings. As a result, we included only new students in further analyses.

Table 1 summarizes the distribution of participants whose data we subjected to the analysis phase following this filtering process. We introduce below the different variables of interest presented in this table.

3.3.3 | Variables of interest

Gender

Like many other technical universities, the institution in this study has put in place a number of measures in order to attract women into engineering programs and increase the diversity of its student cohorts. Despite these efforts, the proportion of women in the engineering programs overall remains relatively low: it was 31% in both years of the experiment. In comparison, the proportion of women participating in the study was 33% and 36% in Year 1 and Year 2, respectively. We controlled for gender when assigning students to the control and experimental groups (see Table 1 for the detailed distribution of the study participants by gender). Since one of our hypotheses is that the flipped format may have an effect on any gap between women's and men's achievement, we analyze the results in terms of this variable in Section 6.2.1.

TABLE 1 Study participants by background, gender, and conditions: This table summarizes the distribution of participating students over the different replications based on their background, gender, and conditions

			Gender		Condition	
Year	**Background**	**Total**	**Women**	**Men**	**Control**	**Flipped**
Year 1	INT-PAM	207	63	144	168	39
Year 1	NAT-PAM	74	16	58	59	15
Year 1	NAT-OTH	70	36	34	54	16
Sum (Year 1)		**351**	**115**	**236**	**281**	**70**
Year 2	INT-PAM	103	38	65	53	50
Year 2	NAT-PAM	48	12	36	24	24
Year 2	NAT-OTH	45	20	25	20	25
Sum (Year 2)		**196**	**70**	**126**	**97**	**99**
Year 1 + 2	INT-PAM	310	101	209	221	89
Year 1 + 2	NAT-PAM	122	28	94	83	39
Year 1 + 2	NAT-OTH	115	56	59	74	41
Total		**547**	**185**	**362**	**378**	**169**

Note: This table only shows the students' data resulting from the filtering process elaborated in Section 3.3.2.

High school background

The first-year bachelor's population in our school is composed of students coming from both the Swiss education system and other international education systems, with varied high school diplomas. As there is no selective entrance exam for Swiss students, students' prior knowledge in different subjects—particularly physics and advanced (reinforced) mathematics—varies significantly. Since we hypothesize that the flipped format may have an impact on differences in attainment between these student cohorts, we classified the incoming bachelor students into three distinct categories based on their high school background as described below:

International PAM (or INT-PAM): This category corresponds to students from a range of international education systems who have completed a high school diploma which includes a strong component of physics and applied mathematics (PAM). All these students are subjected to a selection process, and so they all have a background in PAM and were all high-performing within their respective high school system.

National PAM (or NAT-PAM): Students from the Swiss national secondary education system are not subject to a selection process and so arrive with a diverse set of subject specializations. The NAT-PAM category corresponds to students who studied PAM as their specialization during high school in the Swiss education system.

National Others (or NAT-OTH): Finally, this category corresponds to the students whose high school specialization in the Swiss national education system was in a subject other than PAM (such as philosophy, economics, or biology). These students are also heterogeneous within this category, in that some of them may have followed advanced mathematics courses while others only had basic mathematics courses.

Over the 2 years of the study, 39% of all incoming bachelor students had an INT-PAM background, while 25% had a NAT-PAM background. Students from the INT-PAM group over-volunteered to the study and made up 56% of participants in Year 1 and 53% and Year 2. The proportions of NAT-PAM students were 21% in Year 1 and 24% in Year 2. Table 1 presents the distribution of the study participants by high school background. As a result of the stratification process, the proportions of these subgroups are preserved within the experimental and control groups; therefore, the overrepresentation of INT-PAM students does not have an impact on the findings, as presented in Section 6.2.2.

Prior level of attainment in mathematics

The diverse origin of students and the lack of an entrance exam meant that there were no homogeneous metrics for quantifying students' prior knowledge in mathematics. To be able to take into account students' prior level in mathematics in our analyses, we used the official transcript that students obtain at the completion of their secondary education and that they submit to the university upon admission. Although the content of this transcript slightly varies as to whether students come from international (INT) or Swiss (NAT) education systems, it usually indicates one final grade for each discipline taken by the student. The mathematics professor in the team reviewed the curriculum of the main types of high school programs to determine the mathematical content relevant to linear algebra and to identify which grade to extract from the transcript. Assistants then manually extracted the grades (one grade per student).

We excluded from the further analysis the students who had no grade recorded for mathematics and students from the Swiss education system who had a grade for basic mathematics only since this program does not prepare for linear algebra. It is worth noting that this step resulted in a considerable additional reduction of the size of our experimental population for this analysis as 115 participants had no certified or only basic background in reinforced mathematics.

We then identified a series of score thresholds to partition students for whom we were able to identify a grade in reinforced mathematics into "Low Performing (labeled as Low)" and "High Performing (labeled High)" categories. Given the differences in grading systems, we had to define separate thresholds for INT and NAT students. On a normalized scale of [0, 1], with 0 being the lowest grade and 1 being the highest, the thresholds we used were:

1. NAT: Low: [0, 0.75] High: (0.75, 1.00].
2. INT: Low: [0, 0.85] High: (0.85, 1.00].

For validation purposes, we also performed a median split. This yielded similar results to the qualitative categorization, which suggested that the qualitative analysis and categorization were valid.

The distribution of the study participants by prior level of attainment in mathematics, gender, and condition can be found in Table 6. We present the analysis of students' academic attainment across the control and the flipped conditions based on their prior levels of attainment in reinforced mathematics in Section 6.3.

Note on ethnicity

In the English-speaking world, it is often seen as preferable to report data differentiated by ethnicity. Indeed, studies have identified on occasions that a lack of data on the differentiated impact of policies and practices could reflect a form of institutional racism (Pilkington, 2013; Trust & Parekh, 2000). This position is, however, more problematic outside the English-speaking world, where both linguistic traditions and histories are different from those of the Anglophone world. In many European countries, asking people to identify their ethnicity is regarded as more problematic. For example, in France, the principle of equal treatment means that it is forbidden in most circumstances to collect data on the ethnic origin of people, as is the inclusion of variables on race or religion in administrative files (for a detailed discussion of this issue across Europe, see Farkas, 2017). Although one of the goals of this study was differentiating learning data, our experience in previous research projects in which we did try to collect data differentiated by ethnicity was that, in line with the prevailing local practices, many students regarded collecting data on ethnicity as intrusive, unusual and problematic. Hence, in this project, we chose to collect a more limited range of demographic data.

3.4 | Instrument: Linear algebra end-of-semester exam

We used the linear algebra end-of-semester exam as a measure of student achievement in the study. This exam has a "common" part, which includes 80% of the exam questions in MCQ format, designed collectively by the teachers who teach the different strands of the linear algebra course. The questions for the remaining 20% are separately designed by each teacher for their respective strands. In this paper, we use the common part of the exam only to compute our measure of students' academic performance, as this part of the exam was identical for the control and flipped groups. It included 24 questions in both Year 1 and Year 2. Below, we describe the steps we took to compute this dependent variable.

