Problem#solving before instruction for learning linearalgebra in university mathematics

Journal Article

Author(s):
Baumgartner, Vera; Daguati, Simona; Trninic, Dragan; Akveld, Meike; Caspar, Alexander; Hungerbühler, Norbert; Kapur, Manu

Publication date:
2025-08

Permanent link:
https://doi.org/https://doi.org/10.3929/ethz-b-000732811

Originally published in:
Instructional Science, https://doi.org/10.1007/s11251-025-09709-8

Instructional Science
https://doi.org/10.1007/s11251-025-09709-8

ORIGINAL RESEARCH

Problem-solving before instruction for learning linear algebra in university mathematics

Vera Baumgartner[1] · Simona Daguati[1] · Dragan Trninic[1] · Meike Akveld[2] · Alexander Caspar[2] · Norbert Hungerbühler[2] · Manu Kapur[1]

Received: 30 August 2023 / Accepted: 24 February 2025

Abstract

Problem-solving before instruction has been shown to be a more effective learning design than traditional tell-and-practice for several mathematical concepts at the secondary school level. In particular, the more a problem-solving before instruction design follows the productive failure principles, such as comparing and contrasting student-generated solutions, the higher the effect on students' conceptual understanding and transfer. University mathematics education poses several inherent constraints that complicate the implementation of these principles. In the present study, we implemented a problem-solving before instruction design in a university linear algebra course adhering to the productive failure principles as closely as possible. Participation in the preparatory problems was voluntary. We investigated the effect on students' learning over four one-year iterations in a design-based research approach. Compared to the baseline (aggregate of cohorts prior to the intervention), we observed a significant increase in final exam performance for all four cohorts with effect sizes between Cohen's $d = 0.28$ and $d = 0.59$. For students who agreed to further analyses, our results show that up to 16% of the variance in students' performance can be explained by variance in their participation in the problem-solving before instruction design. As our design did not include a control group, we refrain from conclusions regarding any design components that might have caused these effects. However, these results are promising, given that our implementation involved only minor changes to the original course structure and required little extra time for students.

Keywords Higher education · Mathematics education · Productive failure · Design-based research

✉ Vera Baumgartner
vera.baumgartner@gess.ethz.ch

1 Department of Humanities, Social and Political Sciences, ETH Zurich, Zurich, Switzerland

2 Department of Mathematics, ETH Zurich, Zurich, Switzerland

Published online: 21 April 2025

Introduction

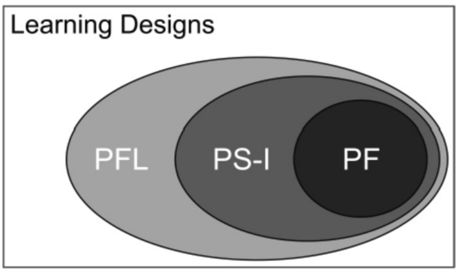
Activation of prior knowledge can be an effective means to prepare students for instruction (e.g., Schwartz & Bransford, 1998). Learning designs that activate students' prior knowledge before formal instruction can lead to deeper conceptual understanding and increased transfer performance than traditional tell-and-practice learning designs. Current research explores the boundary conditions of learning designs structured to activate students' relevant prior knowledge before instruction, such as preparation for future learning (PFL; Schwartz & Martin, 2004), problem-solving before instruction (PS-I; Loibl et al., 2017; Schoenfeld, 1992), and productive failure (PF; Kapur, 2008; Kapur & Bielaczyc, 2012). For mathematical concepts, PS-I has been shown to be effective for a range of topics in primary and secondary school using controlled and classroom experiments (e.g., Kapur, 2014; Loibl & Rummel, 2014a, 2014b; Roll et al., 2009). While some design constraints have already been identified for PS-I, limitations regarding the content and its presentation remain largely unexplored, especially in higher mathematics. In the present study, we investigated whether it is possible to implement a PS-I learning design in a university mathematics course, where the concepts and ideas become more abstract, the mathematical language becomes more formal, and the classrooms become lecture halls.

Design constraints

Schwartz and Bransford (1998) argue that there is an optimal "Time for Telling." That is, students need to be ready to construct knowledge from instruction. If they are not prepared to do so, students may encode new information as mere facts to memorise.

As a means to prepare students to learn from direct instruction, problem-solving is effective (Loibl et al., 2017; Schwartz & Martin, 2004), even if students fail to find the canonical solution to the problems posed (Kapur, 2008; Kapur & Bielaczyc, 2012). PFL, PS-I and PF learning designs focus on the importance of activating students' relevant prior knowledge before the instruction phase. This can be done using general preparatory activities (as in PFL), preparatory problem-solving (PS-I), or preparatory problem-solving activities designed for failure (PF). Following their definitions, PS-I learning designs are a subset of PFL and PF learning designs are in turn a subset of PS-I designs as depicted in Fig. 1. Our classification of these reversed-order learning designs is not exhaustive. For example, DeCaro and Rittle-Johnson (2012) highlighted the exploratory learning aspect of the preparatory activities used in their study where primary school students explored mathematical equivalence problems before instruction on the concept.

Fig. 1 Venn diagram of the relationship of preparation for future learning (PFL), problem-solving before instruction (PS-I), and productive failure (PF) learning designs



In reviews, PS-I and PF designs have been shown to overall positively impact the targeted learning outcomes, especially conceptual understanding and transfer, compared to traditional tell-and-practice designs with instruction prior to problem-solving (I-PS; Darabi et al., 2018; Loibl et al., 2017; Sinha & Kapur, 2019). For procedural knowledge, however, a recent review by Chen and Kalyuga (2020) implies a positive effect of I-PS over PS-I. Sinha and Kapur's (2021) comprehensive meta-analysis shed more light on the comparison of PS-I and I-PS. The authors found a significant pooled effect (Hedge's $g = 0.36$) in favour of PS-I for conceptual knowledge and transfer. For procedural knowledge, they found that PS-I and I-PS lead to similar learning outcomes. Furthermore, Sinha and Kapur were able to partly explain the heterogeneity of previous results with the PF fidelity of the conducted studies. That is, the more closely the PF design principles (Kapur, 2016; Kapur & Bielaczyc, 2012) were implemented in the study design, the higher the positive effect in favour of the PS-I design.

Productive failure design principles

Kapur and Bielaczyc (2012) defined PF as a two-phase learning design that embodies the following four core learning mechanisms: (a) activation and differentiation of prior knowledge, (b) attention to critical features, (c) explanation and elaboration, and (d) organisation and assembly. Based on these mechanisms, Kapur and Bielaczyc (2012) proposed design principles for PF. Split into the two phases of the learning design—the problem-solving phase (Phase 1) and the instruction phase (Phase 2)—the PF principles can be summarized as follows:

Phase 1: Generation and exploration of representations and solution methods for novel problems

(A) that are challenging, not frustrating, and have an affective draw
(B) admit multiple representations and solution methods (RSMs)
(C) activate and differentiate prior and intuitive mathematical knowledge
(D) provide opportunities for explanation and elaboration (e.g., in group work), and a safe space for generation and exploration (e.g., with reassuring prompts)

Phase 2: Consolidation and knowledge assembly

(A) Instruction builds on solutions generated by students (evaluation without judgement)
(B) The lecturer allows for comparing and contrasting of solution methods (e.g., in dialogue)

The core mechanisms of PF through the lens of CID

Recently, Loibl et al. (2024) proposed a framework for the cognitive analysis of composite instructional designs (CID), such as PS-I and PF. Therein, the authors highlight the importance of acknowledging the interplay of the learning mechanisms that are triggered in the phases of an instructional design. The problem-solving phase in PF is designed to allow for the activation of prior knowledge and the recognition of critical features. After the problem-solving phase, students are expected to have noticed critical features of the targeted concept and identified some of their knowledge gaps. If the used materials fail to trigger these mechanisms, students' knowledge after the first phase is affected (intermediate knowledge), which serves as prior knowledge to the second phase. In turn, if students were unable to identify knowledge gaps and recognize critical features, the core mechanisms in the second phase, elaboration and assembly, do not have the desired effects of consolidating students' discoveries. For a successful implementation of PF, it is therefore crucial to carefully design each aspect of the two phases.

Challenges when implementing the PF principles in higher education

In many cases, the setting of the problems used in PF designs relates to students' lives outside of academia to increase affective draw, and thereby, students' engagement (PF principle 1A). For example, in Kapur's (2008) original productive failure study, students are asked to investigate whether a person was speeding based on an eyewitness account before instruction on Newtonian kinematics. In the case of higher mathematics education, the relation of targeted concepts to students' lived experiences is often rather elusive. Particularly in introductory mathematics courses, such as linear algebra, real-world applications of the targeted concepts cannot be used as a stepping stone, because students lack sufficient training in other fields, like physics or computer science, where those principles are applied. The fact that higher mathematical concepts relate less easily to students' prior experiences makes higher mathematical content more abstract for students. We will continue to use the term abstract in that sense throughout the article.

Other studies have highlighted the importance of considering the constraints of the PF principles to evaluate the positive effect of PS-I designs compared to I-PS: Loibl and Rummel (2014a) found that comparing and contrasting student solutions to the canonical solution (PF principle 2B) crucially supports learning during the instruction phase of PS-I. However, discussing student-generated solutions in university courses with hundreds of students is not feasible within a reasonable time frame.

