Technology, Knowledge and Learning
https://doi.org/10.1007/s10758-023-09714-1

ORIGINAL RESEARCH

Voluntary E-Learning Exercises Support Students in Mastering Statistics

Jakob Schwerter[1] · **Taiga Brahm**[2]

Accepted: 11 December 2023

Abstract
University students often learn statistics in large classes, and in such learning environments, students face an exceptionally high risk of failure. One reason for this is students' frequent statistics anxiety. This study shows how students can be supported using e-learning exercises with automated knowledge of correct response feedback, supplementing a face-to-face lecture. To this end, we surveyed 67 undergraduate social science students at a German university and observed their weekly e-learning exercises. We aggregated students' exercise behavior throughout the semester to explain their exam performance. To control for participation bias, we included essential predictors of educational success, such as prior achievement, motivation, personality traits, time preferences, and goals. We applied a double selection procedure based on the machine learning method Elastic Net to include an optimal but sparse set of control variables. The e-learning exercises indirectly promoted the self-regulated learning techniques of retrieval practice and spacing and provided corrective feedback. Working on the e-learning exercises increased students' performance on the final exam, even after controlling for the rich set of control variables. Two-thirds of students used our designed e-learning exercises; however, only a fraction of students spaced out the exercises, although students who completed the exercises during the semester and were not cramming at the end benefited additionally. Finally, we discuss how the results of our study inform the literature on retrieval practice, spacing, feedback, and e-learning in higher education.

Keywords Higher education · Undergraduate students · Retrieval practice · Self-testing · Spaced-out learning · Machine learning application

✉ Jakob Schwerter
jakob.schwerter@tu-dortmund.de

Taiga Brahm
taiga.brahm@uni-tuebingen.de

[1] Center for Research on Education and School Development, and Faculty of Statistics, TU Dortmund University, Martin-Schmeißer-Weg 13, 44227 Dortmund, Germany

[2] Economic Education, University of Tübingen, Melanchthonstr. 30, 72074 Tübingen, Germany

Published online: 21 March 2024

1 Introduction

Statistics is a course in higher education (HE) that students often have trouble learning (Förster et al., 2018; Schwerter, Wortha et al., 2022; Vaessen et al., 2017) and are consequently affected by statistics anxiety (Condron et al., 2018). This is of serious practical concern as statistics is part of the curriculum of many university subjects (Garfield & Ben-Zvi, 2007). Research also indicates that many beginning students have severe difficulties in thinking statistically and face several misconceptions about statistics (Förster et al., 2018). Therefore, it is important to improve statistics learning concept to support students to counteract their learning difficulties. One possibility to improve student learning is the usage of e-learning tools with retrieval practice, video teaching, and similar formats, which have gained relevance in HE (Förster et al., 2018, 2022; Graham et al., 2013; Schwerter, Wortha et al., 2022; Velde et al., 2021). Research and academic literature evaluating this new way of teaching have been growing accordingly (Anthony et al., 2020; Castro & Tumibay, 2021). Recently, learning analytics has become a major trend in HE research (Hellings & Haelermans, 2020). From the many possibilities, this study focuses on students' retrieval practice as it is one of the most robust and efficient methods in learning science (Yang et al., 2021).

As the literature reports, practicing helps people acquire and apply skills more confidently (Jonides, 2004). To make the most out of students' practice time investment, we focus on the most effective learning techniques: Retrieval practice with corrective feedback and variability. The retrieval practice effect has been proven to be one of the most robust results in memory research in cognitive psychology (Karpicke, 2017), both in laboratory (e.g., Karpicke & Blunt, 2011; Lim et al., 2015) and in a few real educational settings (Förster et al., 2018; Roediger et al., 2011; Schwerter, Dimpfl et al., 2022). With the help of increased retrieval practice during the semester, students can easily reflect on whether they are achieving their study goals and monitoring their learning progress. By reviewing their performance on these self-tests, students can reflect on their achievements and identify areas for improvement. Thereby, this approach can support students to develop their self-regulation skills and it empowers students to take charge of their learning by making informed decisions about their study habits (Alexander et al., 2011; Azevedo, 2009; Butler & Winne, 1995). Thus, retrieval practice with direct feedback can serve as a powerful tool for promoting self-regulated learning (Ifenthaler et al., 2023).

However, evidence on the interplay of retrieval practice, spacing behavior, and task variability in real educational settings with problem-solving exercises is missing. Accordingly, in this study we analyzed additional retrieval practice through weekly voluntary online exercises. We examined whether participating in weekly voluntary e-learning exercises with different versions and free choice of when to work on these exercises helps students achieve higher grades at the end of the semester. We observed $N = 67$ students participating in an e-learning environment accompanying an advanced statistics course (on inference statistics) over a whole semester. This third-semester course is designed for undergraduate social science students at a large public university in Germany. To address the challenge of self-selection, we used important predictors of student achievement (such as prior achievement, motivation, personality, and time preferences) as control variables, applying a double-feature selection method (Belloni et al., 2014) to avoid overfitting. This approach corresponds to the call for future research to include affective prerequisites (Förster et al., 2018). Our study thus aims at contributing to the literature on retrieval practice by taking a

closer look at students' usage of voluntary online exercises in a real-life setting, at the same time controlling for important prerequisites.

1.1 Literature on Retrieval Practice and Related Concepts

Retrieval practice (or practice testing, self-testing), i.e., retrieving knowledge under study without any stakes, is one of the most efficient learning techniques for later retention (Donoghue & Hattie, 2021; Dunlosky et al., 2013; Yang et al., 2021). It is a study technique requiring the student to set aside the learning material and try to recall information from memory. This applies desirable difficulties (Bjork, 1994), i.e., it imposes challenging conditions on students, consequently requiring higher cognitive engagement. Although this initially seems to slow down the learning process, it improves later retention and transfer (Roediger III & Karpicke, 2006; Yan et al., 2014). Both are of particular importance in real education settings like university courses as topics within one course and courses within a study program build upon each other. Accordingly, knowledge learned at the beginning (such as statistics) is needed to understand the material at the end of studying. Retrieval practice improves delayed retention compared to re-reading (Roediger III & Karpicke, 2006), note-taking (McDaniel et al., 2009), verbal and visual elaboration of material (Karpicke & Smith, 2012), as well as using concept maps (Karpicke & Blunt, 2011; Lechuga et al., 2015).

Moreover, several studies have highlighted how this retrieval effect can be enhanced. For example, retrieval practice can be more effective by giving learners tasks of higher difficulty requiring comprehension and application rather than just memorizing discrete facts (Jensen et al., 2014). Regarding the difficulty level, it is unclear whether students must perform well during retrieval practice. Higher success in practice phases improved the retrieval effect (Racsmány et al., 2020). However, others have shown that performance in retrieval practice is not essential (Butler et al., 2017; Schwerter, Dimpfl et al., 2022). Additionally, the feedback literature shows that making errors does not harm but helps learners (Butler et al., 2011; Hays et al., 2013; Kornell et al., 2009). For example, Butler and Roediger (2008) found that feedback enables learners to correct incorrectly stored information. Due to the feedback, answers that could not be retrieved were not discarded from the memory (Kornell et al., 2011; Mundt et al., 2020; Wong & Lim, 2022). Feedback can even correct mistakes made with high confidence, also called hypercorrection (Butler et al., 2011). Thus, the students' practice performance might not be crucial if the retrieval practice is accompanied by corrective feedback. Only if the retrieval practice exercises are too difficult, the retrieval practice may be harmful to students learning (Carpenter et al., 2016; Karpicke et al., 2014).

Another option to enhance the retrieval practice effects is spaced learning, i.e., repeated retrieval distributed over time (Rawson et al., 2015). Spacing out the learning over a more extended period is more beneficial for students than cramping before deadlines (Cepeda et al., 2006; Dempster, 1989). Additionally, spaced-out learning over a more extended period is better than cramming before a test because memory traces are reinforced through repetition, also known as the forgetting curve effect (Murre & Dros, 2015). The positive impact of retrieval practice and spacing on learning has been shown in many studies (Baker et al., 2020; Rodriguez et al., 2021a, 2021b)—even independent of prior performance (Rodriguez et al., 2021a, 2021b). The combination of both approaches is particularly helpful for students (Rodriguez et al., 2021a, 2021b; Roediger III & Karpicke, 2006).