Questions were negatively marked (+3 for a correct response, −1 for an incorrect response, and 0 for leaving the question unanswered). In order to ensure the validity of the measure, we applied the following procedure:

1. We removed questions that did not effectively distinguish between students (e.g., too easy, too hard, or confusing in some way). We computed a discrimination index (DI) value (Carneson et al., 2016) for each question in both years and removed the questions with a DI below 0.33. Our choice of this threshold was based on two criteria: (1) minimize

the number of filtered out questions, (2) and the boundary between questions to retain or to filter out should be clear, implying that the DIs of two questions should differ by at least 0.1 (10%). As a result, we removed three questions in Year 1 and two in Year 2.

2. We removed questions designed to examine students on the themes/topics taught during the initial 4 weeks of the course (nonflipped part) of both Year 1 and Year 2 (Part A in Section 3.2 and Figure 1). We analyze the impact of removing these questions on the overall results in Section 6.1. Although Part C in Year 1 was not flipped as such, we felt that the problem-solving methods addressed in Part B would impact the students' learning in Part C. Hence, we retained questions from Part C in Year 1 in the analysis. Part A of the exam included six questions in Year 1 and four questions in Year 2; therefore, we computed the scores on a total of 15 questions in Year 1 and 18 questions in Year 2.
3. Finally, we normalized the scores of each study participant against all the first-semester bachelor students who took linear algebra.

4 | TEACHING DESIGN

As introduced in Section 3.1, linear algebra is a six ECTS course. It is given over a 14-week semester with a weekly schedule of four periods of 45 min of lectures and two periods of recitation or exercise sessions with the teaching assistants (TAs), each split into two nonconsecutive days (two periods of lecture and one period of recitation/exercise session each). In addition, students are also expected to spend about 6 h per week on individual study. We kept this schedule and the overall workload for the flipped strand identical to that of the other strands: we neither reduced nor increased the number of contact hours.

In the following section, we describe the type of learning activities used in the flipped strand for preparatory work before class, in-class during the scheduled contact hours, and after class. To facilitate the categorization of these activities, we use the terms used by Lo and Hew (2019) whenever possible. Then we compare with the activities used in the other strands (control group).

4.1 | Learning activities in the flipped strand

4.1.1 | Preparatory work (pre-class activities)

The teacher sent instructions to students regarding the preparatory work for the whole week on the Friday of the week before. He provided an indicative duration for the different tasks, as well as the deadline by which to complete them (i.e., Day 1 or Day 2 of the scheduled in-class time). The typical preparatory work included a list of sections from a linear algebra MOOC by Professor Donna Testerman with video lectures and online quizzes, as well as an exercise worksheet. The teacher asked the students to take notes while watching the video lectures, like in a traditional lecture. The online quizzes enabled students to self-assess their learning and were not formally graded (although they were scored on the MOOC platform). The teacher took the exercise worksheet from the course material of previous years. He strongly encouraged students to work on the exercises by themselves before class, but they did not have to submit them, and the exercises were not graded.

4.1.2 | In-class activities

In-class time was divided into time with the teacher (twice two periods per week) and time with the TAs (twice one period per week). During the class time with the teacher, there were three types of activities: (1) Quizzes to start the session: these were usually True/False questions, designed to be answered in about 1 min. The percentage of correct responses helped the teacher to identify the common conceptual problems at the start of the session and also enabled the students to review the pre-class learning (this functioned as an interactive review session, in line with the findings of Lo et al. (2017) and Lo and Hew (2019)). (2) Short, problem-solving exercises: students were given some practice exercises, which could be completed in about 10 min each. Based on the percentage of correct responses in the class, the teacher then decided between asking students to work in small groups or managing a class solution with students' interventions. (3) Structured problems or proof-type problems: the teacher asked students to solve longer problems individually in a given time frame, and during this time, interacted with the students and gathered partial responses (of different steps, for example) to enable interaction and discussion at the level of the whole class. The teacher used a classroom response system (or "clickers") to collect students' anonymous answers, provide the class with immediate feedback and adjust the pace of the class.

During the recitation/exercise sessions with the TAs, students either worked individually or in small groups and benefited from one-to-one help by the TAs (roughly one TA for every 28 students). The students had the opportunity to complete the exercise sheet, as well as any uncompleted exercises from the in-class activities. The TAs usually did not present the solution to the exercises but only assisted the students with difficulties. The teacher provided a detailed written solution for all the exercises at the end of the week.

4.1.3 | Follow up work (postclass activities)

After the scheduled class time, students could review the course material and finish any remaining exercises, followed by verifying their work against the detailed solution.

4.2 | Learning activities in the other strands (control group)

In the other eight parallel strands, which followed what we have called a traditional approach, students were not expected to do any preparatory work before coming to class. In-class time with the teacher took the form of frontal lectures, mainly on the blackboard or equivalent. The recitation/exercise sessions were organized exactly in the same way as in the flipped strand (students working individually or in small groups with the help of TAs). However, since no exercises were addressed during the lecture time, overall, students had a much higher number of exercises to complete after the scheduled class time, along with reviewing the course material.

5 | DATA ANALYSES

We have used both parametric and nonparametric statistical procedures to conduct our analysis. Our choice of either parametric or nonparametric test was defined by the criteria elaborated by Harwell (1988). More specifically, we used parametric tests when the tests' underlying assumptions (normality, equality of variance) were met or when the test was robust to departures from these underlying assumptions.

In order to examine differences among the independent variables, we used Welch's sample *t*-test as the parametric test with Cohen's *d* to compute effect sizes (Cohen, 2013; Navarro, 2018). Among the nonparametric tests, we used the Kruskal–Wallis test with the Epsilon-squared method to compute effect sizes (Tomczak & Tomczak, 2014). In Section 6.2.2, following the examination of differences between independent variables using a Kruskal–Wallis test, we conducted a post hoc pairwise comparison between individual pairs. For this purpose, we used a Wilcoxon Rank-Sum test (nonparametric) with Bonferroni corrections to compute adjusted *p*-values (Navarro, 2018).

6 | RESULTS

In this section, our analyses consider the data from Year 1 and Year 2 as two separate experiments, except for the last part of the analysis presented in Section 6.3, where we combine the data from both years.