PS-I and PF for university mathematics

Most studies investigating PS-I and PF learning designs for mathematics learning focused on primary and secondary school contents (e.g., DeCaro & Rittle-Johnson, 2012; Fyfe et al., 2014; Glogger-Frey et al., 2017; Loehr et al., 2014; Schwartz & Martin, 2004). In particular, the affordances and constraints of PS-I and PF learning designs for mathematics learning were mainly investigated in the topics variance, average speed, and linear functions, which are usually taught in secondary school (Hartmann et al., 2020; Kapur, 2010, 2011, 2014; Loibl & Rummel, 2014a, 2014b, 2015; Roll et al., 2009; Schalk et al., 2018). Whereas these findings call for more research looking into more varied mathematical content in general, the use of more abstract university mathematics in a PS-I design remains largely unexplored.

In Sinha and Kapur's meta-analysis (2021), five of the 53 studies included investigated the implementation of a PS-I design for mathematics with university students (Belenky & Nokes-Malach, 2012; D.-W. Chen & Catrambone, 2019; Jarosz et al., 2017; Kerrigan, 2018; Newman & DeCaro, 2019). However, only one of these studies focused on higher mathematical content (Chen & Catrambone, 2019) and there, the authors did not find a difference between the PS-I and the I-PS condition for both immediate and delayed post-test scores. Extending the analysis to affective variables, the authors were able to show higher subjective and objective exploration levels in the PS-I condition during the problem-solving phase, which is in line with previous research. In sum, research investigating PS-I in higher mathematical education is scarce and so far, existing results are inconclusive.

In the broader field of university education in science, technology, engineering, and mathematics (STEM), a couple of studies exist that compared PS-I to I-PS. For example, Weaver et al. (2018) showed that undergraduate physics students' conceptual understanding was significantly higher following a PS-I session on the electric potential than after a comparable I-PS session. The authors did not find a difference in students' procedural knowledge. Similarly, Lai et al. (2017) compared engineering students' learning gains after a PS-I and I-PS learning activity on process engineering and found that the PS-I group performed significantly higher than the I-PS group. These studies imply that PS-I could be effectively implemented for higher STEM education. However, both of these studies focused on short interventions (70–80 min) and only evaluated students' short-term learning gains. Long-term implementations of a PS-I learning design in STEM education in general, and mathematics education in particular, remain to be investigated.

Long-term implementations of PS-I and PF

So far, existing research on PS-I and PF designs focused on short-term interventions (Sinha & Kapur, 2021). It is unclear, therefore, how long-term implementations of PS-I or PF designs impact students' learning: for example, if they are implemented in a complete university course stretching over a whole year. When students become familiar with problem-solving before instruction, they possibly learn to profit from the initial problem-solving sequence over time. Glogger-Frey et al. (2017) found, for example, that practising inventing leads to more exploration during self-regulated preparatory problem-solving. On the other hand, the impact of PS-I could diminish over time as students learn that they will receive formal instruction on the targeted concepts independently of their efforts in the problem-solving phase, especially if participation is voluntary.

The difficulties of learning linear algebra

Linear algebra is foundational to most STEM fields and the course is mandatory in many higher education STEM programs. As a means to formalise the algebraic computations from secondary school mathematics, linear algebra concepts provide a basis for multi-dimensional approximations and algorithmic problem solving; for example, to develop machine learning algorithms or to produce linear models of forces in real-world phenomena as used in physics and engineering.

For most students, linear algebra is their first encounter with advanced mathematical abstractions, where computations are generalised and the validity of an argument is shown by formal proof (Tucker, 1993). This transition from secondary school algebra, where the use of visual and symbolic representations is predominant, to a more formal framework poses difficulties for students (Tall, 2008) and can leave them disoriented and confused (Carlson, 1993). As a result, students are often able to perform linear algebra procedures but lack conceptual understanding after a course in linear algebra (Dorier & Sierpinska, 2002).

Several recent studies tried to address this problem. The survey paper by Stewart et al. (2019) gives a concise overview of the field and shows that most of the empirical research on linear algebra learning is of qualitative nature and focuses on students' difficulties with specific concepts, such as span, linear independence, and eigenvalue-theory (e.g., Hannah et al., 2013; Wawro et al., 2012).

One possible reason for students' difficulties is the heavily formalised axiomatic approach dominating linear algebra course curricula that contrasts with the epistemological development of the theory (Dorier, 2000). Students are confronted with the polished theory without feeling the intellectual need for the presented linear algebra concepts, which is crucial for constructing knowledge in linear algebra (Harel, 2000). Harel (2018) argues that the occurrence of intellectual need depends entirely on the prior knowledge students bring to class. In the case of linear algebra, an example for intellectual need could be students experiencing the benefit of using basis vectors to describe any vector in the vector space leading to them wanting to understand what a basis of a vector space is and what constraints a set of vectors must satisfy in order to form a basis. Similarly, Nathan (2012) argues that if students' prior knowledge is activated with targeted activities, their sense-making within a traditional mathematics lecture can be supported. Thus, activating students' prior knowledge before instruction in a PS-I learning design could increase students' understanding of linear algebra considerably.

The present study

In the present study, we implemented a PS-I design in a university linear algebra course adhering to the PF principles as closely as possible under the constraints of university education. We opted for a design-based research approach (DBR; Cobb et al., 2003) implemented over four iterations to increase PF fidelity over the course of the study. We aimed to answer the following research questions:

RQ1. How does the introduction of preparatory exercises to a university linear algebra course influence students' performance in the final exam?
RQ2. To what extent does students' participation in preparatory exercises predict their performance in the final linear algebra exam and tests on conceptual understanding of linear algebra?

To answer our research questions, we developed ten preparatory exercise sheets for a one-year linear algebra course in the mechanical engineering program at a technical university in Switzerland, in which the enrolled students could participate voluntarily *prior* to the corresponding lecture. Additionally, students were invited to participate in two end-of-semester assessments that tested their conceptual understanding of the targeted concepts. The course structure including lectures, exercise classes, and the final exam with the additional elements of the intervention are displayed in Fig. 2. The figure depicts eight example cycles of (preparatory exercises before) lecture before exercise class out of the total 28 cycles in the course.

For each of the ten targeted concepts, we thus changed the instructional design from instruction with subsequent homework exercises to PS-I, while preserving the subsequent homework exercises. In the first phase of our PS-I design, students solved preparatory exercises as homework that required prior knowledge of previously covered concepts in the course as well as mathematics from secondary school. By engaging with the preparatory exercises, students' relevant prior knowledge was activated, and students were given the opportunity to recognize relevant features of the

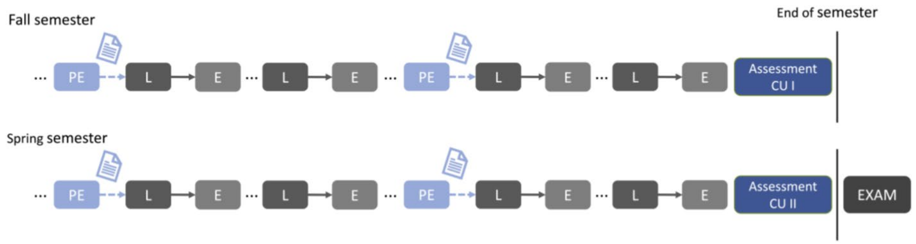


Fig. 2 Original course structure with lectures, homework exercises, and the final exam with intervention elements preparatory exercises and assessments on conceptual understanding added. *CU* Conceptual understanding, *E* homework exercise, *L* lecture, *PE* preparatory exercise

targeted concepts and to explain and elaborate on their thinking (Kapur & Bielaczyc, 2012). After the problem-solving phase, students might have identified knowledge gaps, gained a disorganised conceptual understanding of the targeted concept, and recognized several relevant features. The instruction phase of our PS-I design consisted of the course lecture in a traditional lecture format, where the lecturer presented the contents in front of the class. To build on students' intermediate knowledge after the problem-solving phase, the lecturer discussed the preparatory exercise in the beginning of the lecture in each PS-I cycle. This discussion allowed students to organise and assemble the relevant features to a more complete conceptual understanding of the targeted concept.

Our DBR project encompassed four iterations, from the 2018 Fall semester until the end of the 2022 Spring semester. Throughout, we collected data on students' performance in the end-of-semester assessments and the end-of-year exam, as well as on their participation in the preparatory exercises. Furthermore, we collected aggregated data on students' performance in the other first-year courses for all cohorts. Based on previous research on problem-solving before instruction as discussed in the previous sections, we formulated the following hypotheses:

H1a The introduction of preparatory exercises to the linear algebra course would increase the average performance in the final exam in comparison to cohorts prior to the intervention.

H1b The introduction of preparatory exercises to the linear algebra course would lower the percentage of students failing the final exam in comparison to cohorts prior to the intervention.

H2a Students' participation in the preparatory exercises would positively predict their final exam performance.

H2b The more students would participate in the preparatory exercises, the higher their likelihood to pass the final exam would be.

H3 Students' participation in the preparatory exercises would positively predict their conceptual understanding of linear algebra.

Methods

Participants

Our sample consisted of first-year mechanical engineering students at the Federal Institute of Technology (ETH) Zurich, a technical public university in Switzerland that offers study programs in STEM. ETH Zurich requires a Swiss Matura (or comparable secondary-school education) for enrolment without any additional selection process.

The grades of all first-year mechanical engineering students who participated in the final exam of the linear algebra course from Summer 2014 to Summer 2022 (baseline and intervention years) were anonymously analysed (N = 4463). The grades of 464 students who took part in the linear algebra exam in the Summer of 2017 were excluded due to deviances in the examination process (Online Resource 1). Compared to the other baseline cohorts, Summer 2017 was a clear outlier, potentially due to the change in exam format. We obtained a final sample size of N = 3999 (11.53% Women).