Additionally, it seems advisable in retrieval practice to not use the same question repeatedly but to use different questions targeted at the same learning goal (Butler et al., 2017). In a study in geological sciences, Butler et al. (2017) demonstrated that increasing the variability improves student learning as students can faster transfer their knowledge to new examples of the same concept. One reason for this might be that variability helps students to distinguish the critical features from interchangeable information to better identify the concept being learned (Butler et al., 2017).

Although most literature on retrieval practice used rather simple test materials for measurement like single words, word pairs, text passages, and academic facts (Carpenter, 2012; Su et al., 2020), more challenging outcomes of understanding and comprehension of complex, educationally-relevant learning contents are now also investigated (Butler, 2010; Carpenter, 2014; Karpicke & Aue, 2015). Similarly, the literature expanded from showing improved recognition, cued recall, and free recall (Su et al., 2020) as well as transfer of factual and conceptual knowledge (Butler, 2010; Chan et al., 2006), to the promotion of superior critical evaluation of research articles (Dobson et al., 2018), analogical-problem-solving performance using hypothesis-testing examples (Wong et al., 2019), and to promoting deep conceptual learning in scientific experimentation skills (Tempel et al., 2020). However, in statistics, a topic in which solving exercises are a natural and widely used practice, the retrieval practice effect is seldomly analyzed. One notable exception is a field study using quizzing as retrieval practice in HE (Förster et al., 2018). The quizzes, used during the semester in a statistics class, included multiple-choice questions. If the students participated in the quizzes, their exam performance at the end of the semester improved. Similarly, but for mathematics in HE and using (mostly) open-end questions, Schwerter, Dimpfl et al. (2022) showed that more retrieval practice in mathematics led to more exam points at the end of the semester, depending on students' motivation, personality, time preferences, and prior achievement. In these two studies, it is unclear whether the retrieval practice using multiple-choice or open-end questions improved students' knowledge or whether the (combination of) testing (and feedback on the testing) encouraged spaced learning during the semester. Therefore, more research is needed to clarify whether a retrieval effect is observed in the studies or whether retrieval practice led to more spaced-out learning.

1.2 Prediction of Student Achievement in Higher Education

Exam grades prediction is a prevalent topic in empirical research. This study also contributes to this literature as it includes a variety of predictor variables. Based on conceptual considerations, relevant theoretical, and empirical work related to students performance in higher education, we focused on student information (Benden & Lauermann, 2022), self-set course goals (van Lent & Souverijn, 2020), expectancy-value beliefs (Eccles et al., 1983), achievement goals (Elliot & McGregor, 2001), the Big Five personality traits (Digman, 1990), and time preferences (Frederick & Loewenstein, 2002). For example, student information like prior achievement, employment responsibility and students' gender are essential predictors for exam grades (McKenzie & Schweitzer, 2001; Paechter et al., 2010; Schwerter, Wortha et al., 2022).

Regarding students' motivation, operationalized by students' achievement goals, there are mixed results on the effect of students' level of mastery and performance approach on exam performance (Elliot et al., 1999; Harackiewicz et al., 2002; Plante et al., 2013; Yperen et al., 2014). Exam performance seems to have a negative association solely with mastery and performance avoidance (Baranik et al., 2010; Hulleman et al., 2010; Payne

et al., 2007). Moreover, the relationship between motivation and performance can be demonstrated employing students' expectancy, value, and cost beliefs (e.g., Bailey & Phillips, 2016; Krause et al., 2012; Macher et al., 2015; Marsh & Martin, 2011; Wigfield & Eccles, 2000). Even though achievement goals and expectancy-value theory are related measures of student motivation, Plante et al. (2013) show that explanatory power is increased when variables from both concepts are included. Particularly for the case of e-learning, Dunn and Kennedy (2019) have shown that intrinsically motivated learners are diligent in completing e-learning exercises, while extrinsically motivated learners complete them more frequently.

In addition to motivation, the literature has documented the high importance of the Big Five personality traits on academic success (Komarraju et al., 2009; Rimfeld et al., 2016; Sorić et al., 2017). Last, concerning students' time preferences, i.e. their inclination to prioritize immediate or future benefits, Bisin and Hyndman (2020) have shown that risk-averse students outperform risk-taking students in exams. Further, similar to Plante et al. (2013) in the context of motivation, Becker et al. (2012) underscored that time preferences complement personality traits, and that both contribute to a better explanation of educational achievement. Since these variables serve as control variables in our study, we refer the reader to the cited literature for further details on each concept.

1.3 Present Study & Research Questions

To address the research gap mentioned, we give students weekly retrieval practice exercises in a statistics class and measure their effect on students' exam performance. Contrary to the two studies most similar, Förster et al. (2018) and Schwerter, Dimpfl et al. (2022), we let students decide when to use this additional online learning opportunity. In comparison, Förster et al. (2018) allocated students a whole week to solve 4 or 5 (depending on the semester of the data collection) weekly quizzes, while in Schwerter, Dimpfl et al. (2022), there was a constrained 60-min window on a specific day allocated to students to solve three practice tests. The key distinction in the present study lies in students' autonomy when to work on the exercises, allowing us to observe varying spacing behavior and to examine whether the retrieval effect persists irrespective of spacing during the semester. This is a novel approach not previously explored. Furthermore, the students were offered multiple versions of the same exercises, enabling students to practice the same topic, using different exercise versions. This should enhance the retrieval effect due to exercise variability (Butler et al., 2017). In contrast to Schwerter, Dimpfl et al. (2022) but in line with Förster et al. (2018), we refrained from providing any incentive for engaging in retrieval practice exercises, primarily because retrieval practice is considered a low-stakes practice opportunity. Offering incentives like extra credit points for the exam could have increased students' pressure or even been an inducement to cheat. Additionally, these external incentives might undermine intrinsic motivation (Deci et al., 2001). Hence, with our study design, we further contribute to the literature on retrieval practice opportunities as part of a university course. Lastly, as we observe students in a statistics course in HE, we also contribute to the general retrieval practice literature on applying knowledge to solve novel (target) problems using complex educational materials. The educational material is complex because it is composed of high interactivity of different and interconnected information elements (Karpicke & Aue, 2015; Wong et al., 2019). Analogous problem-solving requires procedural knowledge and successive execution of rules to apply an algorithm to solve a new task (Wong et al., 2019).

Our study corresponds to the call for future research (Carvalho et al., 2022; Förster et al., 2018; Reeves & Lin, 2020; Schwerter, Dimpfl et al., 2022; Wong et al., 2019; Yang et al., 2021) in four ways. (i) We assess problem-solving with exercises in which students do not need to recall the solution but learn the steps to arrive at the solutions and calculate the answer rather than stating whether a hypothesis testing decision is true or false, i.e., knowing how to solve a problem rather than knowing the solution. (ii) We check the difference between spaced-out learning in a HE course in comparison to cramping before the exam with regard to students' exam performance. (iii) We include affective preconditions. (iv) Lastly, we conduct a field analysis in a HE gateway statistics course to increase the ecological validity of laboratory research. We were particularly interested in a statistics class because abundant literature has shown that statistics is a course which many students find troubling to master in HE (Vaessen et al., 2017) and are consequently affected by statistics anxiety (Condron et al., 2018). The specific research questions are as follows.

- RQ1: Do students use the e-learning exercises even though they are voluntary, and no external rewards are given (RQ1a)? When students practice, do they space or cramp the exercises (RQ1b)? Do students only self-test one weekly exercise once or do they have multiple tries per week to make use of the exercise variability? (RQ1c)
- RQ2: Do the weekly retrieval practice (RQ2a), spacing (RQ2b), and multiple tries per week (RQ2c) result in more exam points?
- RQ3: Are the effects of retrieval practice, spacing, and multiple tries per week on exam points robust when controlling for demographic information, prior achievement, expectancy-value variables, achievement goals, personality traits, and time preferences? Or does the effect vanish once the additional controls are included, and hence the effect in RQ2 is only driven by selection?

While different studies highlight that students seldomly use retrieval practice to study (Susser & McCabe, 2013), Förster et al. (2018) and Schwerter, Dimpfl et al. (2022a) showed that students in statistics and mathematics courses in HE do use voluntarily practice opportunities. Thus, we expect that at least some students will use our retrieval practice opportunities. Furthermore, given that students are likely to procrastinate (Baker et al., 2019), we expect that most students have cramped rather than spaced-out their learning. In line with previous research (e.g. Tullis & Maddox, 2020), we also expect that most students will only do one try per week and not multiple tries per week. Next, following Förster et al. (2018), we expect an unconditional practice effect. Finally, following Schwerter, Dimpfl et al. (2022), we expect to find a lowered but still significant conditional retrieval practice effect. However, as this was not studied before, the effects of the free choice to space or cramp on students' practice are unclear which highlights the need for further evidence from authentic HE settings.