6.1 | Overall impact of the flipped class

The first question to address is whether the flipped class approach had any impact on overall student attainment. Table 2 presents the scores of students at the final exam across the two conditions after the removal of questions with

TABLE 2 Scores at the final exam, including all parts (A, B, & C) before normalization: The table illustrates the mean score, median score, and standard deviation (SD) for students in the flipped and control conditions before normalization but after removal of questions with low discrimination index

Scores before normalization (Parts A, B, and C)					
Year	**Condition**	***N***	**Mean**	**Median**	**SD**
Year 1	Control	281	31.00	32.00	15.60
Year 1	Flipped	70	31.70	31.5	15.00
$t(109.67) = -0.37$, $p = .71$; $d = 0.05$					
Year 2	Control	97	33.20	33.00	17.50
Year 2	Flipped	99	31.90	34.00	16.60
$t(192.77) = 0.55$, $p = .58$; $d = 0.08$					

Note: The maximum possible score is 63 for Year 1 and 66 for Year 2. We indicate the difference in students' mean score between the two conditions as Welch's two-sample *t*-tests for both course years (Year 1 and Year 2), where *d* is Cohen's measure of effect size.

TABLE 3 Normalized scores, with or without questions from Part A: The table illustrates the mean score, median score and standard deviation (SD) for students in the flipped and control conditions after normalization against all the first semester bachelor students who took linear algebra

Normalized scores (Parts A, B, and C)						**Normalized scores (Parts B and C)**					
Year	**Condition**	***N***	**Mean**	**Median**	**SD**	**Year**	**Condition**	***N***	**Mean**	**Median**	**SD**
Year 1	Control	281	−0.12	−0.05	0.98	Year 1	Control	281	−0.09	−0.01	0.97
Year 1	Flipped	70	−0.07	−0.08	0.94	Year 1	Flipped	70	−0.03	0.12	0.96
$t(109.67) = -0.37$, $p = .71$; $d = 0.05$						$t(106.80) = -0.42$, $p = .67$; $d = 0.06$					
Year 2	Control	97	−0.19	−0.20	1.01	Year 2	Control	97	−0.20	−0.15	1.00
Year 2	Flipped	99	−0.27	−0.15	0.95	Year 2	Flipped	99	−0.21	−0.08	0.97
$t(192.77) = 0.55$, $p = .58$; $d = 0.08$						$t(193.49) = 0.09$, $p = .92$; $d = 0.01$					

Note: The left side of the table shows the normalized scores, including the questions from Part A (first 4 weeks of the linear algebra course, which were taught as traditional lectures) while the right side of the table presents the normalized scores without Part A. In addition, we indicate the difference in students' mean score between the two conditions as Welch's two-sample *t*-tests for both course years (Year 1 and Year 2), where *d* is Cohen's measure of effect size.

low DI but before normalization against all the first-semester bachelor students who took linear algebra. The maximum possible score is 63 for Year 1 and 66 for Year 2. This table gives an idea of the size of the differences observed in the normalized scores presented in Table 3. In Year 1, we observe that students' scores in the flipped and the control condition do not differ significantly (Welch's two-sample *t*-test: $t[109.67] = -0.37$, $p = .71$; $d = 0.05$). We observe similar results for Year 2, where the differences in students' scores are again not statistically significant (Welch's two-sample *t*-test: $t\ [192.77] = 0.55$, $p = .58$; $d = 0.08$). On the basis of this data, it appears as if the flipped class had no evident effect on the overall attainment of students.

Although the participant filtering process (described in Section 3.3.2) was intended to improve the quality of the data used in the analysis, we also wanted to ensure it did not actually introduce unforeseen bias. This process removed a large portion of our data set. In Year 1, the size of our control group went from 410 to 281, and our experimental group was reduced from 109 to 70. In Year 2, the size of our control group was reduced from 171 to 97, while our experimental group was reduced from 202 to 99 participants (most of these reductions were due to the removal of repeating students). Nonetheless, our analysis indicates that this did not impact the overall pattern of findings.

Since part of the exam in both Year 1 and Year 2 addressed material covered in the early nonflipped weeks in the experimental setting (Part A), we also wanted to assess the impact of removing these questions from the analysis. We summarize this analysis in Table 3. This shows that the removal of these results does not change the overall findings: the flipped class format did not have any evident impact on the final attainment scores of students. Indeed, the removal of these scores did not affect the overall pattern of results to any notable extent. Questions from Part A are excluded from the analyses thereafter.

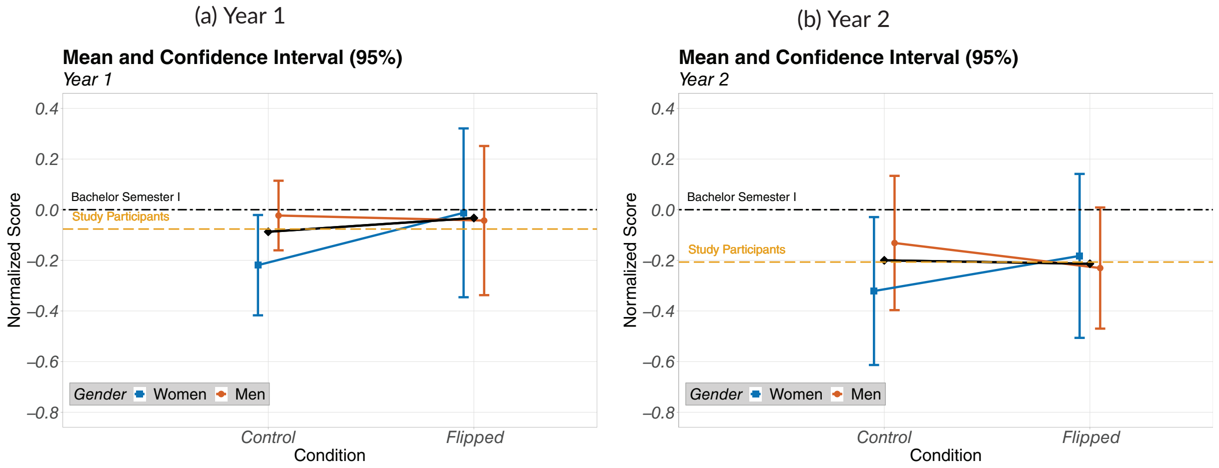
6.2 | Inclusiveness of the flipped class

Since our experimental population comprises several student cohorts with different gender and background characteristics, we also wanted to explore if the flipped format had a different impact on these cohorts. In this section, we analyze the differential effects of the flipped format on different student groups.

6.2.1 | Impact across gender

Prior to undertaking this study, men outperformed women on average in the traditionally taught linear algebra course. We hypothesized that this "gender gap" would be reduced in the flipped condition. Therefore, as a first step, we analyzed the differential impact of the flipped format for men and women.