For further analysis that could not be conducted completely anonymously, we recruited first-year mechanical engineering students who were enrolled in the linear algebra lecture in the four academic years between Fall 2018 and Spring 2022 to volunteer their data. We highlight that participation in the intervention was independent of students' sharing their data for further analysis. In this part, we excluded students repeating the course from analysis, as their learning gains cannot be attributed to a single academic year. The sample sizes are depicted in Table 1.

An a priori power analysis was conducted using G*Power version 3.1.9.6 (Faul et al., 2009). Results indicated the required sample size to achieve 80% power for detecting a small effect, at a significance criterion of $\alpha = 0.05$, was $N = 1845$. Thus, the obtained sample size of $N = 3999$ is adequate for our planned comparison. For the detailed analysis, the results indicated that our sample sizes are satisfactory for detecting an effect of $R^2 = 0.05$ at 80% power for our planned regressions.

Materials

For every academic year between 2018 and 2022, we collected data on students' participation in the preparatory exercises (independent variable), their performance in tests on conceptual understanding, and their end-of-year exam performance (dependent variables). The changes made to the course structure are detailed in the following.

Table 1 Sample sizes for in-depth analysis for intervention cohorts

Academic year	Original sample	Repeating students	Final sample	% Women
2018/2019	344 (72%)	58	**286**	**12.94%**
2019/2020	417 (85%)	57	**360**	**14.17%**
2020/2021	441 (95%)	85	**356**	**13.20%**
2021/2022	403 (78%)	90	**313**	**13.42%**

The percentage of the sample of students who consented to the extended analysis with respect to all students who partook in the end-of-year exam are given in parentheses

Course structure

Before 2018, the course consisted of a two-hour lecture and a one-hour exercise class per week with an end-of-year exam. After each lecture, students were given a homework exercise sheet to practice the content of the lecture. The homework sheets could voluntarily be handed in to the tutors of the exercise classes for feedback. In the exercise classes, tutors revisited key concepts from the lecture and discussed the solutions to the homework sheets.

During our intervention from 2018 to 2022, ten preparatory exercise sheets and two tests on conceptual understanding were added to the course structure (Fig. 2). The preparatory exercises were not regularly distributed over the semester, as they were administered before the corresponding lecture where the targeted concept was introduced. Throughout, the course was conducted in German.

The lecturer remained the same throughout the baseline and the intervention years. Except for the materials added for the intervention, which consisted of the preparatory exercises and the tests on conceptual understanding (Fig. 2), the course materials remained the same throughout the baseline and the intervention years.

Preparatory exercises

For the development of the preparatory exercises, we analysed the final exam performance two years prior to our interventions and identified ten core concepts foundational to linear algebra yet challenging for the students. The ten targeted concepts are listed in Table 2. For each of these concepts, we developed a preparatory exercise sheet that built on students' prior knowledge and experiences in mathematics from secondary school or from previous linear algebra course content to establish relevant features of the novel concept (PF principle 1C). The exercises were designed for students to explore novel mathematical phenomena and to formulate and elaborate on possible explanations using familiar mathematical language (PF principle 1D). The exercises prompted the students to notice novel features of and connections between already-known mathematical objects and did not include any new definitions or statements.

In contrast to the regular homework sheets after formal instruction, we designed the preparatory exercises to regularly prompt the students to explain their thinking explicitly in the exercise instruction (e.g., "Justify your answer.", "Explain."; PF principle 1D).

Participation in the preparatory exercises was voluntary. If students decided to participate, they were required to hand in their solutions prior to the targeted lecture to ensure engagement prior to instruction. The preparatory exercise sheets are included in Online Resource 2 (German originals and English translations).

Table 2 The ten core linear algebra concepts targeted by preparatory exercises

1st semester	2nd semester
PE1. Rank of matrices	PE6. Eigenvalues and Eigenvectors
PE2. Geometric interpretation of linear systems of equations	PE7. Inner product in vector spaces
	PE8. Orthogonal projection
PE3. Non-commutativity of matrix multiplication	PE9. Kernel and image of a linear map
PE4. Determinant of matrices	PE10. Orthogonal diagonalization of matrices
PE5. Bases of a vector space	

PE Preparatory exercises

The preparatory exercises were designed to trigger two core mechanisms of PF, activation of prior knowledge and recognition of critical features. For example, the first preparatory exercise included seven contrasting cases of linear systems of equations. The contrasting cases were designed for students to activate their prior knowledge of the meaning of equations. Moreover, students were expected to recognize that an equation which is a linear combination of others in the same system does not add any information and can therefore be disregarded. The same idea applies to the rank of a matrix, where only linearly independent rows (or columns) are counted. In the instruction phase, the lecturer built on students' intermediate knowledge gained in the preparatory exercise by extending the concept from linear systems of equations to the rank of matrices, which was the targeted knowledge in this PS-I cycle.

In general, the lecturer discussed a possible solution to the preparatory exercises and highlighted common misconceptions and potential pitfalls at the beginning of the subsequent lecture in each PS-I cycle.

Tests on conceptual understanding

At the end of each of the two semesters, students were tested on their conceptual understanding of the targeted concepts (Table 2). Participation in the tests was voluntary. In the first and second cohorts, the tests included three multiple-choice items with four answer options and one correct choice of increasing difficulty per concept. In the third and fourth cohorts, the tests included two problems with four true/false items per concept of varying difficulty. In all cases, each item was graded as correct/incorrect. The tests were developed in close collaboration with the lecturer of the course, who also co-authored the present article. The items were chosen based on conceptual learning goals set by the lecturer and the distractors were chosen based on common misconceptions as observed in linear algebra exams from previous years.

The second version of the tests are included in Online Resource 3 (German originals and English translations) and they were piloted with the tutors of the course and a group of second-year students (Online Resource 1).

Except for Spring semester 2020 (Covid-19 restrictions), the end-of-semester tests on conceptual understanding were administered on-site. Starting from Spring semester 2020, they were administered online during an exercise class under supervision by the tutors. Students were also able to log-in from outside the university. However, the tests included a final question on whether students used forbidden additional resources.

Our variables operationalising performance in the end-of-semester tests on conceptual understanding were obtained as follows. Each item in the tests was coded as either true or false. The performance was then defined to be the total number of correctly answered items. For our analysis, we converted students' performance in the assessments on conceptual understanding into percentages. Students who exceeded the time limit, did not finish the test, or admitted to having used forbidden additional resources were excluded from the analysis.

Participation variable and bonus grade points

Participation in the preparatory exercises and the tests on conceptual understanding was voluntary and awarded with bonus grade points for their final grade. In total, students were able to attain P out of 12 points for participation that translated to min

(0.25, P/9*0.25) bonus grade points in the exam, where 0.25 grade points are equivalent to 5%. The bonus points were awarded for participation regardless of students' performance. Students' final grade in the linear algebra course was determined as the sum of their grade in the end-of-year exam and the bonus grade points they achieved.

In case of the preparatory exercise sheets, students were required to hand in their solutions before the corresponding lecture to be awarded a bonus point. The hand-ins were evaluated by tutors and rewarded with one point for participation if students provided answers for at least 90% of the preparatory exercise. The correctness of the answers did not influence grading. To ensure uniform grading across all tutors, the tutors were trained by the authors at the beginning of the semester based on detailed criteria and examples. Over the course of the semester, tutors discussed ambiguous cases with the authors via an online platform. For the tests on conceptual understanding, students were awarded a bonus point if they handed in their answers on time regardless of their performance.

In our analysis, we operationalised participation in the intervention as the total number of preparatory exercises the students participated in ranging from 0 to 10. Thus, our participation variable consists of the bonus points the students achieved without the tests on conceptual understanding. To analyse students' performance in the end-of-year exam, we disregarded the bonus grade points.

End-of-year exams

The development and grading of the exam were conducted independently of this study and in the same manner as in years prior to the intervention. The exam, which covered the same course materials as in the baseline years, was developed by the coordinator of the course, a doctoral student in mathematics who was not part of this research group, in the same way prior to and during the intervention. The lecturer reviewed the exam and set the grading scale before the exam took place based on the his judgement of the difficulty of the exam and his previous experience as a lecturer (> 12 years teaching this course). As all previous exams by the lecturer are available to students to study, the passing grade tends to increase over the years. All end-of-year exams were administered on-site and solved on paper.

Throughout the baseline and the intervention years, the end-of-year exams targeted both procedural knowledge and conceptual understanding with a ratio of around 2:1. Up until Summer 2017, the exams consisted of 10 multiple-choice questions and four open questions and from Summer 2017 onwards, the exams consisted of 20 multiple-choice questions and three open questions. The ratio of the targeted knowledge types remained comparable. Because the coordinator of the course was a different doctoral student every year, the foci of the content of the exams vary. A link to all end-of-year exams is provided in Online Resource 3.

To ensure reasonable comparability between students' exam performance over the years, exam performance was measured using the grades awarded by the lecturer. For the baseline cohorts, we thus considered their final course grade and for the experimental cohorts, we considered their exam grade disregarding any achieved bonus grade points. The university uses the Swiss grading system, ranging from 1 (not measurable) over 4 (pass) to 6 (excellent) and with quarter-grade increments (ETH Zurich, 2025).