2 Methods

2.1 Course Information

The topics of the course, *Social Science Statistics 2*, are inference statistics. This course builds on the course *Social Science Statistics 1* from the preceding semester and spans 15

weeks, with 13 lectures. The lectures are accompanied by a weekly tutorial session with mandatory attendance in which tutors present solutions to the problem sheets. If the students miss more than two sessions, they cannot take the exam at the end of the semester. Thereby, the requirement is not to miss a tutorial, not whether they are prepared or actively participate. A general overview of the course topics and respective dates during the semester can be seen in Appendix Table 8.

Then, at the center of the research design, students can practice the week's topic with the help of e-learning exercises. These exercises cover one to three weekly tutorial exercises with the same frame or wording as those in the tutorial but with new examples, following the concept of variability (Butler et al., 2017). The number of exercises depends on the respective length and difficulty.

The official exam took place at the end of the semester. The exam was divided into a first and second trial, with the first being the main trial. The first trial took place one week after the end of the lectures, and the second trial would have occurred one week before the new semester started. However, due to the COVID-19 pandemic, the second trial was postponed several weeks into the next semester. Because of this unique situation, we exclude it in the analysis.

2.2 Design of E-Learning Exercises

This study aimed to investigate whether participating in additional e-learning exercises enhanced students' achievement in inference statistics. The exercises were provided weekly and voluntarily with direct automatic corrective feedback in the online management system of the university. Additionally, students saw how many points they earned at the end of the exercises. This direct feedback guided them on which topics required further attention.

Within the e-learning exercises, students mostly needed to apply or transfer knowledge from the lectures by calculating exercises. There were also some multiple-choice questions to avoid open-ended questions.

The e-learning exercises were uploaded weekly, but it was up to the students to decide if, when, and in which order they worked on the e-learning exercises. If students crammed, they could work on all e-learning exercises in the last week or final days before the exam. Students were further allowed to retake the exercises as often as they wanted to improve their performance or refresh their memory right before the exam.

Additionally, each e-learning exercise had five different versions, i.e., students who repeated exercises did not necessarily receive the same exercise. Participating in the e-learning exercises was not connected to an additional external reward. We refrained from using external incentives because they may have undermined intrinsic motivation (Deci et al., 2001).

The duration of time students could work on each exercise was limited by a timer. Thereby, we wanted to ensure that students focused on the exercises. Additionally, the timer also resembled the setting of the exam. However, students had twice as much time to work with their learning material compared to the exam.

2.3 Participants

Data were collected in 2019 during the second of two mandatory statistics courses for social sciences students at a large German public university. Data collection was restricted

Table 1 General sample information

Specific group information of (sub) sample	Number of observations
Registered for the exam (R)	80
Attended the exam (A)	67
At least one question of the survey answered (S)	83
Worked on at least one e-learning exercise (E)	51
A∩S	55
A∩E	46

Students who did not work on the e-learning exercises could still be used for the regressions. We recoded non-participation as zero

to students who took the exam at the end of the semester. About 80 students had registered for the exam, but only 67 ultimately took the exam. Of these students, 53 answered the survey (at least partly), summarized in Table 1. More than half of the students were female (58%).

Table 2 Descriptive statistics for the outcome, variables of interest, and demographic information

	N	*Mean*	*SD*	*Min*	*Max*
Outcome					
Points on end exam	67	51.62	19.56	1.00	87.00
Practice variables					
Retrieval practice attempts	67	5.29	4.77	0.00	13.00
Practice performance	67	43.74	31.95	0.00	92.64
Mean number of trials per week	67	1.16	0.84	0.00	3.92
Spacing	67	0.94	1.13	0.00	6.00
Face-to-face tutorial preparation	67	1.46	0.86	0.00	3.27
Missed face-to-face tutorials	67	3.55	3.81	0.00	12.00
Individual characteristics (*Char*)					
Female	55	0.58	0.50	0.00	1.00
Number of semesters	55	4.24	2.05	1.00	11.00
Retaking Statistics 2	55	0.15	0.36	0.00	1.00
High school GPA	55	2.69	0.64	1.00	4.00
Standardized points in Statistics 1	67	0.16	0.77	-1.59	1.58
Exam in Statistics 1 written	67	0.93	0.26	0.00	1.00
Self-set goals (*SG*)					
Number of e-learning exercises	55	7.51	3.92	0.00	12.00
Aspired exercise performance	55	0.72	0.20	0.00	1.00
Aspired spacing	55	1.45	0.66	1.00	3.00
Aspired grade in the exam	55	2.26	0.65	1.00	4.00

The table shows the number of observations (*N*), mean, standard deviation (*SD*), minimum (Min), and maximum (Max) for each variable. Descriptive statistics for a complete case sample, including only individuals without any missing information, are reported in Table 9

2.4 Data

We collected the survey information within the first week of the course with an online survey. The data were saved when students worked on the e-learning exercises on the online-study website ILIAS. When students solved the same weekly exercise again, only the best attempt was saved. Table 2 shows the descriptive statistics of the variables we employed.

For the analysis, the outcome variable was the number of points earned in the final exam. The maximum number of points on the exam was 90. The best student earned 87 points, and the passing cut-off was 40. The *retrieval practice* variable was the sum of the weekly e-learning exercises in which students participated. Therefore, the mean of 5.20 shows that, on average, students worked on five to six e-learning exercises out of 13. Performance was assessed by mean points per exercise of the sessions the students self-tested (*performance*). For *mean number of trials per week*, we measured how often students repeated a specific exercise of any week. Although we designed five different versions for each weekly exercise, most students used only one version. Then, for *spacing,* we summed the number of times students worked on the e-learning exercises within the first two weeks of their respective publications. The mean of spacing below one shows that only a fraction of students spaced their learning, and most crammed it in the last week before the exam.

Next, we also collected the self-reported preparation for the weekly face-to-face tutorials. At the beginning of these face-to-face tutorials, students had to sign a list to prove attendance. When students signed the list in the tutorial, we additionally asked them, on a scale from 1 to 4, to what extent they had prepared themselves for the tutorial (1 = not at all, 4 = fully prepared). Thus, the variable *face-to-face tutorial preparation* was the mean preparation of students (between zero and four) over the 13 tutorial weeks. Then, the attendance rate (*missed face-to-face tutorials*) measured the number of tutorials students missed, which were two to three on average. Within the sample, some students retook the exam, and thus the mean of the dummy *Retaking Statistics 2* was slightly above zero. The mean high school GPA was about 2.6 (in Germany, the GPA ranges from 1, best, to 4, worst). For a subject-specific ability measure, we used the performance on the course *Social Science Statistics 1*, which students should have taken the semester before. We standardized the number of points for the specific exam date to make it comparable. Further, we included a variable indicating whether an individual had not yet passed the exam or whether they had not yet taken the exam.

Additionally, to the abovementioned variables, we asked students about their self-set goals for the practice and exam. We asked how many of the e-learning exercises they planned to solve, whether they planned to take them weekly, how well they wanted to perform on them, and what grade they aimed to earn on the final exam. Students wanted to complete seven to eight e-learning exercises at the beginning of the semester. Lastly, on average, students aimed for a 2.3 on the exam.

Finally, we surveyed standard measures of expectancy-value theory (Gaspard et al., 2017, adapted to the university context and course), achievement goals (Elliot & Murayama, 2008, translated and adapted for the specific context), the big five personality traits (Schupp & Gerlitz, 2014, taken as is) and present bias preferences (Frederick & Loewenstein, 2002, translated). Summary statistics and the Cronbach's α are presented in Table 3. The respective measures are further described in Tables 10 and 11. Only for the big five personality traits, Cronbach's α was below 0.7 but still above 0.6 for some constructs.