In fact, regardless of the condition, the gender differences in students' scores are not significant in our data (using a Kruskal–Wallis test—Year 1: $\chi^2(\mathrm{df} = 1) = 2.13$, $p = .14$, $\varepsilon^2 = 0.006$; Year 2: $\chi^2(\mathrm{df} = 1) = 0.46$, $p = .49$, $\varepsilon^2 = 0.002$). There are also no significant differences between the attainment of women and that of men in the flipped and control conditions. However, a closer examination of the data (see Figure 2a,b) shows that women perform less well than men in the control condition on average, but this difference is reduced and inverted in the flipped class (see Table 4 and Figure 2). Despite the lack of statistical significance, which can be attributed to the small proportion of women in our study, the fact that a similar pattern emerges in both years is in itself notable. Provided that ours is a "real-world" study, with a significantly smaller proportion of women as compared to men, this repeating pattern for the flipped


(a) Year 1

(b) Year 2

FIGURE 2 Gender differences in achievement across conditions: This figure shows the mean and confidence interval values for differences in normalized scores across conditions and gender. The black dash-dotted horizontal line at $y = 0.0$ corresponds to the mean score in linear algebra for all Bachelor Semester I students. The orange dashed horizontal line represents the mean score of all study participants (irrespective of the condition). Finally, the black solid line with diamond markers represents the weighted mean across conditions.

TABLE 4 Gender differences in achievement across conditions: This table summarizes the mean score, median score, and standard deviation across gender and conditions

		Control				Flipped			
Year	Gender	*N*	Mean	Median	SD	*N*	Mean	Median	SD
Year 1	Women	92	−0.22	−0.18	0.97	23	−0.01	−0.01	0.82
Year 1	Men	189	−0.02	0.08	0.96	47	−0.04	0.16	1.03
Year 2	Women	35	−0.32	−0.22	0.88	35	−0.18	0.06	0.98
Year 2	Men	62	−0.13	0.02	1.07	64	−0.23	−0.08	0.98

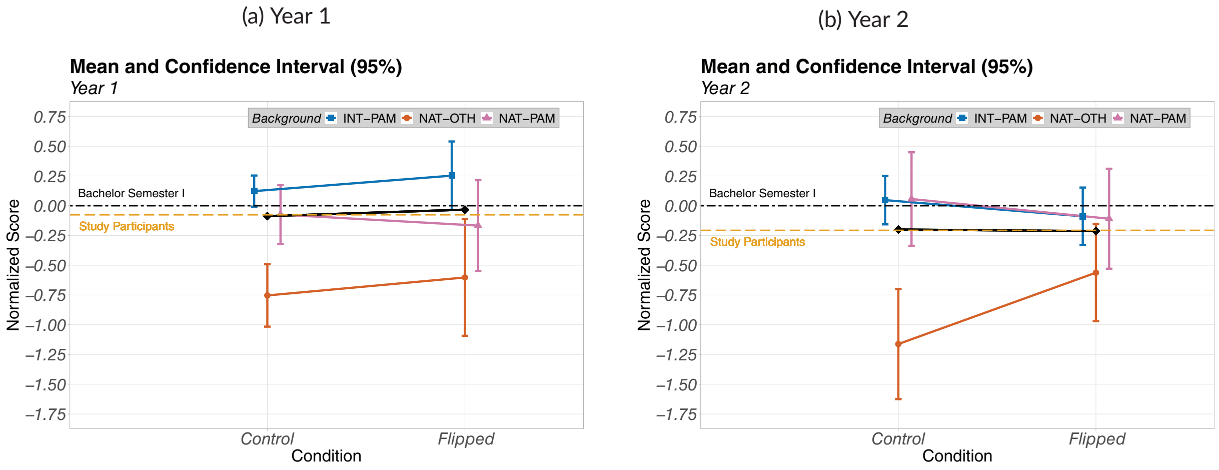
Note: The gender differences across flipped and control conditions are also illustrated graphically in Figure 2.

condition emphasizes that the reduced gender gap should probably be taken into account despite the lack of statistical significance (discussed further in Section 7).

6.2.2 | Impact across the educational background

In this section, we now look at the differences in achievement across educational backgrounds (see Table 5 and Figure 3). In both Year 1 and Year 2, we find significant differences between student cohorts with different high school backgrounds, irrespective of the conditions (Kruskal–Wallis test—Year 1: $\chi^2(\text{df} = 2) = 39.28$, $p < .001$, $\varepsilon^2 = 0.112$; Year 2: $\chi^2(\text{df} = 2) = 19.43$, $p < .001$, $\varepsilon^2 = 0.099$). In Year 1, post hoc tests show a statistically significant difference between INT-PAM and NAT-OTH (Wilcoxon Rank Sum test, with Bonferroni adjusted p-value $p < .001$) and between NAT-PAM and NAT-OTH (adjusted $p = .001$). However, the difference between INT-PAM and NAT-PAM is not significant (adjusted $p = .20$). Similarly, in Year 2, post hoc tests show a significant difference between INT-PAM and NAT-OTH (adjusted $p < .001$) and between NAT-PAM and NAT-OTH (adjusted $p = .002$). However, the difference is again not significant for INT-PAM and NAT-PAM (adjusted $p = 1$). In conclusion, regardless of the condition, INT-PAM students and NAT-PAM students tend to outperform those with NAT-OTH backgrounds in both years.

Similar to our previous analysis of the gender gap, we examined these cohorts separately across the two conditions. We observe a decrease in the gap between the scores of these three student groups in the flipped condition as compared


(a) Year 1

(b) Year 2

FIGURE 3 Background differences in achievement across conditions: This figure shows the mean and confidence interval values for differences in normalized scores across conditions and background. The black dash-dotted horizontal line at $y = 0.0$ corresponds to the mean score in linear algebra for all Bachelor Semester I students. The orange dashed horizontal line represents the mean score of all study participants (irrespective of the condition). Finally, the black solid line with diamond markers represents the weighted mean across conditions.

TABLE 5 Background differences in achievement across conditions: This table summarizes the mean score, median score, and standard deviation across background and condition

		Control				Flipped			
Year	**Background**	***N***	**Mean**	**Median**	**SD**	***N***	**Mean**	**Median**	**SD**
Year 1	INT-PAM	168	0.12	0.16	0.87	39	0.25	0.25	0.92
Year 1	NAT-OTH	54	−0.75	−0.86	0.98	16	−0.60	−0.39	1.00
Year 1	NAT-PAM	59	−0.08	−0.01	0.97	15	−0.17	0.08	0.76
Year 2	INT-PAM	53	0.05	−0.01	0.76	50	−0.09	−0.01	0.87
Year 2	NAT-OTH	20	−1.16	−1.42	1.06	25	−0.56	−0.70	1.04
Year 2	NAT-PAM	24	0.06	0.30	0.98	24	−0.11	0.23	1.05

Note: The background differences across flipped and control conditions are also illustrated graphically in Figure 3.