Performance in other first year courses

To get an impression of the overall performance of the cohorts, we looked at trends in our students' exam performance in analysis, their other first-year mathematics course, as well as their overall performance in all first-year courses. In addition to linear algebra and analysis, the first-year courses are mechanics, machine elements and innovation process, engineering materials and production, chemistry, and computer science. Their overall performance determines whether students are allowed to proceed to the second year of the mechanical engineering program and is computed as a weighted average of the exam performance in all first-year courses. This data was obtained from the university directly. The data was anonymized and given to us as an aggregate. We excluded the cohort 2016/2017 due to deviances in the linear algebra exam that impacted the validity of a comparison to other cohorts. Statistical analysis confirmed this cohort as a clear outlier (Online Resource 1).

Implementation of PF principles

We implemented the four PF principles for the problem-solving phase as follows. The preparatory exercises were developed in collaboration with the lecturer of the linear algebra course and refined based on students' feedback. We thereby ensured that the exercises were challenging, but not frustrating for the students. Moreover, if the students were completely stuck in an exercise, they were invited to approach their tutor for help. For affective draw, we chose graphical examples and applications in physics, if possible, to raise students' engagement. We are thus confident that we were able to follow PF principle 1A.

The preparatory exercises allowed the students to explore the solution space by offering them the opportunity to discover features of novel concepts before formal instruction. However, we were unable to develop exercises that truly admit multiple representations and solution methods as demanded by PF principle 1B, as most of the exercises pose open questions that still target one specific answer.

The exercises build on students' prior knowledge from secondary school and the preceding part of the linear algebra course. To solve the exercises, the students needed to activate and differentiate their prior knowledge as postulated in PF principle 1C.

PF principle 1D asks for opportunities for explanation and elaboration as present in group work, and a safe space for generation. As the preparatory exercises were given to the students as facultative homework exercises, we did not control whether and to what extent the students worked in groups. However, we frequently reminded the students that their performance in the preparatory exercises was of no importance and that they should "just try". Thereby, we aimed to reassure the students that potential failure in solving the exercises was part of the learning process.

In the instruction phase of our design, the lecturer was unable to build directly on students' solutions due to the vast number of students in the course. However, the lecturer discussed the preparatory exercises at the beginning of the subsequent lecture and elaborated on potential pitfalls and common misconceptions. The lecturer furthermore emphasized critical features of the novel concept and referred frequently to the preparatory exercises during instruction. Our approach was similar to the comparing of typical student solutions to the canonical one in Loibl and Rummel (2014a). We thereby enabled the students to compare and contrast their solutions to the canonical ones. During the lecture and the following session with their tutor, the students were encouraged to ask questions about the exercises to forgo any confusion. We thus adapted PF principles 2A and 2B due to the size of the course.

Procedure

Students enrolled in the linear algebra course were invited to participate in 10 preparatory exercises and two end-of-semester assessments in addition to the original course structure as detailed in the previous section. We estimated that this additional coursework amounted to around 10 extra hours over the year. This is an estimate of an average workload per student and the individual workload per student presumably varied.

Research design

The study consisted of four DBR iterations (Cobb et al., 2003). After each iteration, we analysed students' participation in the preparatory exercises and their performance in the end-of-semester assessments of conceptual understanding. After the first and second years of the study, we revised the materials to account for conceptual understanding items in the assessments where students performed unexpectedly low and increase fidelity to the PF design principles. In these cases, we included more contrasting cases, refined the wording, or changed the framing of an exercise. For example, we included contrasting matrices in the third preparatory exercise after Year 1 to allow students to discover the difference between diagonal matrices and multiples of the unit matrix within the exercise after discovering that students struggled to differentiate between them in the assessment of conceptual understanding. We thus allowed for a better opportunity for students to differentiate their knowledge (PF principle 1C).

In addition to the data we collected throughout the year, we based these revisions on informal feedback from students, tutors, and the lecturer. The tutors collected students' remarks and questions to the preparatory exercises and the lecturer collected students' questions raised during the lecture. This approach allowed us to determine preparatory exercises that were too challenging for students, and we adapted the exercises accordingly. For example, we changed the framing of a proof in preparatory exercise 8 after Year 2. Originally students struggled to find the proof on their own. We transformed the exercise into a more guided approach, where students had to identify the correct order of the statements in the canonical proof. We thereby strengthened the fidelity of the exercise to PF principle 1A stating that the exercises should be challenging, but not frustrating. This example illustrates our approach to balance guidance in the exercises with opportunities exploration to account for the abstract nature of the subject.

Moreover, the tutors reported how well students in their group solved the exercises in regular meetings with the research group. If there were parts of an exercise that many students had neglected, we adapted the instruction or the answer format. For example, many students did not answer all questions posed in preparatory exercise 6 in Year 1. We changed the answer format from open text to a table, such that missing answers were directly obvious to the students. In Year 2, students were thus given the opportunity to discover more aspects of the targeted concept as some questions could not as easily be overlooked.

The targeted concepts as well as the general structure of the exercises remained unchanged. For the fourth iteration, all exercises remained the same as in the third iteration.

Data analysis

To answer our research questions, we split the analysis into two parts. First, we compared the exam performance of our targeted cohorts to the exam performance of students of previous years as a baseline. For this analysis, we used anonymized data from all students who participated in the linear algebra exams in the summer sessions 2014–2022 and we conducted an ANOVA with planned contrasts using JASP (Version 0.17.3 Intel; JASP Team, 2023). All plots were produced in RStudio (Version 2023.6.1.524; Posit team, 2023) using the ggplot2 package (Wickham, 2016).

In the second analysis, we investigated the effect of participation in the preparatory exercises on students' performance in the end-of-semester tests of conceptual understanding and the final exam. For this second part, we used the data of a subsample of students in the intervention cohorts that volunteered for the analysis and excluded all repeating students. We conducted linear regressions with performance as the dependent variable for each test and the final exam in every intervention year. Furthermore, we conducted logistic regressions with passing the final exam as the dependent variable.

To strengthen our results, we also conducted Bayesian linear regressions with performance as the outcome and performed a sensitivity analysis. In our sensitivity analysis, we investigated whether our data included any highly influential cases or relevant outliers and whether the classification of our independent variable (participation in the preparatory exercises) as a continuous instead of a ratio variable influenced the model significantly.

Results

RQ 1: Experimental cohorts' exam performance compared to previous cohorts

Descriptive statistics of the exam performance are displayed in Table 3. Detailed descriptive statistics of the baseline are included in Online Resource 1.

The effect of the intervention on students' grades

An ANOVA showed that there was a significant effect of the intervention years on exam performance, $F(4) = 58.37$, $p < 0.001$, with an overall effect size of $\eta^2 = 0.06$. Planned contrasts using Tukey's method between the separate intervention years and the baseline

Table 3 Descriptive statistics of the exam performance of the intervention cohorts Year 1 to Year 4 compared to the baseline

	Exam performance (exam grade)				
	N	*Mdn*	*M*	*SD*	Pass in %
Baseline	2045	4.00	3.84	1.17	55.84
Year 1	480	4.50	4.27	1.03	70.00
Year 2	492	4.75	4.52	1.07	75.81
Year 3	463	4.25	4.15	1.04	65.66
Year 4	519	4.75	4.45	1.13	71.29

The exam performance was measured using Swiss grades (Min: 1, Max: 6, Pass: 4). The baseline consists of an aggregate of four cohorts prior to the intervention

Table 4 Planned contrasts of ANOVA of intervention years on exam performance (grade)

	Mean Difference	SE	df	t	Cohen's *d*	*p* (tukey)
Comparison to baseline						
Year 1	0.42	0.06	3994	7.43	0.38	< .001
Year 2	0.67	0.06	3994	11.85	0.59	< .001
Year 3	0.31	0.06	3994	5.39	0.28	< .001
Year 4	0.60	0.06	3994	10.92	0.54	< .001

Cohen's *d* does not correct for multiple comparisons

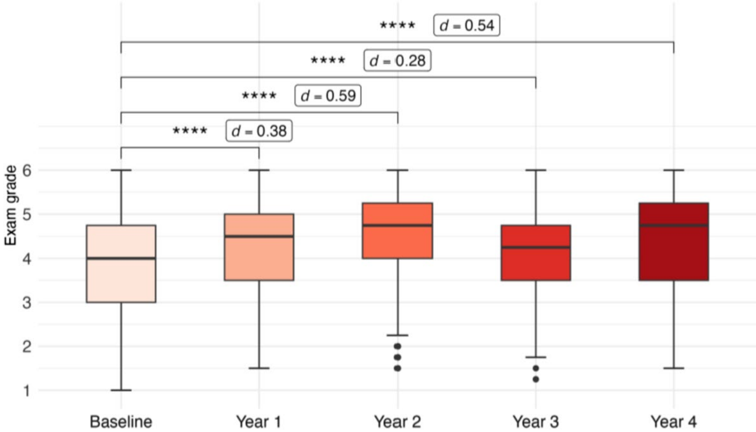


Fig. 3 Boxplot of exam performance baseline and intervention years Year 1 to Year 4. ****p < .0001; Effect sizes indicated with Cohen's *d*

revealed that the exam performance in each year of the intervention differed significantly from the baseline, all $p < 0.001$, with effect sizes of Cohen's *d* between 0.28 and 0.59. The results of the four planned contrasts are listed in Table 4 and summarized in Fig. 3.

The results provide evidence for our hypothesis H1a that the addition of preparatory exercises to the course structure significantly improves students' exam performance in linear algebra.

The effect of the intervention on the percentage of passing grades

An ANOVA showed that there was a significant effect of the intervention years on the percentage of passing grades, $F(4) = 27.32$, $p < 0.001$, with an overall effect size of $\eta^2 = 0.03$. Planned contrasts using Tukey's method between the separate intervention years and the baseline revealed that the exam performance in each year of the intervention differed significantly from the baseline, all $p \leq 0.001$, with effect sizes of Cohen's *d* between 0.21 and 0.42. The results of the four planned contrasts are listed in Table 5.