Table 3 Descriptive statistics for additional control variables

	N	Mean	*SD*	Cron. α
Expectancy-value beliefs (*EVT*)				
Self-concept	55	2.50	0.56	0.88
Intrinsic value/Dispositional interest	55	2.56	0.84	0.89
Attainment value	55	2.48	0.51	0.82
Utility value	55	3.45	0.82	0.92
Cost	55	2.15	0.70	0.82
Big five (*BF*)				
Conscientiousness	53	1.83	1.14	0.66
Extraversion	54	2.10	1.21	0.80
Agreeableness	53	3.05	0.92	0.63
Openness	53	5.17	0.97	0.68
Neuroticism	53	1.45	1.18	0.67
Achievement goals (*AG*)				
Mastery approach	53	5.69	0.99	0.75
Mastery avoidance	52	4.92	1.39	0.83
Performance approach	50	3.96	1.64	0.87
Performance avoidance	51	3.65	1.82	0.91
Present bias preferences (*PBP*)				
Risk	54	0.69	0.18	
Discount factor	52	0.93	0.24	
Present bias	52	1.15	0.66	

The table shows the number of observations (*N*), mean, standard deviation (*SD*), and Cronbach's α for each variable

3 Statistical Analysis

OLS regression was performed to estimate the relationship between practice and exam points:

$$points_i = \mu + \rho' practice_i + \beta_1' Char_i + \beta_2' EVT_i + \beta_3' AG_i + \beta_4' BF_i + \beta_5' PBP_i + \beta_6' SG_i + \varepsilon_i, \quad (1)$$

where index i stands for the individual and ε_i is the idiosyncratic error term. The outcome variable $points_i$ is the number of points of the exam. To estimate the students' practice behavior, the vector $practice_i$ included a (sub) set of the practice variables introduced in Table 2. Confounders may have influenced the practice variables. For example, motivation might have increased in additional practice and exam points. Thus, the practice variables might have not only measured the practice effect but also included the underlying motivation. Therefore, we included variables in $Char_i, EVT_i, AG_i, BF_{i,}$, *and* SG_i as presented in Table 2.

However, we faced the problem of too many variables per student. Hence, we used variable selection methods to achieve a sufficient sparse set of control variables. We followed the double selection procedure introduced by Belloni et al. (2014) for this purpose. Their suggestion was a two-stage selection procedure: First, variables were selected that explained exam points and all practice variables. Thereby, we acquired a sparse set of essential variables from $Char_i, EVT_i, AG_i, BF_{i,}, PBP_i$, *and* SG_i explaining students' exam points and practice behavior. Second, we ran an OLS regression that included the

pre-selected variables and the practice variables on exam points. Assuming that the most important variables were surveyed in the first place, we interpreted the estimated practice coefficients cautiously causally. We followed Belloni et al. (2014) and used the machine learning method LASSO for the feature selection. LASSO selects variables by imposing the L1 penalty $\lambda[\Sigma_i(|\beta_i|)]$ on the regression coefficients. This penalty sets some coefficients to be exactly zero, effectively removing the corresponding predictors from the model. The amount of shrinkage applied to the coefficients is controlled by the λ tuning parameter. We follow the standard rule and use cross-validation to find the λ that is 1 standard error higher than the minimizing λ to prevent the model from over-fitting (Friedman et al., 2001). By setting the coefficients to zero, LASSO is useful for situations where there are many predictors and only a subset of them is relevant. However, it can struggle with highly correlated predictors. The Elastic Net combines LASSO and Ridge Regression to overcome LASSO's difficulty with highly correlated predictors. It is a hybrid of these two methods, including additionally the Ridge penalty $\lambda[\Sigma_i(\beta_i\prime\beta_i)]$, also called L2-penalty. Like LASSO, Elastic Net can generate models with zero coefficients, resulting in sparse selection. However, it also incorporates the penalty of ridge regression, which helps handle situations with highly correlated variables (Hastie et al., 2009). The Elastic Net equation is as follows:

$$points_i = \mu + \rho' practice_i + \beta_1' Char_i + \beta_2' EVT_i + \beta_3' AG_i + \beta_4' BF_i + \beta_5' PBP_i + \beta_6' SG_i + \varepsilon_i + \lambda \left[\sum_i \left((1-\alpha)\beta_i'\beta_i + \alpha|\beta_i| \right) \right], \quad (2)$$

in which λ is the penalty weight and α is the weight for either Ridge (L_2 normalization: $\beta_i\prime\beta_i$) or LASSO penalization (L_1 normalization: $|\beta_i|$). Hence, Eq. (2) reduced to LASSO with $\alpha = 1$ and to Ridge if $\alpha = 0$. For the selection, we chose a value $\alpha = \{1, 0.8, 0.6, 0.4, 0.2\}$, i.e., using LASSO as well as elastic with varying mixture between LASSO and elastic net. We did not choose $\alpha = 0$ because Ridge does not help select a sparse set of variables.

For the post-double selection regressions, we used multiple imputations to include all 67 observations as some students did not respond to all questions in the survey. Therefore, we used 100 imputations and then pooled the results. While the standard in educational literature using *R* is the package *mice* (Lüdtke et al., 2017), we used classification and a tree-based method. Akande et al. (2017) and Murray (2018) reasoned against using *mice PMM* because it is too inflexible and recommended using tree-based methods instead, especially for mixed data types and non-linear interactions between variables, as well as to cope with high-dimensional data. Further, Madley-Dowd et al. (2019) showed that using tree-based methods reduces bias even when the proportion of missing values is large. Therefore, we applied *missForest* (Stekhoven & Bühlmann, 2012) to all variables. MissForest is a random forest-based imputation method. It is used to handle missing values in data sets containing different types of variables. By averaging over unpruned classification or regression trees, missForest performs multiple imputations. It estimates the imputation error using the out-of-bag error estimates of random forest, eliminating the need for a test set (Stekhoven & Bühlmann, 2012). Descriptive statistics did not reveal any differences between the sample of complete and incomplete cases. Tables comparing both samples are available upon request.

Correlation plot

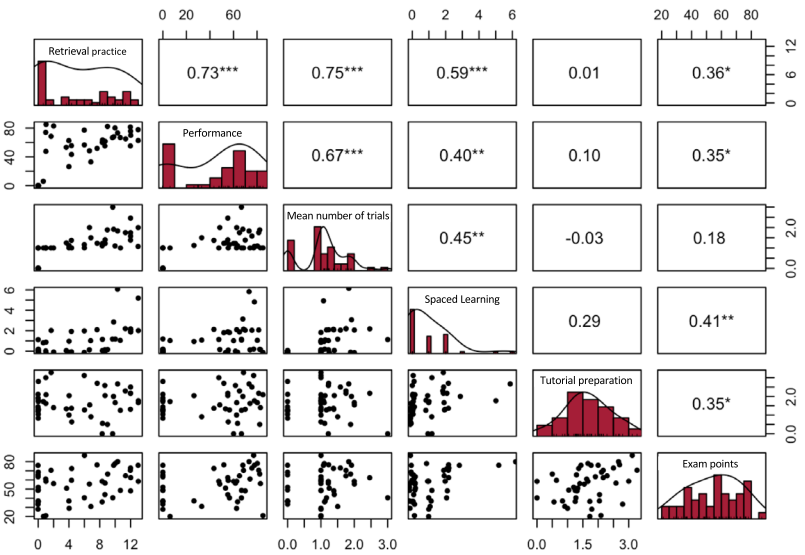

Fig. 1 Correlation plot
Note: The diagonal shows the distribution of the respective one-dimensional distributions. The lower half shows the two-dimensional scatterplot, and the upper half shows the correlation $^{*}p < 0.05$,$^{**}p < 0.01$,$^{***}p < 0.001$

4 Results

4.1 Participation in Exercises and Correlates

First, the results of the correlation analyses are shown. Figure 1 illustrates the relationship between five practice variables and exam grades. The correlation between (the number of) retrieval practice attempts and (the mean) performance (in the retrieval practice attempts) was high ($r = 0.73$) because students who never participated in the e-learning exercises also had no performance and we did not observe students with high participation and low performance in these exercises. This was similar to the number of trials that correlated with retrieval practice attempts ($r = 0.75$) and the respective performance ($r = 0.67$). The correlation between spaced learning and retrieval practice was $r = 0.59$, between spaced learning and performance was $r = 0.40$, and between spaced learning and the mean number of trials per week was $r = 0.45$.

The correlation between face-to-face tutorial preparation and retrieval practice ($r = 0.08$), performance ($r = 0.16$) and spacing ($r = 0.29$) was very low and insignificant, and even negative for the mean number of trials per week ($r = -0.03$). All practice variables were positively correlated with exam points, whereas only the mean number of trials per week was statistically insignificant.