to the control condition. We used Kruskal–Wallis tests to assess the differences between the scores of INT-PAM, NAT-PAM, and NAT-OTH students in the different conditions. In Year 1, we observe a statistically significant difference in the scores of these student groups (see Figure 3a and Table 5) both in the control ($\chi^2(\text{df}=2)=30.86$, $p<.001$, $\varepsilon^2=0.11$) and in the flipped ($\chi^2(\text{df}=2)=9.97$, $p=.007$, $\varepsilon^2=0.14$) conditions. In Year 2 (see Figure 3b and Table 5), we observe again a statistically significant difference in the scores of the three student groups in the control condition ($\chi^2(\text{df}=2)=18.25$, $p<.001$, $\varepsilon^2=0.19$). However, this difference is not significant in the flipped condition ($\chi^2(\text{df}=2)=4.07$, $p=.13$, $\varepsilon^2=0.04$). To assess the pairwise differences between the scores of INT-PAM, NAT-PAM, and NAT-OTH students separately for the control and the flipped conditions, we used the Wilcoxon Rank-Sum test with Bonferroni correction as the post hoc test. In the control condition, we observe statistically significant differences in the scores of NAT-OTH and INT-PAM students (adjusted $p<.001$ in Year 1, adjusted $p<.001$ in Year 2) and of NAT-OTH and NAT-PAM students (adjusted $p=.002$ in Year 1, adjusted $p=.002$ in Year 2). The difference is not significant for the scores of NAT-PAM and INT-PAM students (adjusted $p=.6$ in Year 1, adjusted $p=1$ in Year 2). On the other hand, in the flipped condition, the only significant difference we observe is in Year 1 between the scores of NAT-OTH and INT-PAM students (adjusted $p=.008$) while it is not significant in Year 2 (adjusted $p=.2$). Neither the differences between the scores of NAT-OTH and NAT-PAM students (adjusted $p=.4$ in Year 1, adjusted $p=.4$ in Year 2) nor the

TABLE 6 Differences in achievement taking into account gender and prior mathematics levels: This table summarizes the mean score, median score, and standard deviation across prior mathematics level, gender, and conditions

		Control				Flipped			
Prior math level	Gender	*N*	Mean	Median	SD	*N*	Mean	Median	SD
High	Women	48	0.29	0.37	0.71	21	0.13	0.06	0.66
High	Men	107	0.27	0.42	0.86	53	0.19	0.19	0.85
High	**Sum**	**155**	**0.27**	**0.42**	**0.81**	**74**	**0.17**	**0.16**	**0.80**
Low	Women	45	−0.38	−0.26	0.85	20	0.11	0.23	0.85
Low	Men	89	−0.31	−0.26	1.00	34	−0.41	−0.42	0.96
Low	**Sum**	**134**	**−0.33**	**−0.26**	**0.95**	**54**	**−0.22**	**−0.15**	**0.95**

Note: The differences in students' prior level across their gender and condition are also illustrated in Figure 4.

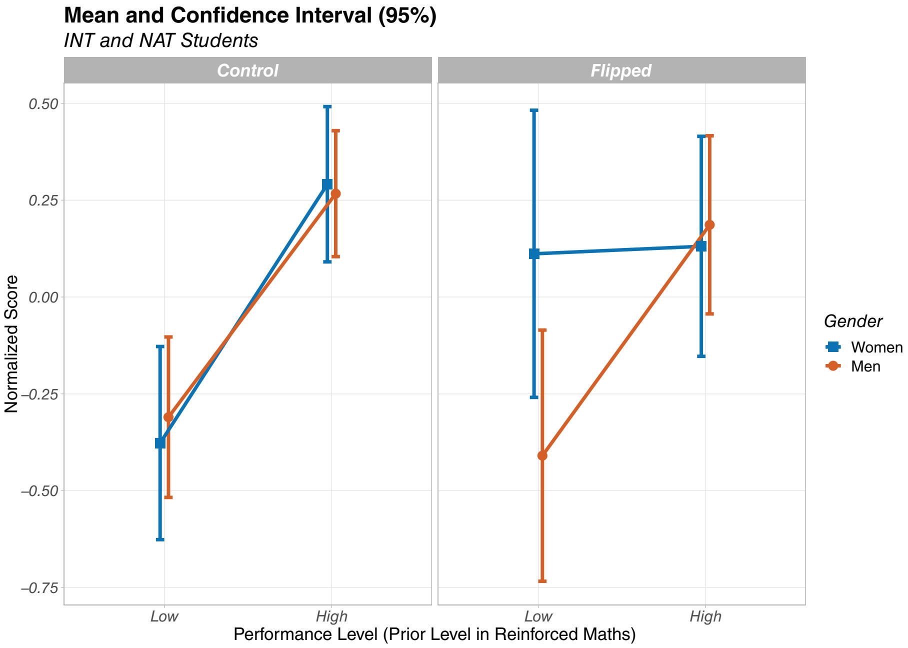


FIGURE 4 Differences in achievement taking into account gender and prior mathematics levels: This plot illustrates the mean score and confidence interval for the scores of students across gender, prior mathematics level, and conditions.

differences between the scores of NAT-PAM and INT-PAM students (adjusted $p = .4$ in Year 1, adjusted $p = 1$ in Year 2) are significant. Overall, the NAT-OTH group seems to benefit particularly from the flipped condition, even more so in Year 2. More generally, as illustrated in Figure 3, the gap in the score of the three student groups is smaller in the flipped condition for both Year 1 and Year 2 and is even smaller in Year 2. These findings may suggest that the flipped format is particularly beneficial in contexts with students from heterogeneous high school backgrounds and varying levels of prior knowledge in mathematics.

6.3 | Impact across prior mathematics level

We studied the effects of the different conditions on students with varying prior levels of attainment in mathematics. As described in Section 3.3.3, the construction of this variable resulted in the exclusion of a significant portion of our sample. Because we observed recurrent patterns in Year 1 and Year 2 as described in the previous sections, we considered it would be acceptable to combine the two datasets into one in order to carry out this analysis. Therefore the following results are to be interpreted as if the two successive years were one single experiment.

Table 6 and Figure 4 illustrate the mean attainment scores of students broken down by their prior level in reinforced mathematics. As expected, students with a stronger level in reinforced mathematics ("High") attained significantly higher scores in the end-of-semester exams (Kruskal–Wallis test: $\chi^2(\text{df} = 1) = 36.38$, $p < .001$, $\varepsilon^2 = 0.087$).

Examining the differences in the end-of-semester scores of students with High and Low prior knowledge across different conditions revealed that there is a statistically significant difference in the end-of-semester scores of students depending on their prior level in reinforced mathematics both in the control condition (Kruskal–Wallis test: $\chi^2(\text{df} = 1) = 32.67$, $p < .001$, $\varepsilon^2 = 0.11$) and in the flipped condition, where this difference is smaller (Kruskal–Wallis test: $\chi^2(\text{df} = 1) = 4.77$, $p = .03$, $\varepsilon^2 = 0.038$).

Finally, we examined if there were gender differences among students with High and Low prior-knowledge levels and whether these groups performed differently in the end-of-semester exam. As illustrated in Figure 4, we observe no gender differences in the control condition, where the end-of-semester performances of women and men in both Low and High prior-levels are equivalent. However, in the flipped condition, women with Low prior-levels in high school mathematics perform better than men with Low prior-levels, although the difference is marginally nonsignificant (Kruskal–Wallis test: $\chi^2(\text{df} = 1) = 3.61$, $p = .06$, $\varepsilon^2 = 0.068$).