These results provide evidence for our hypothesis H1b that the addition of preparatory exercises to the course structure significantly increases the percentage of passing grades.

Table 5 Results of planned contrast in ANOVA of intervention years on percentage of passing grades in the final exam

	Mean Difference	SE	df	t	Cohen's *d*	*p* (tukey)
Comparison to baseline						
Year 1	0.14	0.02	3994	5.86	0.30	< .001
Year 2	0.20	0.02	3994	8.35	0.42	< .001
Year 3	0.10	0.02	3994	4.00	0.21	.001
Year 4	0.15	0.02	3994	6.60	0.32	< .001

Cohen's *d* does not correct for multiple comparisons

Comparison of trends in exam performance of first-year mathematics courses

The descriptive statistics of the weighted average of all first-year exams (Block) and of the linear algebra and the analysis exams for the baseline and the experimental cohorts are included in Table 6 and displayed in Fig. 4. In contrast to the linear algebra course, where the lecturer was the same for all cohorts and there no bonus grade points were awarded in the baseline cohorts, the lecturers of the other courses varied over the years. Moreover, some of these lecturers awarded bonus grade points and others did not. Regardless of the course, the bonus always lay between 0 and 0.25 grade points. As we only had access to the aggregated data for all courses except linear algebra, we could not deduct the bonus grade points for these courses. For better comparability between the courses, we therefore decided to include the grade points in this analysis for the linear algebra course as well.

We observed that up until the fourth baseline cohort (B4), students' exam performance in the two mathematics courses were lower or equal to their overall performance. During the intervention, exam performance in analysis lay close to overall performance, whereas exam performance in linear algebra increased noticeably and remained distinctly above overall exam performance. However, as the lecturers of some other first-year courses, including Analysis, changed over the years, we are cautious to directly compare the exam grades. Still, the data implies that our intervention had an

Table 6 Descriptive statistics of first-year exam performance for baseline and experimental cohorts

Cohort	Block			Linear Algebra			Analysis		
	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
B1	529	4.00	0.82	531	3.89	1.10	532	3.49	1.20
B2	510	4.02	0.92	513	3.84	1.29	513	3.74	1.39
B3	539	3.98	0.96	543	3.67	1.20	542	3.73	1.36
B4	534	4.05	0.90	534	3.98	1.05	536	4.07	1.08
Year 1	502	4.27	0.88	505	4.48	1.05	503	4.20	1.07
Year 2	509	4.24	0.88	509	4.75	1.04	509	4.33	1.04
Year 3	480	4.14	0.98	484	4.38	1.05	484	4.02	1.26
Year 4	535	4.32	0.91	537	4.66	1.12	535	4.40	1.16

The number of students vary between Linear Algebra, Analysis and the aggregate of all first-year-exams (Block), because some students quit the exam session between exams

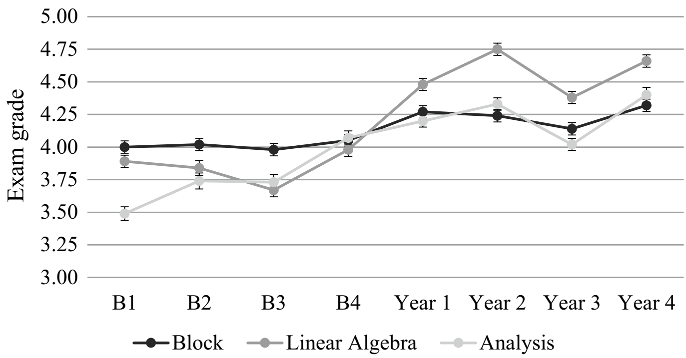


Fig. 4 Exam performance in first-year mathematics courses Linear Algebra and Analysis and overall performance (Block) in the baseline and experimental cohorts (B1–B4 and Years 1–4; with standard error and including bonus grade points achieved by additional course work in both courses)

impact on students' linear algebra exam performance that cannot be solely attributed to a performance difference to the cohorts before the intervention.

RQ2: The influence of participation in preparatory exercises on performance

Descriptive statistics of the performance and participation variables for all iterations (Year 1 to Year 4) are displayed in Table 7 and the corresponding distributions are displayed in Fig. 5. Even though the participation variables were recorded on an ordinal level, we entered them as a covariate into our linear regression models. In Online Resource 4, we include the results of the corresponding sensitivity analysis showing that if included as a factor, the size of the effect of participation on performance is comparable. Furthermore, the residuals of some variable pairings showed positive autocorrelation (Test CU I and II with participation in Year 4). However, according to Sharma (1987) and our power analysis, our sample sizes were still large enough to observe an effect and thus, we proceeded with the planned linear regressions. The results of all Durbin-Watson tests are included in Online Resource 4. In addition, we performed Bayesian linear regressions due to the skewed distributions of our participation variables.

Linear regressions for each of the intervention years showed that the number of preparatory exercises the students participated in significantly predicted their exam grade with coefficients β between 0.16 and 0.31. This means that for every preparatory exercise the students participated in, they advanced their grade by a maximum of 0.31 grade points. The corresponding adjusted R^2 values were between 0.03 and 0.16. As some of the variables have skewed distributions, we also performed Bayesian linear regressions. The Bayesian factors BF_{10} lay between 11.80 and 10^9 and the corresponding R^2 values were between 0.03 and 0.16. This means that at least 3% of the variance in the exam performance can be explained by students' participation in the preparatory exercises. These results provide evidence for our hypothesis H2a that participation in the preparatory exercises positively predicts final exam performance. All results are summarized in Table 8.

Logistic regressions for each iteration (Year 1 to Year 4) showed that students' odds to pass the final exam significantly increased by 103%, 49%, and 58% for Year 1, Year 2, and Year 3, when they participated in one preparatory exercise. In Year 4, students' odds increased by 21% to pass the exam per preparatory exercise; at $p = 0.1$, this result

Table 7 Descriptive Statistics of performance variables and participation variables for intervention cohorts

	N	*Mdn*	*M*	*SD*
Performance variables				
Exam				
Year 1	269	4.50	4.50	0.96
Year 2	332	5.00	4.64	1.02
Year 3	294	4.50	4.32	0.97
Year 4	273	5.00	4.71	1.04
Assessment CU I Fall				
Year 1	271	0.79	0.77	0.15
Year 2	336	0.71	0.72	0.16
Year 3	263	0.78	0.78	0.10
Year 4	283	0.78	0.78	0.08
Assessment CU II Spring				
Year 1	266	0.67	0.62	0.17
Year 2	260	0.60	0.62	0.19
Year 3	178	0.65	0.68	0.13
Year 4	195	0.68	0.69	0.13
Participation variables				
Participation PE Total				
Year 1	269	8	8.27	1.22
Year 2	332	9	8.91	1.36
Year 3	294	8	8.35	1.51
Year 4	273	9	8.71	1.19
Participation PE Fall				
Year 1	271	5	4.76	0.54
Year 2	336	5	4.78	0.58
Year 3	263	5	4.71	0.62
Year 4	283	5	4.77	0.65
Participation PE Spring				
Year 1	266	5	4.77	0.52
Year 2	260	5	4.84	0.48
Year 3	178	4	3.91	0.96
Year 4	195	4	4.04	0.86

Exam performance was measured using Swiss grades (1–6) and performance in the CU assessments are indicated in percentages. Students who did not finish the CU tests or used additional resources were excluded from analysis. Participation points were measured on a scale from 0 to 10. *CU* Conceptual understanding, *PE* preparatory exercises

was not significant. These results mostly support our hypothesis H2b that participation in the preparatory exercises increases the likelihood to pass the final exam. All results of the logistic regressions are displayed in Table 9.

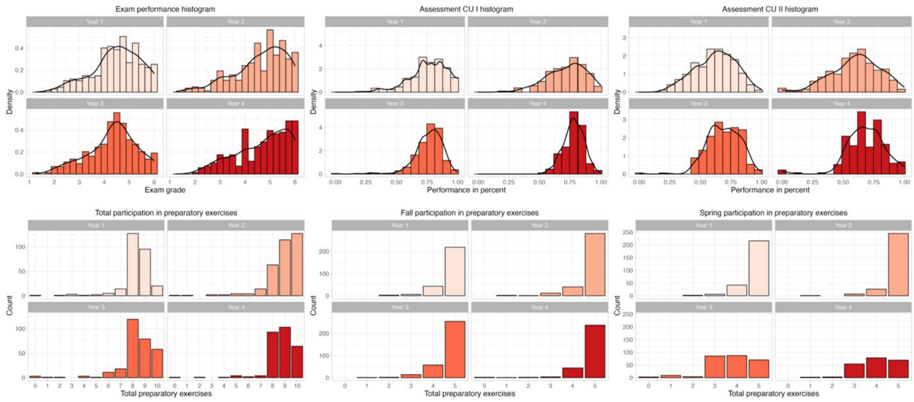


Fig. 5 Distribution of performance variables (top row) and participation variables (bottom row; total, first semester Fall, second semester Spring). *CU* Conceptual understanding. The Effect of Participation in Preparatory Exercises on Exam Performance

The effect of participation in preparatory exercises on end-of-semester assessments

Linear regressions showed that in the first two iterations (Year 1 and 2), the number of preparatory exercises the students participated in during the semester (participation in PE) significantly predicted their performance in the end-of-semester assessments on conceptual understanding (CU) with standardized coefficients β between 0.18 and 0.21. In the third iteration, the linear regression was significant only in the case of the first assessment of conceptual understanding with a standardized coefficient $\beta = 0.13$. In the fourth iteration, none of the predictions were significant. As the participation variables have highly skewed distributions, we also performed Bayesian linear regressions. Confirming the results of the parametric linear regressions, we obtained Bayesian factors greater than 1 for both tests during the first two iterations with $BF_{10} = 11.39$, $BF_{10} = 216.29$ and $R^2 = 0.03$, $R^2 = 0.05$ (Assessment CU I; Year 1, Year 2) and $BF_{10} = 8.79$, $BF_{10} = 27.66$ and $R^2 = 0.03$, $R^2 = 0.04$ (Assessment CU II; Year 1, Year 2). In the third iteration (Year 3), only the Bayesian factor of the regression predicting the performance in assessment CU I favoured our hypothesis very slightly with $BF_{10} = 1.04$ and $R^2 = 0.02$. In the fourth iteration (Year 4), both Bayesian factors favoured the null model. The results of Year 1 and Year 2 support our hypothesis H3 that participation in the preparatory exercises positively predicts conceptual understanding of the targeted concepts. In Year 3 and Year 4, we did not find evidence for this hypothesis. All results are summarized in Tables 10 and 11.