Regarding RQ1, we found that about two-thirds of the students who attended the exam used the e-learning exercises to practice for the exam (RQ1a). Most students, however, used the e-learning exercises to repeat the topics right before the exam

and did not space out their learning (RQ1b). Figure 1 also provides initial evidence of the positive relationship between more retrieval practice and spaced learning and exam performance (RQ2). Within the next section we focus on the practice variables retrieval practice attempts, spaced learning and mean FTF tutorial preparation, as these variables are the most important predictors in multivariate regressions (see Table 12). Additionally, Table 13 shows that the results are also robust when we control for the selection into using the e-learning exercises at least once. Since the regression results are robust, we exclude this variable in the subsequent analysis.

4.2 Effects of Retrieval Practice Variables on Exam Performance

Table 4 presents the regression results for the practice variables on the end exam points without any additional control variables. The first column includes only the variable *retrieval practice attempts* to show whether the retrieval practice with several e-learning exercises predicts more points in the end exam. The coefficient is equal to 1.917 and is highly significant. Thus, students who practiced one additional e-learning session increased their points on average by around 2 points. Since there were 13 sessions, students with full participation improved their grades by 24.92 points, equaling more than one entire grade. In column (2), we include the *mean FTF tutorial preparation*, which reduces the *retrieval practice attempts* coefficient to 1.695 to proxy students' offline practice behavior. Next, the mean of the *number of trials per practice* of the participated e-learning sessions in column (3) does not substantively change the regression, and the coefficient itself is statistically insignificant. This missing significance is likely due to the very low variation already reported in Table 1. Lastly, once we include students' spacing during the semester in column (4), the *retrieval practice attempts* coefficients decrease again slightly to 1.239. Additionally, the coefficient for *spaced learning* is equal to 2.866 and significant at the 10 percent level. Working on one additional e-learning exercise within two weeks after publication would increase the exam points

Table 4 Main practice regression—sequential inclusion of practice variables

	Dependent variable: Points on end exam			
	(1)	(2)	(3)	(4)
Retrieval practice attempts	1.917*** (0.430)	1.695*** (0.421)	1.919*** (0.574)	1.239** (0.495)
Mean FTF tutorial preparation		6.868*** (2.213)	6.742*** (2.246)	6.203*** (2.269)
Number of trials per practice			− 1.713 (3.142)	
Spaced learning				2.866* (1.487)
Constant	41.485*** (3.511)	32.648*** (4.204)	33.639*** (4.896)	33.331*** (4.206)
Observations	67	67	67	67
R^2	0.219	0.306	0.309	0.329
Adjusted R^2	0.207	0.285	0.276	0.297

Heteroskedastic robust standard errors in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

by almost 3 points. Since the adjusted R^2 is highest for column (4), which included *retrieval practice attempts*, *mean FTF tutorial preparation,* and *spaced learning*, we will focus on these practice variables from now. The estimated coefficients above should be interpreted cautiously because they could be biased due to omitted variables. Therefore, we add additional control variables in the following subsection. Since spaced learning is additional information on how students self-tested, we will first look at post-double selection regression without spacing in Table 5 and include spacing in Table 6.

4.3 Post-double Selection Regression Results

In the second step, the results of the post-double selection regression analyses were reported. The selected control variables are each a subset of the selected variables of the column to the left. This means that LASSO selected most variables, and the subsequent Elastic Net picked a subset of these variables. Introducing additional control variables in Table 5 columns (1) to (5) showed a reduced but stable effect between 1.25 and 0.99 for the retrieval practice attempts. Thus, the retrieval practice effect was almost halved but still robustly statistically significant and important, even after including a rich set of control variables. Furthermore, the adjusted R^2 increased from 0.285 in Table 4 column (2) without covariates up to 0.610 in Table 5 column (2). Thus, the covariates explained a slightly higher amount of the variance in the dependent variable than that in the practice variables. This high adjusted R^2 in Table 4 is further reassurance that we captured important variables explaining exam grades, making it less likely that the estimated effect was driven solely by unobserved selection.

For the FTF tutorial preparation effect, including the control variables led to a meaningful change. The effect decreased to 2.917 in column (5) and was no longer significant. In contrast to the retrieval practice, this implies that the estimation in Table 4 column (2) was upward-biased and partly explained by our rich set of control variables. The preparation might still be beneficial, but the effect was driven by a selection into more preparation.

Table 5 Post-double selection regression results without spacing

	Dependent variable: Points on end exam				
	(1)	(2)	(3)	(4)	(5)
	α = 1	α = 0.8	α = 0.6	α = 0.4	α = 0.2
Retrieval practice attempts	1.251*** (0.433)	1.230*** (0.426)	1.133*** (0.412)	1.093*** (0.410)	0.989*** (0.348)
Mean FTF tutorial preparation	3.672 (2.536)	3.817 (2.365)	4.241* (2.036)	3.944* (1.893)	2.917 (1.965)
Constant	20.407 (36.239)	7.754 (34.383)	7.609 (39.613)	28.519 (33.833)	17.915 (16.049)
Add. Control Var	22	21	19	15	6
Observations	67	67	67	67	67
R^2	0.749	0.746	0.716	0.687	0.618
Adjusted R^2	0.606	0.610	0.583	0.578	0.566

α refers to the elastic net parameter for the post-double feature selection. The elastic net is a mixture of LASSO and Ridge, and the value for α, which must be between 0 (Ridge) and 1 (LASSO) defines the mixture between LASSO and Ridge. Heteroskedastic robust standard errors in parentheses. The full estimation results are reported in Table 14. Results are robust if multiple imputation is not used and are reported in the appendix, Table 15. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Table 6 Post-double selection regression results with spacing

	Dependent variable: Points on end exam				
	(2)	(3)	(4)	(5)	(6)
	α = 1	α = 0.8	α = 0.6	α = 0.4	α = 0.2
Retrieval practice exercises	0.809* (0.486)	0.794* (0.481)	0.624 (0.439)	0.748* (0.446)	0.690* (0.379)
Mean FTF tutorial preparation	3.191 (2.374)	3.356 (2.288)	3.710* (2.184)	3.436 (2.298)	2.500 (2.171)
Spacing	2.952* (1.517)	2.905* (1.581)	3.598** (1.614)	2.185* (1.246)	2.046* (1.048)
Intercept	18.668 (37.185)	5.053 (36.454)	5.227 (40.215)	26.532 (33.675)	20.600 (16.095)
Add. Control Var	22	21	19	15	6
Observations	67	67	67	67	67
R^2	0.765	0.762	0.742	0.699	0.630
Adj. R^2	0.622	0.625	0.612	0.586	0.571

Heteroskedastic robust standard errors in parentheses. The full estimation results are reported in Table 14. Results are robust if multiple imputation is not used and are reported in Table 16. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Next, we re-ran the post-double selection regression, including the spacing variable in the post-regression. The results are presented in Table 6. First, for the retrieval practice attempts, the coefficient decreased to between 0.624 in column (4) and 0.809 in column (2) and was only significant at the 10% level (except column 3). Thus, one additional weekly e-learning self-test would only yield an increase of around 0.7 or 0.8 points. However, the reduction in the retrieval practice coefficient due to the spacing variable might have occurred for two reasons. The first is that retrieval practice is necessary for spacing. Thus, the variable *spacing* captures part of the retrieval practice effect as depicted by the high correlation shown in Fig. 1. Without the retrieval practice, there would not been any spacing in our model. Second, the number of observations might have been too small for both practice variables and additional control variables. Additionally, the spacing coefficient was between 2.046 (column 5) and 3.598 (column 3), significant at the 5% or 10% level. Therefore, we conclude that retrieval practice with the help of our weekly e-learning exercises is helpful and even more so if students' practice is spaced out during the semester.