7 | DISCUSSION

There is very substantial interest in flipped class approaches in higher education and science and engineering education within the context of a broader enthusiasm for interactive approaches to teaching. While there is growing evidence that interactive teaching has a more positive effect on student learning and performance than traditional teaching in STEM disciplines (Freeman et al., 2014), weaker effects have been found in studies focusing specifically on flipped class approaches. Lo and Hew (2019) found an effect size of only $g = 0.29$ in their work on flipped classes in engineering education, while Cheng et al. (2019) found even weaker effects ($g = 0.19$), with a weak positive effect for mathematics courses ($g = 0.21$) and a very weak negative effect for engineering courses ($g = -0.08$). Hattie (2009) has identified that, for educational interventions, effect sizes of less than 0.40 should be regarded as indicating a low effect. Despite the fact that our replication differed from 1 year to another in terms of intervention time, our data is consistent with prior findings in that it shows no effect on average attainment from the flipped class—the grades of the control and experimental groups were effectively the same in both Year 1 and Year 2 of the study (see Table 2).

This finding is not surprising, given the apparently weak impact of flipped classes in general and the short time frame of the intervention (the intervention took place in one-third of one semester in Year 1 and two-thirds of one semester in Year 2). This finding may also be explained in part by virtue of the nature of assessment: studies on interactive teaching which measure impacts using exams have found weaker effects compared to those using concept tests (Freeman et al., 2014). It may be that interactive teaching is more relevant when the focus is on the application of concepts to physical scenarios and less relevant when the focus is on mathematical thinking and proofs. Finally, class size may also be an issue. We had 109 students in the flipped strand in Year 1 and 202 in Year 2 (before filtering), where most existing studies on flipped classes are with classes in the 20–50 student size range (see Section 2). It may be possible that stronger effects could be achieved with smaller classes.

There has been a growing interest in so-called “null results” (such as ours) in education studies, given that null results appear to be so common when moving from the “efficacy” studies in highly controlled labs to more ecologically valid field-based randomized controlled “effectiveness” trials (Kim, 2019). Kim (2019) notes that in one review of effectiveness trials designed to evaluate a previously validated educational intervention, only 11 of 90 trials yielded positive results. Jacob et al. (2019) argue that, rather than seeing these null results as an indication that something does not work, when designed and interpreted appropriately, null results have the potential to yield valuable information. In particular, Jacob et al. (2019) note that interventions which do not show a significant positive impact may still be worthwhile if the intervention is desirable for some other reason.

With this in mind, what would be the implication for practice? The flipped class format is very popular with the students, as evidenced by the fact that the flipped strand has been considerably oversubscribed each year that we have offered it. While there is considerable work involved in switching from a traditional to a flipped class approach, materials, once developed, can be reused. Discounting the initial costs over time in this way suggests that the implementation of flipped class teaching may well be regarded as cost-effective (Lo & Hew, 2019). Indeed, taking into account that learning to give traditional lectures does, in itself, involve a steep learning curve, our experience suggests that the effort involved in becoming proficient in flipped class teaching is probably no greater than the effort involved in becoming proficient in traditional teaching. If so, and given that many students are looking for this kind of alternative to traditional teaching, if appropriate training can be offered to new faculty, and given that there is no real evidence of negative impact, it would seem strange not to offer this option to students. In the case of our institution, this study has led to increased uptake of the flipped format by other instructors. This format is now offered to first-year students not only in linear algebra but also in calculus and general physics courses in classes of circa 200 students.

Replication trials, even those that show little overall impact, can also shed light on previously unidentified interactions. One such interaction that is worthy of attention is the relationship between flipped teaching, gender, and prior mathematical background. While we know that there is some evidence that interactive teaching can reduce the so-called “gender gap” in science education (Haak et al., 2011; Lorenzo et al., 2006), existing reviews on flipped classes in engineering settings (e.g., Lo & Hew, 2019), in STEM education more generally (Lo et al., 2017), or in higher education (O'Flaherty & Phillips, 2015) do not address this issue well. Theobald et al. (2020) addressed the question of the differential impact of interactive teaching on different student groups in their review; however, they found that the data was not sufficient to examine the problem of gender differences. The data presented here suggests that it is possible that flipped classrooms may have positive impacts on the learning of women in engineering curricula. While women performed a little worse than men in each of the 2 years in our control group, they performed identically to men in the flipped class in both years. While the differences between the attainment of men and women were not statistically significant, the fact that the same pattern emerged in both years was notable.

A similar stronger pattern emerges when differences in students' prior education are considered. While in the control group, there are significant differences in attainment between students who have studied technical disciplines in high school (INT-PAM and NAT-PAM students in our sample) and those who have not taken a scientific strand in high school (NAT-OTH students in our sample), these differences were reduced and became nonsignificant in the flipped class group (see Figure 3 and Table 5). This pattern is even clearer when one looks at the experience of women who enter with comparatively low grades in high school mathematics. In the control group, both men and women who come in with lower high school grades in mathematics tend to have a similarly weak performance in their end-of-semester exams. In the flipped class, this pattern does not hold true for women; women who come into the flipped class with lower high school mathematics grades tend to do as well in the flipped class as both men and women who come in with stronger high school mathematics grades in both the flipped and control classes. While the difference between control and flipped settings is marginally nonsignificant ($p = .06$), the pattern is quite notable.

8 | LIMITATIONS

Nonetheless, this study does have its limitations. In Year 1, a side effect of the remarkably high number of students who volunteered compared to the class size we had planned for is the unbalanced size of the control and experimental groups. While it would have been preferable to have more balanced groups, we addressed this issue within the statistical analysis by virtue of the stratification we used when assigning students to the groups. The overrepresentation of some groups of students in our sample compared to our population, as noted in Section 3.3, is also potentially a limitation of our study, although it does not impact the statistical analysis, again thanks to the stratification.

Classroom heterogeneity and the need to control for students' prior educational trajectory meant that although we had quite large numbers of volunteers, we were left with rather fewer students in the analysis (351 participants in Year 1 and 196 in Year 2). We would contend that this is a reasonable outcome given the desire to achieve both internal and ecological validity in our study design. Nonetheless, it should be recognized as a limitation and one which may well have impacted our ability to identify statistically significant findings from the data.

Another limitation of our work is that we use binary and mutually exclusive categories of woman/man to describe the gender of participants. This limitation arises because this is the way gender is represented in the university's academic database from which we drew the data. However, data from other studies in the university suggest that only circa 2% of members of the university community identify as a gender other than binary woman/man. As such, and taking into account the sample size, even had other gender identifiers been included, it is unlikely that it would have been possible to draw meaningful conclusions from the additional data.

In addition, due to the incremental way in which we introduced the flipped class approach over time, the actual length of the flipped component in both years was comparatively short. This may well have lessened the potential impact of the flipped approach. While we would contend that this approach was a realistic way of implementing a pedagogical change in the context of a large, mandatory, and high-stakes course, it should be recognized that a more systematic and consistent use of flipped class approach may well have a deeper impact on students' learning.