Discussion

To assess the efficacy of our intervention, we compared students' performance in the final exam to an aggregate of the final exam performance of four previous years. We found that in the first year, the mean exam performance increased by 0.42 grade points (8.4%) with a significant effect of $d = 0.38$, and 14% more students passed the exam. The increase in performance is in line with the effect size found by Sinha and Kapur (2021) of Hedge's $g = 0.36$ favouring PS-I over I-PS for conceptual understanding and transfer.

Table 8 Results of model comparisons of frequentist and Bayesian linear regressions to null model for participation in preparatory exercises predicting exam performance for all intervention years Year 1 to Year 4

		ANOVA			Regression Coefficients							Bayesian	
	Adj.R^2	F	df	p	Effects	β	SE	Std.β	95% CI	t	p	BF_{10}	R^2
Iteration													
Year 1	0.16	49.62	(1, 267)	< .001	Intercept	1.93	0.37		[1.21, 2.66]	5.26	< .001	> 10^8	0.16
					Participation in PE	0.31	0.04	0.40	[0.22, 0.40]	7.04	< .001		
Year 2	0.11	41.53	(1, 330)	< .001	Intercept	2.40	0.35		[1.71, 3.09]	6.84	< .001	> 10^7	0.11
					Participation in PE	0.25	0.04	0.33	[0.17, 0.33]	6.44	< .001		
Year 3	0.11	34.73	(1, 292)	< .001	Intercept	2.57	0.30		[1.97, 3.16]	8.51	< .001	> 10^6	0.11
					Participation in PE	0.21	0.04	0.33	[0.14, 0.28]	5.89	< .001		
Year 4	0.03	9.55	(1, 271)	.002	Intercept	3.31	0.46		[2.41, 4.21]	7.23	< .001	11.80	0.03
					Participation in PE	0.16	0.05	0.18	[0.06, 0.26]	3.09	.002		

Exam performance was measures using Swiss grades (1–6). *PE* Preparatory exercises

Table 9 Results of model comparisons of logistic regressions to null model for participation in preparatory exercises predicting passing the final exam

	Nagelkerke R^2	Chi-Squared Test			Regression Coefficients						
		χ	df	p	Effects	β	SE	e^{β}	95% CI e^{β}	Wald Stat	p
Iteration											
Year 1	0.16	29.18	267	< .001	Intercept	4.45	1.33	85.63	[22.65, 323.76]	11.21	< .001
					Participation in PE	0.71	0.16	2.03	[1.73, 2.39]	18.95	< .001
Year 2	0.08	18.11	330	< .001	Intercept	2.12	0.87	8.33	[3.49, 19.89]	5.87	.020
					Participation in PE	0.40	0.10	1.49	[1.35, 1.65]	15.91	.001
Year 3	0.12	24.77	292	< .001	Intercept	2.87	0.91	17.64	[7.10, 43.82]	9.93	.002
					Participation in PE	0.46	0.11	1.58	[1.42, 1.77]	17.93	< .001
Year 4	0.02	2.66	271	.1	Intercept	0.28	0.99	1.32	[0.49, 3.56]	0.08	.780
					Participation in PE	0.19	0.11	1.21	[1.08, 1.35]	2.73	.1

The final exam performance was measured using Swiss grades (1–6). Students passed the exam if they achieved a grade greater or equal to 4. *PE* Preparatory exercises

Table 10 Results of model comparisons of frequentist and Bayesian linear regressions to null model for participation in preparatory exercises in the first semester predicting performance in the first assessment of conceptual understanding for all intervention years Year 1 to Year 4

		ANOVA			Regression Coefficients							Bayesian	
	Adj.R^2	F	df	p	Effects	β	SE	Std.β	95% CI	t	p	BF_{10}	R^2
Iteration													
Year 1	0.03	9.47	(1, 269)	.002	Intercept	0.51	0.08		[0.35, 0.68]	6.26	< .001	11.39	0.03
					Participation in PE	0.05	0.02	0.18	[0.02, 0.09]	3.08	.002		
Year 2	0.05	15.91	(1, 334)	< .001	Intercept	0.44	0.07		[0.30, 0.58]	6.21	< .001	216.29	0.05
					Participation in PE	0.06	0.01	0.21	[0.03, 0.09]	3.99	< .001		
Year 3	0.01	4.31	(1, 257)	.040	Intercept	0.68	0.04		[0.60, 0.77]	15.28	< .001	1.04	0.02
					Participation in PE	0.02	0.01	0.13	[0.01, 0.04]	2.08	.040		
Year 4	0.00	0.09	(1, 281)	.760	Intercept	0.77	0.04		[0.69, 0.84]	20.80	< .001	0.14	0.00
					Participation in PE	0.002	0.01	0.02	[-0.01, 0.02]	0.31	.760		

PE Preparatory exercises

Table 11 Results of model comparisons of frequentist and Bayesian linear regressions to null model for participation in preparatory exercises in the second semester predicting performance in the second assessment of conceptual understanding for all intervention years Year 1 to Year 4

		ANOVA			Regression Coefficients							Bayesian	
	Adj.R^2	F	df	p	Effects	β	SE	Std.β	95% CI	t	p	BF_{10}	R^2
Iteration													
Year 1	0.03	8.90	(1, 264)	.003	Intercept	0.34	0.09		[0.16, 0.53]	3.62	< . 001	8.79	0.03
					Participation in PE	0.06	0.02	0.18	[0.02, 0.10]	2.98	.003		
Year 2	0.04	11.38	(1, 258)	< .001	Intercept	0.23	0.12		[0.10, 0.46]	1.98	.05	27.66	0.04
					Participation in PE	0.08	0.02	0.21	[0.03, 0.13]	3.37	< .001		
Year 3	0.00	0.40	(1, 176)	.53	Intercept	0.65	0.04		[0.57, 0.73]	16.08	< .001	0.20	0.00
					Participation in PE	0.01	0.01	0.05	[-0.01, 0.03]	0.64	.53		
Year 4	0.00	0.61	(1, 193)	.44	Intercept	0.66	0.04		[0.57, 0.75]	14.63	< .001	0.21	0.00
					Participation in PE	0.01	0.01	0.06	[-0.01, 0.03]	0.78	.44		

PF Preparatory exercises

For the second year, we revised the preparatory exercises to increase the fidelity to PF design principles. For these revisions, we collected informal feedback from students, tutors, and the lecturer on the wording, the difficulty level and the length of the exercises. Furthermore, for each item in the end-of-semester tests on conceptual understanding where students' performance in the first year was low, we revised the corresponding preparatory exercise to include more guidance towards the targeted conception.

In the second year, we found that 20% more students passed the exam and the exam performance in the second year increased by 0.67 grade points (13.2%) with a significant effect of $d=0.59$ compared to the baseline, which can be considered a large effect for an educational intervention (Kraft, 2020).

Even though students faced lockdown measures due to the Covid-19 pandemic in the Spring of the second year, their exam performance did not drop. This is in line with research from Spain, showing that, in the 2020 final exams, students' performance was not negatively impacted by lockdown measures (Iglesias-Pradas et al., 2021).

Interestingly, although the preparatory exercises only underwent minor changes for the third year of the intervention, exam performance dropped compared to the second year. In the third year, we only had an increase of 0.31 grade points (6.2%) compared to the baseline with an effect of $d=0.28$ and an increase of passing grades of 10%. The drop in performance might be due to the Covid-19 pandemic measures that influenced student life significantly during this academic year. In contrast to Year 2, where lockdown measures impacted part of the spring semester and exam preparation, students in Year 3 started their university studies in the midst of the pandemic in Fall 2020. Students were split up into bubbles of 25 and lectures were only available online to reduce in-person interactions. These restrictions potentially reduced students' opportunities to build friendships and form study groups. Along with fear and uncertainty caused by the pandemic, the restrictions likely hurt our students' mental health, as shown for university students in Italy (Cesco et al., 2021) and the UK (Ihm et al., 2021) and thus presumably also negatively impacted learning (Online Resource 5). This hypothesis is supported by the fact that students' performance in analysis, their other first-year mathematics course, also dropped in this cohort.