Table 7 shows which variables were selected by the respective elastic nets, which directions they had and their significance level. Prior achievement measured by the standardized grade for Statistics I, self-concept, utility value, performance-avoidance, conscientiousness, and neuroticism are always selected for all specifications. The standardized grade for Statistics I, utility value and mastery approach, retaking Statistics II, present-bias, and openness also have a particularly high predictive power, shown by a robust significant effect, in the specifications they are selected in. The results support that these variables are complements rather than substitutes, as they are each selected. This is also in line with Plante et al. (2013) for EVT and achievement goals, as well as Becker et al. (2012) for personality and time preferences. More specifically, prior achievement in Statistics I was always selected and always had a positive statistically significant relation with exam performance. Retaking Statistics II and openness, if they were selected, also had a positive statistically significant relation. While students' utility value was always selected, it demonstrated a positive statistically significant relation with

Table 7 Feature selection results

Selected variables	Elastic net feature selection with varying α				
	α = 1	α = .8	α = .6	α = .4	α = .2
From individual characteristics					
Statistics 1 grade	+***	+***	+***	+***	+***
Retaking statistics 2	+**	+**	No	No	No
Female	−	−	−	−	No
Semesters	+	+	+	+	No
From expectancy-value beliefs					
Self-concept	+	+	+	+	+
Utility value	+**	+**	+	+**	+
Attainment value	+	+	+	+	No
Intrinsic value	−	−	−	No	No
Costs	−	−	+	+	No
From achievement goals					
Mastery approach	−***	−***	−*	−*	No
Performance approach	+	+	−	No	No
Performance avoidance	−	−	−	−	−
From big five					
Agreeableness	−	No	No	No	No
Conscientiousness	+	+	−	+	−
Neuroticism	+*	+*	+	+	+
Openness	+***	+***	+***	No	No
From time preferences					
Present-bias	−**	−**	−***	−***	No
Discount factor	−	−	No	+	No
From self-set course goals					
Number of e-learning exercises	+	+	+	+	No
Aspired exercise performance	+	+	+	No	No
Aspired spacing	+	+	+	No	No
Aspired grade in the exam	−	−	−**	−*	No

α refers to the elastic net parameter which adjusts the mixture of LASSO (variable selection) and RIDGE (coefficient reduction towards zero). Particularly if the number of covariates is not much smaller than the number of observations, RIDGE first reduces the coefficients which leads to a lower set of selected variables

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$

exam performance only in three out of five instances. Students' mastery approach and present bias preferences exhibited a negative statistically significant relation with exam performance in all post-double selection regression except for α = 0.2. Additionally, the negative statistically significant relation of the aspired grade in the exam in two specifications means that students who aspired a better (= lower) grade had a better exam performance.

It is important to note that deviations from the literature could be driven by the high number of control variables. Although not the primary scope of this analysis, it is possible to conduct regressions per variable group to examine if all variables go into the expected

direction. For example, when considering the negative impact of mastery approach, when also including students' self-concept and prior achievement, it may be because students with high self-concept and high prior achievement also have a high mastery approach. In such cases, self-concept and prior achievement might account for most of the explanatory power, leaving the remaining explanatory power of mastery approach to exhibit a negative effect. This could mean that students with lower self-concept and lower prior achievement who are still determined to master every topic may have lower exam points.

5 Discussion, Limitations and Conclusion

This study analyzed the effect of voluntary, non-rewarding e-learning exercises on students' exam points at the end of the semester. In a university inference statistics course, additional exercises were offered to undergraduate social science students to practice the topics of the lectures and tutorials. Students' practicing behavior was analyzed with regard to the frequency and spacing of the usage of these exercise. Our study results highlight the potential of e-learning tools in higher education teaching. Particularly, we found that taking part in additional e-learning exercises improves students' achievement. In contrast to most studies in this area, which were solely based on surveys (O'Brien & Verma, 2019), we added to the few examples of Förster et al. (2018) and Schwerter, Dimpfl et al. (2022) who additionally collected data within e-learning environments. Thus, we made use of the new possibilities of learning analytics to improve teaching and learning (Hellings & Haelermans, 2020).

Our study provided some proof for the saying 'practice makes perfect' in a natural educational environment to the extent that students benefited from more retrieval practice attempts in the additional e-learning exercises. Most students used our designed e-learning practice (RQ1a); however, in line with the literature on procrastination (Bisin & Hyndman, 2020), only a fraction of students spaced out their use of the exercises (RQ1b). Students also rarely used different versions of the same exercise, but practiced a self-test only once per weekly topic (if at all), which is in line with current literature showing that additional practice opportunities are rarely used (Tullis & Maddox, 2020) (RQ1c). For future research, it would be worthwhile to investigate whether this selective usage of retrieval practice depends on students' self-estimated abilities to remember the information. Earlier research showed that learners do not continue to study the best-learned information, but instead restudy the worst learned content; however, they selectively test the learned content (Karpicke, 2009).

The positive effect of retrieval practice attempts on exam performance (RQ2a) confirms general e-learning practice literature for lower-order learning, using quizzes (Collins et al., 2018; Landrum, 2007; Panus et al., 2014), and higher-order learning (Förster et al., 2018; Schwerter, Dimpfl et al., 2022), as well as the general retrieval practice literature (Baker et al., 2020; Hartwig & Dunlosky, 2012; Park et al., 2018; Rodriguez et al., 2021a, 2021b). More specifically, retrieval practice with one additional weekly e-learning exercise increased the student's final exam points in our study by 1 to 2 points. Our results do not only confirm Förster et al.'s (2018) study, they also extend it by including various important predictors of student achievement (such as motivation, personality traits, time preferences and goals). After including these control variables, the effects of additional practice were reduced but remained statistically significant and of importance (RQ3). Overall, our results confirm that with the help of digital technology, in particular, online quizzes, students can learn more efficiently and effectively (Morrison & Anglin, 2005) and are most likely to retrieve more information

(Roediger III & Karpicke, 2006). The results are particularly interesting as social science students are known to have trouble with statistics (Vaessen et al., 2017). They could, thus, also be used to design interventions to counter-act statistics anxiety.

Several factors are likely to drive the positive effect of retrieval practice on exam performance. First, practice leads to a more efficient encoding of the information to be retrieved, stored and/or recalled (Jonides, 2004; Roediger III & Karpicke, 2006). Second, experiencing knowledge gaps can lead to additional learning to fill these gaps (Karpicke, 2009). This additional learning results in the potentiation effect (Hays et al., 2013). Third, even if students failed to recall how to solve the problem correctly, students might learn from the error they made when solving the respective exercise (Kornell et al., 2009). Third, in this study, students also received knowledge of correct response feedback immediately after each self-test. This most likely added to the positive effect of the retrieval practice since research on feedback has demonstrated its effect in correcting errors or misconceptions (Hattie & Timperley, 2007; Wisniewski et al., 2020) and its general effect on student achievement (Attali, 2015; Attali & van der Kleij, 2017). Thus, the feedback likely increased the error generation effect (Kornell et al., 2009) as students already learned about their mistakes immediately after the self-test. Additionally, feedback may have also helped guide students in their learning (Kirschner et al., 2006).

Finally, our results showed that students who spaced out the self-tests had an additional benefit in their learning, i.e., one additional exercise done within the respective week yielded around three additional points (RQ2b). This can be explained by students' deeper processing of the content, particularly if their learning was spread out over the whole semester (Collins et al., 2018; Jonides, 2004). Especially at the beginning of the semester, doing the additional exercises might help students follow the upcoming weeks' topics, explaining this large effect. This also relates to the error generation effect (Kornell et al., 2009) mentioned above. Students who spaced out their learning might have benefited more from the following lectures as they built on the past topics. Also, the weekly topics built on each other, which is why practicing during the semester also meant some repetition of topics from earlier weeks. In this regard, it is noteworthy that "using retrieval practice selectively for well-learned information, rather than for all information, may be the most effective use of retrieval practice because benefits of testing occur only when students successfully retrieve information" (Tullis & Maddox, 2020 p. 140). Students benefited from the forgetting curve, i.e., the repetition of earlier study topics helped reinforce memory traces. However, the interpretation of the spacing effect is limited since only one student managed to work on 6 of 13 exercises within the first two weeks after publication. The relatively low spacing realizations, in turn, align with students' well-known procrastination behavior in HE (Denny et al., 2018). Altogether, our results showed that the survey and intervention results in retrieval practice and spacing can be transferred to natural educational settings in which additional e-learning exercises indirectly promote spacing.

Our results of selected covariates are mostly in line with the literature as follows: Prior achievement (Förster et al., 2018; Rodriguez et al., 2021a, 2021b), utility (Brisson et al., 2017; Gaspard et al., 2017; Wigfield & Eccles, 2020), openness (Ziegler et al., 2018), and higher exam goals (van Lent & Souverijn, 2020) are important positive predictor for students exam achievement. Additionally, as higher present-bias preferences are known to explain students probability to procrastinate (Bisin & Hyndman, 2020), we find a negative relation to exam performance. Lastly, we add to the mixed results of the direction of mastery approach: Contrary to Elliot et al. (1999) and Harackiewicz et al. (2002) but in line with Plante et al. (2013), we find a negative effect of mastery approach.