We did not investigate conceptual understanding (e.g., using a concept inventory) either, and this might be a valuable path to explore for future studies; however our choice of using a real end-of-semester exam as an instrument is an important factor in the ecological validity of our study. Finally, the duration and scope of our study are also limited since only one institution was involved over the course of 2 years.

9 | CONCLUSION

Cheng et al. (2019) note that many of the existing studies of flipped classes are of questionable design and that quite a few do not provide adequate information about the study design to be effectively used to draw conclusions. Our aim in this study was to provide a clear account of both the research design and the instructional design to allow others to draw conclusions from our data. Our study explores what happens when flipped class approaches are used in a real teaching and learning setting in engineering education, with a high-stakes course, addressing complex technical content. Such real-life contexts can be messy, with students drawn from a variety of backgrounds and trajectories. These kinds of ecologically valid studies also have many potentially intervening variables, including students' prior knowledge, self-efficacy beliefs, and motivations, as well as teacher's skill and behavior. This study was designed so that it meets the criteria for high-quality studies already in use in the field, specifically, comparability between control and experimental groups in terms of assessment, students, and instructors (see Freeman et al., 2014). The quality of the design of our study makes our results all the more important, considering that they show that the flipped format did not have a particular impact on students' achievement overall. Null results like ours are not frequently published, which is an important source of bias in educational studies. Despite the lack of statistical significance in our results, our analysis has uncovered some trends worth investigating in terms of the inclusiveness of the flipped format. Our two modified replications show three recurring patterns: (a) the flipped format resulted in smaller differences in the achievement of women and men; (b) the flipped format resulted in smaller differences in the achievement of students with different high school backgrounds; (c) women with weaker prior math attainment achieved better results in the flipped condition. We think that this gives us some indication as to how to provide heterogeneous classes with a better learning experience and to better retain both women and students with nonscientific high school backgrounds in engineering education, and therefore, increase diversity in the field.

Our study also suggests some future trajectories for further research. It is notable that flipped class strategies seem to have, on average, a less positive impact on learning than the use of other types of interactive strategies. Lo and Hew (2019) also suggest that some review component in the flipped class approach seems to have a positive impact on attainment (there was a review component in the design of the flipped class in this study). Rather than simply comparing flipped with traditional courses, future research may wish to focus on whether different approaches to flipping classes may have different impacts. Given the significant challenges facing engineering education in attracting and retaining women and students (Lichtenstein et al., 2015), with nonscientific prior educational backgrounds (Aeby et al., 2019; Lichtenstein et al., 2015) and given that this issue has been largely neglected by existing research on flipped classes (see Lo & Hew, 2019) we suggest this should be a priority for future research.

ACKNOWLEDGMENTS

We would like to thank the statistics office of EPFL, in particular Aline Blasco Stampfli, as well as Patrick Jermann and Francisco Pinto, for their help in extracting the data from the academic database; Ludovic Bonivento, for his help with the grading data of the linear algebra exam; Jessica Dehler Zufferey and Melissa Skweres, for their helpful feedback and proofreading; Anne Remillet, for her help in managing the manual processing of the high school transcripts; Daniel Kressner, former director of the Section of Mathematics, for his support; Donna Testerman for her linear algebra MOOC. Open access funding provided by Ecole Polytechnique Federale de Lausanne.

ORCID

Cécile Hardebolle https://orcid.org/0000-0001-9933-1413
Himanshu Verma https://orcid.org/0000-0002-2494-1556
Roland Tormey https://orcid.org/0000-0003-2502-9451
Simone Deparis https://orcid.org/0000-0002-2832-6630

REFERENCES

Aeby, P., Fong, R., Vukmirovic, M., Isaac, S., & Tormey, R. (2019). The impact of gender on engineering students' group work experiences. *International Journal of Engineering Education*, *35*(3), 756–765.

APA. (2020). *Dictionary of psychology*. Retrieved from https://dictionary.apa.org/replication

Atadero, R. A., Paguyo, C. H., Rambo-Hernandez, K. E., & Henderson, H. L. (2018). Building inclusive engineering identities: Implications for changing engineering culture. *European Journal of Engineering Education*, *43*(3), 378–398. https://doi.org/10.1080/03043797.2017.1396287

Barnard, S., Hassan, T., Bagilhole, B., & Dainty, A. (2012). 'They're not girly girls': An exploration of quantitative and qualitative data on engineering and gender in higher education. *European Journal of Engineering Education*, *37*(2), 193–204. https://doi.org/10.1080/03043797.2012.661702

Beasley, M. A., & Fischer, M. J. (2012). Why they leave: The impact of stereotype threat on the attrition of women and minorities from science, math and engineering majors. *Social Psychology of Education*, *15*(4), 427–448. https://doi.org/10.1007/s11218-012-9185-3

Carneson, J., Delpierre, G., & Masters, K. (2016). *Designing and managing multiple choice questions* (2nd ed.). https://doi.org/10.13140/RG.2.2.22028.31369

Cheng, L., Ritzhaupt, A. D., & Antonenko, P. (2019). Effects of the flipped classroom instructional strategy on students' learning outcomes: A meta-analysis. *Educational Technology Research and Development*, *67*(4), 793–824. https://doi.org/10.1007/S11423-018-9633-7

Chiquito, M., Castedo, R., Santos, A. P., López, L. M., & Alarcón, C. (2020). Flipped classroom in engineering: The influence of gender. *Computer Applications in Engineering Education*, *28*(1), 80–89. https://doi.org/10.1002/cae.22176

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic Press.

DeLozier, S. J., & Rhodes, M. G. (2017). Flipped classrooms: A review of key ideas and recommendations for practice. *Educational Psychology Review*, *29*(1), 141–151. https://doi.org/10.1007/s10648-015-9356-9

Farkas, L. (2017). *Analysis and comparative review of equality data collection practices in the European Union: Data collection in the field of ethnicity* (European Commission & Directorate-General for Justice and Consumers, Eds.). https://doi.org/10.2838/447194

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Gross, D., Pietri, E. S., Anderson, G., Moyano-Camihort, K., & Graham, M. J. (2015). Increased preclass preparation underlies student outcome improvement in the flipped classroom. *CBE—Life sciences Education*, *14*(4), ar36. https://doi.org/10.1187/cbe.15-02-0040

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, *332*(6034), 1213–1216. https://doi.org/10.1126/science.1204820

Harwell, M. R. (1988). Choosing between parametric and nonparametric tests. *Journal of Counseling & Development*, *67*(1), 35–38. https://doi.org/10.1002/j.1556-6676.1988.tb02007.x

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Jacob, R. T., Doolittle, F., Kemple, J., & Somers, M.-A. (2019). A framework for learning from null results. *Educational Researcher*, *48*(9), 580–589. https://doi.org/10.3102/0013189X19891955