In the fourth year of the intervention, starting in the Fall semester of 2021, measures against the Covid-19 pandemic were largely lifted and student life entered the *new normal* era. Although the preparatory exercises remained unchanged compared to the third year, exam performance increased again, with 15% more students passing the exam compared to the baseline. We observed an increase of 0.60 grade points (12.2%) compared to the baseline with an effect of $d=0.54$. One interpretation of these results could be that the intervention returned to the effectiveness seen in the second year, providing more evidence for our hypothesis that the results in the third year were heavily influenced by the pandemic. This interpretation is supported by the trends in students' first-year courses, mainly in analysis. There, performance also dropped during the third year of our intervention and regained a similar value as in the second year.

Considering the relationship between the number of exercises the students participated in and their exam performance, the difference between the years is not easy to explain. It is worth noting that there is a selection bias in this part of our data as we could only include the data of students in our regression model who volunteered for the additional analysis, and we observed that more high-performing students decided to enter the study. We were thus unable to analyse the whole cohort which might have reduced the amount of variance in the participation variables and could be a reason for the difference in prediction power of participation on exam performance in the four years of our intervention. Furthermore, in Years 2 to 4, some repeating students joined the study allowing us to retroactively use their data as first-timers in Years 1 to 3. This phenomenon broadened our data sets of low-performing students in Years 1 to 3. This could explain why the fourth iteration was overall more effective than the third, even though our detailed analysis does not reflect that. Modelling the relationship between participation and performance more precisely should be an object of future research.

An explanation for the difference in effect size between Year 4 and the preceding years could lie in their participation behaviour. Our participation variable only reflects whether students handed in their solution to the preparatory exercises to their tutor before the corresponding lecture. We had no means of assessing whether students engaged with the exercises after the allotted time span or engaged with the exercises without submitting a solution to their tutor. It might be that in Year 4, the number of students who engaged with the exercises without submitting a solution was greater than in the previous cohorts. Future research should include an instrument to investigate students' engagement with the preparatory exercises in addition to recording the submissions.

Another explanation of the decreasing relationship between participation and performance over the iterations could be the effect of sharing resources between cohorts. Repeating students might have distributed solutions to the preparatory exercises amongst their peers, reducing students' motivation to engage with the materials themselves during the allotted time span. As a result, more students might have handed in copied solutions to the preparatory exercises in Years 2 to 4 than in the first year, confounding the analysis. In future implementations, the content of the preparatory exercises might have to be alternated between subsequent cohorts to reduce this effect. Again, this could mean that some students did not engage with the preparatory exercises within the time span set by the intervention. Moreover, it might indicate that other aspect of the intervention, such as the discussion of the preparatory exercises in the beginning of the lectures supported students' learning, as exam performance remained high in Year 4, even though these students potentially had the highest number of solutions available from other students.

In the first two years of the intervention, participation significantly predicted students' performance in the assessments on conceptual understanding with R^2 between 0.03 and 0.05. Despite being significant, the strength of this relationship was smaller than we expected. On the one hand, the selection bias mentioned above presumably played a role. On the other hand, the fact that the assessments were facultative, and students' performance had no influence on their exam grades complicates the interpretation of these results. Instead of doing their best during the assessments, students might have guessed on some answers and engaged less with more difficult items.

While for the first assessment in the third year, the prediction remained significant with $R^2 = 0.01$, we obtained non-significant results for the second assessment in the third year and both assessments in the fourth year. Furthermore, the Bayesian factors of these comparisons favoured the null model. As before, selection bias and the effort students put into solving the items correctly might have been important factors. Moreover, the students were aware that they would be able to use the assessment materials as a resource for studying before the final exam independent of their participation or performance in the assessments. We thus feel that students' performance in the final exam better reflected their learning outcome than their performance in the facultative assessments, even though we were unable to differentiate between procedural knowledge and conceptual understanding in the final exam.

Following the operationalisation in the meta-analysis of Sinha and Kapur (2021) of the PF design principles, our study scored around 57% PF design fidelity. Our observed effect size of $d = 0.58$ and $d = 0.52$ on exam performance in the second and fourth year of the intervention surpasses the expected effect size according to the meta-analysis. This might be due to the age of the participants and the length of the intervention, as both of these factors were observed to be significant by Sinha and Kapur.

The fact that linear algebra is particularly well-suited to support students' learning in a PS-I learning design might have added to the effect of our intervention on students' performance. Dealing with problems targeting base concepts before formal instruction might have increased students' intellectual need to understand critical features of linear algebra concepts, which could have been easily overlooked when linear algebra was presented as a unified and polished theory (Harel, 2000).

Limitations and further research

Our work is based in a real university classroom serving hundreds of students and thus is limited by the vast number of variables we did not control for or include in our models. We did not collect any data on students' demographic background, their affective states, or their mathematical performance before university. We concede that a more detailed investigation of the incoming characteristics of students might allow for more precise statements as to which groups of students were affected by our intervention. Such an analysis might provide insights as to whether preparatory exercises in higher education provide advantages only for certain students and leave others behind. However, we were able to show in the present study that by introducing preparatory exercises to a higher mathematics course, an overall increase in learning outcomes can be achieved.

Furthermore, our design did not include a control group. The correlational analyses used to investigate the relationship between participation and performance are limited in that we cannot exclude the possibility of an unknown third factor being responsible for the detected effects. In particular, the effect of participation on exam performance could partly be attributed to a selection effect. It is likely that more motivated students participated more often in the preparatory exercises and also performed better in the final exam. In future research, affective variables that might impact the relationship between participation and performance should be investigated.

We included all students who participated in the final exam in our overall comparison of the final exam performance of the intervention cohorts to the baseline cohorts, regardless of whether they participated in the preparatory exercises. However, most students in the intervention cohorts did participate in the preparatory exercises. We found that the percentage of students who took the final exam but did not submit a solution to any of the preparatory exercises was consistently lower than 6% in each of the intervention. Given that the number of these students was consistently low, their impact on the results is minor.

Even though our design is based on research on problem-solving before instruction and productive failure, our results show a positive effect of the addition of the preparatory exercises to the course, but we cannot infer whether a similar effect could have been observed if the students solved the exercises *after* having received the corresponding instruction. Thus, we cannot make a statement about the importance of the sequence in which the learning materials were administered. However, we developed the additional exercises specifically to serve as preparation for instruction, and their structure and guided nature might inhibit any meaningful use after the targeted instruction in a real-world university course.

A more in-depth analysis of students' solutions to the preparatory exercises might have provided valuable insights into possible causes of our observed effects and the interplay of the learning processes in the two phases of the PS-I design. It would have allowed for an understanding of students' intermediate knowledge which in turn might have permitted clearer conclusions. Due to the high number of hand-ins (around 2000 per cohort), such an additional analysis might have exceeded the scope of the present article. We encourage a detailed discussion of students' solutions in further research as well as an investigation of the learning mechanisms involved.

Moreover, our intervention was developed collaboratively with the course lecturer, and he was not blind to our hypotheses, which might have influenced his teaching. As the goals of our study did not encompass the replication of the positive effect of PS-I on conceptual understanding and transfer (Loibl et al., 2017; Sinha & Kapur, 2021), but to test whether PS-I can be implemented in a university mathematics course in a DBR approach, we aimed at an ecologically valid implementation that includes collaboration with the lecturer.

Conclusion

In sum, we implemented a problem-solving before instruction (PS-I) design for a first-year linear algebra course for mechanical engineering students at a technical university in a design-based research approach. Over the four years of this implementation, we drew on the productive failure (PF) principles as stated by Kapur and Bielaczyc (2012). Some principles, such as building the instruction on students' solutions, needed to be adapted to the higher education context. Our implementation involved only minor changes to the original course structure and required little extra time for students. We observed a substantial increase in exam performance compared to prior years. We conclude that our intervention seemingly supported student learning. Due to the lack of a control group, we cannot draw any conclusions concerning the effects of specific design components.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11251-025-09709-8.

Author contributions Vera Baumgartner: Conceptualisation; data curation; formal analysis; investigation; methodology; project administration; resources; supervision; validation; visualisation; writing—original draft, review & editing. Simona Daguati: Conceptualisation; data curation; investigation; project administration; resources; Supervision; validation; writing—review. Dragan Trninic: conceptualisation; funding acquisition; methodology; resources; writing—review & editing. Meike Akveld: Conceptualisation; resources; writing—review. Alexander Caspar: Conceptualisation; resources; writing—review. Norbert Hungerbühler: Conceptualisation; funding acquisition; resources; supervision; writing—review. Manu Kapur: Conceptualisation; funding acquisition; methodology; resources; supervision; writing—review & editing.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich. The study was conducted within a project of the Future Learning Initiative of the Open ETH program and funded by ETH Zurich.

Data availability The datasets generated and analysed during the current study are available in the ETH Zurich research collection, https://doi.org/https://doi.org/10.3929/ethz-b-000625066.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose. The authors have no competing interests to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

Ethical approval The study was approved by the ethics commission of ETH Zurich (2019-N-50).

References

Belenky, D. M., & Nokes-Malach, T. J. (2012). Motivation and transfer: The role of mastery-approach goals in preparation for future learning. *Journal of the Learning Sciences, 21*(3), 399–432. https://doi.org/10.1080/10508406.2011.651232

Carlson, D. (1993). Teaching linear algebra: Must the fog always roll in? *The College Mathematics Journal, 24*(1), 29–40. https://doi.org/10.1080/07468342.1993.11973503

Cesco, S., Zara, V., Toni, A. F. D., Lugli, P., Betta, G., Evans, A. C. O., & Orzes, G. (2021). Higher education in the first year of covid-19: Thoughts and perspectives for the future. *International Journal of Higher Education, 10*(3), 285. https://doi.org/10.5430/ijhe.v10n3p285

Chen, D.-W., & Catrambone, R. (2019). Productive failure and subgoal scaffolding in novel domains. In R. A. Sottilare & J. Schwarz (Eds.), *Adaptive Instructional Systems* (pp. 282–300). Springer International Publishing.