This study is limited given the relatively small sample. Also, we observed only one cohort of social science students in one university, which, simultaneously, meant that we

did not have a teacher or an institutional effect that needed to be controlled. Nevertheless, although the external validity was already enhanced by the natural setting of the study, more research with a larger sample is needed to replicate the results. A larger sample would also enable us to better estimate the effects of retrieval practice attempts, the number of trials per week, the respective performance, and its development and spacing. Furthermore, we only measured students' preparation for the weekly face-to-face tutorial, which was supposed to capture students' non-digital learning behavior. However, this variable was only self-reported and did not capture additional learning outside the e-learning environment. The literature shows that measurement error is a potential problem in student self-report measures. However, it is more likely to occur when students provide sensitive information such as GPAs (Wilson & Zietz, 2004) and when responding to items that address the main topic of the survey (Brenner & DeLamater, 2017). We would argue that (a) the question whether students prepared for the face-to-face tutorial is not sensitive, and (b) it was not the primary focus of the survey. Therefore, we expect self-reported measurement error to be low in this context. In our setting, there was no possibility of assessing it in any other way.

There are two concerns when interpreting the positive effects of retrieval practice. Given our design and the ethical requirements, we could not conduct an RCT study that would have allowed some students access to the retrieval practice exercises while not allowing others to do so. Further, we could not distinguish the testing-enhancing effect and the respective feedback. Future studies could use an RCT to determine the importance of feedback when students self-test. Altogether, we add to the literature on retrieval practice, which thus far has mainly used surveying or promoting study techniques: e-learning exercises that promote retrieval practice and include feedback that help students learn and results in higher achievement. Although limited by contextual constraints, such as the lack of a traditional control group due to ethical concerns, the study demonstrated the added value of additional e-learning exercises in a natural, i.e., 'noisy', setting. Given the problems of replicating experimental research, it is important to show the robustness of the effects in different contexts.

Since only about two thirds of our students used the additional practice opportunities, it is noteworthy to reflect upon the practical implications of our research. We showed—in line with previous research—that additional practice improves students' exam performance. Also spacing out the participation further enhances students' learning. Thus, future (statistics) courses could be designed in such a way that students are motivated to (a) utilize the e-learning exercise and to (b) space out their learning to benefit from the potentiation effect. In our study, we refrained from providing further incentives for students to participate in the additional practice to avoid crowding out intrinsic motivation. Nevertheless, providing students with reasoning on why practice is important might already support their participation and reduce procrastination. Ideally, such e-learning exercises can also support faculty (or teaching assistants in higher education) and help them to support students in their learning.

Appendix

See Tables 8, 9, 10, 11, 12, 13, 14, 15 and 16.

Table 8 Semester structure

Week	Date of lecture	Topics	Tutorial
1	2019-10-14	Probability (1)	Pretest and survey
2	2019-10-21	Discrete random variables (2)	Exercise sheet 1
3	2019-10-28	Continuous random variables (2)	Exercise sheet 2
4	2019-11-04	Specific discrete distributions (3)	Exercise sheet 3
5	2019-11-11	Specific continuous distributions (3)	Exercise sheet 4
6	2019-11-18	Two-dimensional distributions (4)	Exercise sheet 5
7	2019-11-25	Theorems and sample mean (5)	Exercise sheet 6
8	2019-12-02	Point estimation (6)	Exercise sheet 7
9	2019-12-09	Interval estimation (7)	Exercise sheet 8
10	2019-12-16	Statistical testing and p-value (8)	Exercise sheet 9
	2020-01-06	National holiday (affected only the lecture)	Old exam questions
11	2020-01-13	Rep. (6), (7) and (8), and introduction of (9)	Exercise sheet 10
12	2020-01-20	Regression analysis (9)	Exercise sheet 11
13	2020-01-27	Regression analysis, examples (9)	Stata Session
14	2020-02-03	Question session	Old exam questions
15	2020-02-10	Exam	

The tutorials were held on Wednesday and Thursday of the same week of the corresponding lecture. Week 11 includes a repetition of topics (6), (7), and (8) to show the connection of the topics and why they are essential for topic (9). Every exercise sheet had an additional e-learning exercise. There were thirteen e-learning exercises

Table 9 Descriptive Statistics for full and complete observation samples

	Full sample				Complete observations			
	N	Mean	SD	Cron. α	N	Mean	SD	Cron. α
Outcome								
Points in end exam	67	51.62	19.56		46	53.08	20.03	
Practice variables								
Retrieval practice	67	5.29	4.77		46	5.40	4.82	
Practice performance	67	43.74	31.95		46	45.33	31.85	
Number of tries per practice	67	1.16	0.84		46	1.08	0.71	
Spacing	67	0.94	1.13		46	0.96	1.33	
Face-to-face tutorial preparation	67	1.46	0.86		46	1.64	0.78	
Missing dates face-to-face tutorial	67	3.55	3.81		46	2.61	2.82	
Individual characteristics (*Char*)								
Female	55	0.58	0.50		46	0.54	0.50	
Number of semesters	55	4.24	2.05		46	4.35	2.15	
Retaking Statistics 2	55	0.15	0.36		46	0.15	0.36	
High school GPA	55	2.69	0.64		46	2.60	0.63	
Standardized points in Statistics 1	67	0.16	0.77		46	0.32	0.73	
Exam in Statistics 1 written	67	0.93	0.26		46	0.93	0.25	
Expectancy value theory (*EVT*)								
Self-concept	55	2.50	0.56	0.88	46	2.40	0.40	0.88
Intrinsic value/Dispositional Interest	55	2.56	0.84	0.89	46	2.46	0.76	0.90
Attainment value	55	2.48	0.51	0.82	46	2.42	0.38	0.86
Utility value	55	3.45	0.82	0.92	46	3.38	0.81	0.93
Cost	55	2.15	0.70	0.82	46	2.09	0.62	0.83
Big five (*BF*)								
Conscientiousness	53	1.83	1.14	0.66	46	1.80	1.18	0.73
Extraversion	54	2.10	1.21	0.80	46	2.09	1.22	0.77
Agreeableness	53	3.05	0.92	0.63	46	3.06	0.98	0.66
Openness	53	5.17	0.97	0.68	46	5.25	1.01	0.80
Neuroticism	53	1.45	1.18	0.67	46	1.38	1.15	0.60
Achievement goals (*AG*)								
Mastery approach	53	5.69	0.99	0.75	46	5.64	1.01	0.81
Mastery avoidance	52	4.92	1.39	0.83	46	4.99	1.29	0.79
Performance approach	50	3.96	1.64	0.87	46	3.96	1.67	0.89
Performance avoidance	51	3.65	1.82	0.91	46	3.64	1.79	0.91
Present bias preferences (*PBP*)								
Risk	54	0.69	0.18		46	0.70	0.18	
Discount factor	52	0.93	0.24		46	0.94	0.25	
Present bias	52	1.15	0.66		46	1.16	0.70	
Self-set goals (*SG*)								
How many e-learning exercises	55	7.51	3.92		46	7.48	3.99	
How good in the e-learning exercises?	55	0.72	0.20		46	0.72	0.18	
Solving the e-learning exercises weekly?	55	1.45	0.66		46	1.50	0.69	
Which grade in the exam?	55	2.26	0.65		46	2.30	0.69	

The table shows the descriptive statistics of all variables for two different sample sets. The first one shows the full sample with a different number of observations per variable due to missing information of the stu-

Table 9 (continued)

dents on some of these variables. The second includes only the students for which complete information is given. Comparing both samples does not reveal a clear selection into missingness

Table 10 Example items

Variables	Number of items	Example item
Expectancy value theory (*EVT*)		
Self-concept	4	I am good in statistics
Intrinsic value	4	Statistics makes fun
Attainment value	4	Statistics are of no importance to me
Utility value	4	Good knowledge of statistics will help me in my later career
Cost	6	Dealing with statistics costs me a lot of energy
Big five (*BF*)		
		I am someone who …
Conscientiousness	3	works thoroughly
Extraversion	3	… communicative, talkative
Agreeableness	3	… is sometimes a bit rough with others
Openness	3	… is original, comes up with new ideas
Neuroticism	3	… often worries
Achievement goals (*AG*)		
Mastery approach	3	My aim is to completely master the material presented in statistics
Mastery avoidance	3	My goal is to avoid learning less in statistics than I could
Performance approach	3	I strive to do well in statistics compared to other students
Performance avoidance	3	My goal is to avoid doing poorly in statistics compared to other students

Table 11 Present bias preferences (PBP)

Question 1
Imagine following situation:
- Option A: You get 50€ right away
- Option B: You get 100€ right away with a probability of p_1, or 0€ with a probability of 1-p_1

With which value of p_1 between 0 and 100 you would start to prefer Option B?