Kerr, B. (2015). *The flipped classroom in engineering education: A survey of the research*. Paper presented at the International Conference on Interactive Collaborative Learning (ICL), Firenze, Italy. https://doi.org/10.1109/ICL.2015.7318133

Kim, J. S. (2019). Making every study count: Learning from replication failure to improve intervention research. *Educational Researcher*, *48*(9), 599–607. https://doi.org/10.3102/0013189X19891428

Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. *The Journal of Economic Education*, *31*(1), 30–43. https://doi.org/10.1080/00220480009596759

Lichtenstein, G., Chen, H. L., Smith, K. A., & Maldonado, T. A. (2015). Retention and persistence of women and minorities along the engineering pathway in the United States. In A. Johri & B. Olds (Eds.), *Cambridge handbook of engineering education research* (pp. 311–334). Cambridge University Press. https://doi.org/10.1017/CBO9781139013451.021

Lo, C. K., & Hew, K. F. (2019). The impact of flipped classrooms on student achievement in engineering education: A metaanalysis of 10 years of research. *Journal of Engineering Education*, *108*(4), 523–546. https://doi.org/10.1002/jee.20293

Lo, C. K., Hew, K. F., & Chen, G. (2017). Toward a set of design principles for mathematics flipped classrooms: A synthesis of research in mathematics education. *Educational Research Review*, *22*, 50–73. https://doi.org/10.1016/j.edurev.2017.08.002

Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, *74*(2), 118–122. https://doi.org/10.1119/1.2162549

Marra, R. M., Rodgers, K. A., Shen, D., & Bogue, B. (2009). Women engineering students and self-efficacy: A multiyear, multi-institution study of women engineering student self-efficacy. *Journal of Engineering Education*, *98*(1), 27–38. https://doi.org/10.1002/J.2168-9830.2009.TB01003.X

Mason, G. S., Shuman, T. R., & Cook, K. E. (2013). Comparing the effectiveness of an inverted classroom to a traditional classroom in an upper-division engineering course. *IEEE Transactions on Education*, *56*(4), 430–435. https://doi.org/10.1109/TE.2013.2249066

McKenna, A. F., Froyd, J., & Litzinger, T. (2014). The complexities of transforming engineering higher education: Preparing for next steps. *Journal of Engineering Education*, *103*(2), 188–192. https://doi.org/10.1002/jee.20039

Minin, S., Varodayan, D., Schmitz, C., Faulkner, B., San Choi, D., & Herman, G. L. (2016). *Minority merit: Improving retention with cooperative learning in a first-year electronics course*. Paper presented at the IEEE Frontiers in Education Conference (FIE), Erie, PA, USA. https://doi.org/10.1109/FIE.2016.7757611

Navarro, D. (2018). *Learning statistics with R: A tutorial for psychology students and other beginners: Version 0.6*. University of Adelaide Adelaide. Retrieved from https://learningstatisticswithr.com/

Nosek, B. A., & Smyth, F. L. (2011). Implicit social cognitions predict sex differences in math engagement and achievement. *American Educational Research Journal*, *48*(5), 1125–1156. https://doi.org/10.3102/0002831211410683

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., Kesebir, S., Maliszewski, N., Neto, F., Olli, E., Park, J., Schnabel, K., Shiomura, K., Tulbure, B. T., Wiers, R. W., … Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*(26), 10593–10597. https://doi.org/10.1073/pnas.0809921106

O'Flaherty, J., & Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *The Internet and Higher Education*, *25*, 85–95. https://doi.org/10.1016/j.iheduc.2015.02.002

Pilkington, A. (2013). The interacting dynamics of institutional racism in higher education. *Race Ethnicity and Education*, *16*(2), 225–245. https://doi.org/10.1080/13613324.2011.646255

Powell, A., Dainty, A., & Bagilhole, B. (2012). Gender stereotypes among women engineering and technology students in the UK: Lessons from career choice narratives. *European Journal of Engineering Education*, *37*(6), 541–556. https://doi.org/10.1080/03043797.2012.724052

Ryan, M. D., & Reid, S. A. (2016). Impact of the flipped classroom on student performance and retention: A parallel controlled study in general chemistry. *Journal of Chemical Education*, *93*(1), 13–23. https://doi.org/10.1021/acs.jchemed.5b00717

Schiltz, G., Feldman, G., & Vaterlaus, A. (2019). Active-learning settings and physics lectures: A performance analysis. *Journal of Physics: Conference Series*, *1286*, 012019. https://doi.org/10.1088/1742-6596/1286/1/012019

Secules, S. (2017). *Beyond diversity as usual: Expanding critical cultural approaches to marginalization in engineering education* (Doctoral Dissertation). Available from the Digital Repository at the University of Maryland (College Park, MD). https://doi.org/10.13016/M22008

Secules, S., McCall, C., Mejia, J. A., Beebe, C., Masters, A. S., Sánchez-Peña, M. L., & Svyantek, M. (2021). Positionality practices and dimensions of impact on equity research: A collaborative inquiry and call to the community. *Journal of Engineering Education*, *110*(1), 19–43. https://doi.org/10.1002/jee.20377

Seymour, E., & Hewitt, N. M. (1997). *Talking about leaving: Why undergraduates leave the sciences*. Westview Press.

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintrón, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones, L., Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., … Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences*, *117*(12), 6476–6483. https://doi.org/10.1073/pnas.1916903117

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences*, *1*(21), 19–25.

Trust, R., & Parekh, B. C. (Eds.). (2000). *The future of multi-ethnic Britain: Report of the commission on the future of multi-ethnic Britain*. Profile Books.

AUTHOR BIOGRAPHIES

Cécile Hardebolle is a Pedagogical Advisor and Specialist in Learning Sciences in the Teaching Support Centre (CAPE) at Ecole polytechnique fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland; cecile.hardebolle@epfl.ch.

Himanshu Verma, formerly a Scientific Collaborator in the Center for Learning Sciences (LEARN) at Ecole polytechnique federale de Lausanne (EPFL) in Lausanne, Switzerland, is currently an Assistant Professor of Industrial Design and engineering at Delft University of Technology, Landbergstraat 15, 2628 CE Delft, The Netherlands; h.verma@tudelft.nl.

Roland Tormey is a Senior Scientist in Learning Sciences and Head of the Teaching Support Centre (CAPE) at Ecole polytechnique fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland; roland.tormey@epfl.ch.

Simone Deparis is an Adjunct Professor in Mathematics and Executive Director of the Prepaedeutic Center (CePRO) at Ecole polytechnique fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland; simone.deparis@epfl.ch.

How to cite this article: Hardebolle, C., Verma, H., Tormey, R., & Deparis, S. (2022). Gender, prior knowledge, and the impact of a flipped linear algebra course for engineers over multiple years. *Journal of Engineering Education*, 1–21. https://doi.org/10.1002/jee.20467