Chen, O., & Kalyuga, S. (2020). Exploring factors influencing the effectiveness of explicit instruction first and problem-solving first approaches. *European Journal of Psychology of Education, 35*(3), 607–624. https://doi.org/10.1007/s10212-019-00445-5

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13.

Darabi, A., Arrington, T. L., & Sayilir, E. (2018). Learning from failure: A meta-analysis of the empirical studies. *Educational Technology Research and Development, 66*(5), 1101–1118. https://doi.org/10.1007/s11423-018-9579-9

DeCaro, M. S., & Rittle-Johnson, B. (2012). Exploring mathematics problems prepares children to learn from instruction. *Journal of Experimental Child Psychology, 113*(4), 552–568. https://doi.org/10.1016/j.jecp.2012.06.009

Dorier, J.-L. (2000). Epistemological analysis of the genesis of the theory of vector spaces. In J.-L. Dorier (Ed.), *On the Teaching of Linear Algebra* (pp. 3–81). Kluwer Academic Publishers.

Dorier, J.-L., & Sierpinska, A. (2002). Research into the teaching and learning of linear algebra. In D. Holton, M. Artigue, U. Kirchgräber, J. Hillel, M. Niss, & A. Schoenfeld (Eds.), *The Teaching and Learning of Mathematics at University Level* (pp. 255–273). Kluwer Academic Publishers.

ETH Zurich. (2025*). Degrees, credits & grading system, legal basis. Degrees, Credits & Grading S*ystem, Legal Basis | ETH Zurich. Retrieved February 17, 2025, from https://ethz.ch/en/studies/legal-principles-degrees.html

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2014). An alternative time for telling: When conceptual instruction prior to problem solving improves mathematical knowledge. *British Journal of Educational Psychology, 84*(3), 502–519. https://doi.org/10.1111/bjep.12035

Glogger-Frey, I., Gaus, K., & Renkl, A. (2017). Learning from direct instruction: Best prepared by several self-regulated or guided invention activities? *Learning and Instruction, 51*, 26–35. https://doi.org/10.1016/j.learninstruc.2016.11.002

Hannah, J., Stewart, S., & Thomas, M. (2013). Emphasizing language and visualization in teaching linear algebra. *International Journal of Mathematical Education in Science and Technology, 44*(4), 475–489. https://doi.org/10.1080/0020739X.2012.756545

Harel, G. (2000). Three principles of learning and teaching mathematics. In J.-L. Dorier (Ed.), *On the Teaching of Linear Algebra* (pp. 177–189). Kluwer Academic Publishers.

Harel, G. (2018). The learning and teaching of linear algebra through the lenses of intellectual need and epistemological justification and their constituents. *Challenges and strategies in teaching linear algebra* (pp. 3–27). Springer.
Hartmann, C., Gog, T., & Rummel, N. (2020). Do examples of failure effectively prepare students for learning from subsequent instruction? *Applied Cognitive Psychology, 34*(4), 879–889. https://doi.org/10.1002/acp.3651
Iglesias-Pradas, S., Hernández-García, Á., Chaparro-Peláez, J., & Prieto, J. L. (2021). Emergency remote teaching and students' academic performance in higher education during the COVID-19 pandemic: A case study. *Computers in Human Behavior, 119*, 106713. https://doi.org/10.1016/j.chb.2021.106713
Ihm, L., Zhang, H., van Vijfeijken, A., & Waugh, M. G. (2021). Impacts of the Covid-19 pandemic on the health of university students. *The International Journal of Health Planning and Management, 36*(3), 618–627. https://doi.org/10.1002/hpm.3145
Jarosz, A. F., Goldenberg, O., & Wiley, J. (2017). Learning by invention: Small group discussion activities that support learning in statistics. *Discourse Processes, 54*(4), 285–302. https://doi.org/10.1080/0163853X.2015.1129593
JASP Team. (2023). *JASP (Version 0.17)[Computer software]* [Computer software]. https://jasp-stats.org/
Kapur, M. (2008). Productive failure. *Cognition and instruction, 26*(3), 379–424.
Kapur, M. (2010). Productive failure in mathematical problem solving. *Instructional Science, 38*(6), 523–550. https://doi.org/10.1007/s11251-009-9093-x
Kapur, M. (2011). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science, 39*(4), 561–579. https://doi.org/10.1007/s11251-010-9144-3
Kapur, M. (2014). Productive failure in learning math. *Cognitive Science, 38*(5), 1008–1022. https://doi.org/10.1111/cogs.12107
Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist, 51*(2), 289–299. https://doi.org/10.1080/00461520.2016.1155457
Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences, 21*(1), 45–83. https://doi.org/10.1080/10508406.2011.591717
Kerrigan, J. (2018). *Productive failure in the flipped mathematics classroom* [PhD Thesis]. Rutgers University-Graduate School of Education.
Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253. https://doi.org/10.3102/0013189X20912798
Lai, P. K., Portolese, A., & Jacobson, M. J. (2017). Does sequence matter? Productive failure and designing online authentic learning for process engineering: Productive failure and designing online authentic learning. *British Journal of Educational Technology, 48*(6), 1217–1227. https://doi.org/10.1111/bjet.12492
Loehr, A. M., Fyfe, E. R., & Rittle-Johnson, B. (2014). Wait for it… Delaying instruction improves mathematics problem solving: A classroom study. *The Journal of Problem Solving*. https://doi.org/10.7771/1932-6246.1166
Loibl, K., Leuders, T., Glogger-Frey, I., & Rummel, N. (2024). CID: a framework for the cognitive analysis of composite instructional designs. *Instructional Science*. https://doi.org/10.1007/s11251-024-09665-9
Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review, 29*(4), 693–715. https://doi.org/10.1007/s10648-016-9379-x
Loibl, K., & Rummel, N. (2014a). Knowing what you don't know makes failure productive. *Learning and Instruction, 34*, 74–85. https://doi.org/10.22318/cscl2013.1.296
Loibl, K., & Rummel, N. (2014b). The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. *Instructional Science, 42*(3), 305–326. https://doi.org/10.1007/s11251-013-9282-5
Loibl, K., & Rummel, N. (2015). Productive Failure as strategy against the double curse of incompetence. *Learning: Research and Practice, 1*(2), 113–121. https://doi.org/10.1080/23735082.2015.1071231
Nathan, M. J. (2012). Rethinking formalisms in formal education. *Educational Psychologist, 47*(2), 125–148. https://doi.org/10.1080/00461520.2012.667063
Newman, P. M., & DeCaro, M. S. (2019). Learning by exploring: How much guidance is optimal? *Learning and Instruction, 62*, 49–63. https://doi.org/10.1016/j.learninstruc.2019.05.005
Posit team. (2023). *RStudio: Integrated development environment for R* [Manual]. Posit Software, PBC. http://www.posit.co/

Roll, I., Aleven, V., & Koedinger, K. R. (2009). Helping students know'further'-increasing the flexibility of students' knowledge using symbolic invention tasks. Proceedings of the 31st Annual Conference of the Cognitive Science Society, 1169–1174.

Schalk, L., Schumacher, R., Barth, A., & Stern, E. (2018). When problem-solving followed by instruction is superior to the traditional tell-and-practice sequence. *Journal of Educational Psychology, 110*(4), 596–610. https://doi.org/10.1037/edu0000234

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. *Handbook of Research on Mathematics Teaching and Learning*, *334370*. https://doi.org/10.1177/002205741619600202

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*(4), 475–5223. https://doi.org/10.1207/s1532690xci1604_4

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22*(2), 129–184. https://doi.org/10.1207/s1532690xci2202_1

Sharma, S. C. (1987). The effects of autocorrelation among errors on the consistency property of OLS estimator. *Journal of Statistical Computation and Simulation, 28*(1), 43–52. https://doi.org/10.1080/00949658708811007

Sinha, T., & Kapur, M. (2019). When productive failure fails. *Proceedings of the Annual Meeting of the Cognitive Science Society, 41*, 2811–2817.

Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research, 91*(5), 761–798. https://doi.org/10.3102/00346543211019105

Stewart, S., Andrews-Larson, C., & Zandieh, M. (2019). Linear algebra teaching and learning: Themes from recent research and evolving research priorities. *ZDM, 51*(7), 1017–1030. https://doi.org/10.1007/s11858-019-01104-1

Tall, D. (2008). The transition to formal thinking in mathematics. *Mathematics Education Research Journal, 20*(2), 5–24. https://doi.org/10.1007/BF03217474

Tucker, A. (1993). The growing importance of linear algebra in undergraduate mathematics. *The College Mathematics Journal, 24*(1), 3–9. https://doi.org/10.1080/07468342.1993.11973500

Wawro, M., Rasmussen, C., Zandieh, M., Sweeney, G. F., & Larson, C. (2012). An inquiry-oriented approach to span and linear independence: The case of the magic carpet ride sequence. *Primus, 22*(8), 577–599. https://doi.org/10.1080/10511970.2012.667516

Weaver, J. P., Chastain, R. J., DeCaro, D. A., & DeCaro, M. S. (2018). Reverse the routine: Problem solving before instruction improves conceptual knowledge in undergraduate physics. *Contemporary Educational Psychology, 52*, 36–47. https://doi.org/10.1016/j.cedpsych.2017.12.003

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org