Question 2
Imagine following situation:
- Option C: You get 50€ right away
- Option D: You get 100€ in 8 weeks with a probability of p_2 or 0€ with a probability of 1-p_2

With which value of p_2 between 0 and 100 you would start to prefer Option D?

Question 3
Imagine following situation:
- Option E: You get 50€ in 8 days
- Option F: You get 100€ in 16 weeks with a probability of p_3 or 0€ with a probability of 1-p_3

With which value of p_3 between 0 and 100 you would start to prefer Option F?

Variable generation:
- Risk = $p_1/100$; with Risk > 0.5 implies risk aversion
- Discount factor = p_1/p_3; with Discount factor < 1 means someone is impatient
- Present bias = p_3/p_2; with Present bias < 1 makes individuals more impatient when the present is involved

Table 12 Regression results including practice variables sequentially

	Dependent variable: Points on end exam						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Retrieval practice attempts	1.917*** (0.430)	1.695*** (0.421)	1.247* (0.712)	1.919*** (0.574)	1.698*** (0.425)	1.239** (0.495)	0.925 (0.805)
Mean FTF tutorial preparation		6.868*** (2.213)	6.463*** (2.176)	6.742*** (2.246)	7.227** (2.897)	6.203*** (2.269)	5.899** (2.758)
Retrieval practice performance			0.089 (0.120)				0.157 (0.131)
Number of trials per practice				−1.713 (3.142)			−4.303 (3.269)
Missing dates FTF tutorial					0.139 (0.650)		0.302 (0.679)
Spaced learning						2.866* (1.487)	3.94** (1.379)
Constant	41.485*** (3.511)	32.648*** (4.204)	31.717*** (4.513)	33.639*** (4.896)	31.614*** (6.795)	33.331*** (4.206)	32.832*** (7.022)
Observations	67	67	67	67	67	67	67
R^2	0.219	0.306	0.314	0.309	0.307	0.329	0.354
Adjusted R^2	0.207	0.285	0.282	0.276	0.274	0.297	0.289

Table 12 shows the regression results for the exercise variables on the final exam scores without additional control variables. The first column contains only the number of retrieval practice attempts to show whether more self-tests with e-learning exercises predicts more points in the final exam. In column (2), the average preparation for the FTF tutorial is taken into account, which leads to a slight decrease in the retrieval practice coefficient. Next, we included the performance of the e-learning exercises. The inclusion leads to a slight decrease in the coefficient for retrieval practice. Only retrieval practice and preparation are significant, implying that retrieval practice might be sufficient for learning, regardless of how well they completed the tasks. Including the mean of the number of tries for each weekly self-testing e-learning session attempt in column (4) does not change the regression significantly, and the coefficient itself is not statistically significant. This lack of significance could either be an indication of the unimportance of repeating the same exercise or, which we think is more likely, could be due to the low variation in this variable. The estimate remains robust when we include the attendance rate at tutorials in column (5) and whether students spaced out their learning in column (6). In the last column (7), all exercise variables are included in the regressions, with the coefficient for retrieval practice slightly below 1.0 and no longer significant

Since the correlation between the e-learning practice variables in Fig. 1 was so high, we suspect that the number of observations is not sufficient to measure the coefficients in their entirety. Since the adjusted R^2 is highest for column (6), which includes retrieval practice, mean tutorial FTF preparation and spaced learning, we focus on these practice variables from now on. Heteroskedastic robust standard errors in parenthesis. *$p<0.1$; **$p<0.05$; ***$p<0.01$

Table 13 Regression results including whether people have practised once

	Dependent variable: Points on end exam							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ever practiced once	3.449 (7.382)	1.463 (7.469)	−4.171 (6.100)	−5.426 (6.063)	−5.444 (8.284)	−6.919 (8.380)	−9.034 (7.609)	−9.724 (7.545)
Retrieval practice (attempts)	1.674*** (0.640)	1.112 (0.693)	1.078* (0.573)	1.585*** (0.547)	1.706** (0.740)	1.425* (0.819)	1.138 (0.759)	1.482** (0.723)
Spaced learning		0.542** (0.266)	0.479** (0.211)			0.401 (0.356)	0.335 (0.269)	
Mean FTF tutorial preparation		5.959** (2.351)	2.357 (2.028)	3.121 (2.045)		6.764* (3.540)	5.783* (3.210)	7.003** (2.840)
Standardised points in statistics 1			15.237*** (3.201)	15.299*** (3.213)			13.230*** (4.720)	13.138*** (4.666)
Exam in statistics 1 written			−5.679 (9.866)	−5.208 (9.924)			0.486 (11.918)	2.544 (11.880)
Constant	40.350*** (4.391)	33.198*** (4.712)	45.664*** (10.691)	44.808*** (10.833)	49.542*** (4.511)	38.745*** (6.635)	38.317*** (13.868)	34.756** (13.610)
Observations	67	67	67	67	46	46	46	46
Adjusted R^2	0.197	0.289	0.537	0.523	0.103	0.218	0.422	0.423

The variable 'ever practiced' includes selection into the e-learning exercises. The regression results show that the estimates for the total sample (column (1) to (4)) and the full case sample (column (5) to (8)) are robust compared to Table 12 and Tables 4 and 5. We only added standardised statistic 1 points as a control variable because it was shown to be particularly important in the feature selections. Together with the fact that the corresponding coefficient for selection is not significant, we conclude that our results should not be affected by students sorting them into using the e-learning exercises. We have only included prior achievement here as an additional control variable, as it was most significant in the feature selections. Heteroskedastic robust standard errors in parenthesis. *$p<0.1$; **$p<0.05$; ***$p<0.01$

Table 14 Full post-selection regression table

	Dependent variable: Points on end exam									
	Without spacing					With spacing				
	α = 1	α = 0.8	α = 0.6	α = 0.4	α = 0.2	α = 1	α = 0.8	α = 0.6	α = 0.4	α = 0.2
Retrieval practice attempts	1.251** (0.433)	1.230** (0.426)	1.133** (0.412)	1.093** (0.410)	0.989** (0.348)	0.809^{+} (0.486)	0.794^{+} (0.481)	0.624 (0.439)	0.748^{+} (0.446)	0.690^{+} (0.379)
Spacing						2.952^{+} (1.517)	2.905^{+} (1.581)	3.598* (1.614)	2.185^{+} (1.246)	2.046^{+} (1.048)
FTF preparation	3.672 (2.536)	3.817 (2.441)	4.241^{+} (2.405)	3.944^{+} (2.311)	2.917 (2.175)	3.191 (2.374)	3.356 (2.288)	3.710^{+} (2.184)	3.436 (2.298)	2.500 (2.171)
Stand. points in statistics 1	12.768*** (3.176)	12.582*** (3.105)	11.968*** (3.286)	12.323*** (3.033)	12.133*** (2.957)	12.636*** (3.188)	12.438*** (3.125)	11.844*** (3.268)	12.344*** (3.032)	12.194*** (2.964)
Self-concept (EVT)	3.663 (5.030)	4.272 (5.024)	4.092 (5.333)	3.225 (3.941)	4.197^{+} (2.454)	3.994 (4.903)	4.646 (4.956)	4.677 (5.159)	3.074 (3.871)	3.857 (2.468)
Utility value (EVT)	9.437* (3.752)	8.756* (3.742)	6.553^{+} (3.489)	8.651* (3.362)	6.313 (3.926)	8.165* (3.584)	7.451* (3.634)	5.415 (3.507)	8.454* (3.387)	5.972 (3.925)
Female (Char)	−1.210 (5.416)	−0.430 (5.362)	0.862 (5.489)	−0.845 (5.396)		−2.463 (5.029)	−1.602 (5.008)	−1.082 (4.959)	−1.572 (5.410)	
Costs (EVT)	−1.320 (5.317)	−0.511 (5.599)	1.828 (6.417)	1.168 (5.012)		−1.041 (4.872)	−0.172 (5.296)	1.954 (5.890)	0.962 (4.713)	
Perf. avoid. (AG)	−3.338* (1.667)	−3.033^{+} (1.630)	−1.736 (1.489)	−0.402 (1.123)	−0.186 (1.012)	−2.467 (1.764)	−2.152 (1.751)	−1.020 (1.567)	−0.357 (1.170)	−0.135 (1.019)
Semesters (Char)	0.844 (0.915)	0.795 (0.913)	1.443 (0.904)	1.134 (0.813)		0.803 (0.870)	0.751 (0.871)	1.222 (0.850)	0.994 (0.811)	
Mast. appr. (AG)	−7.845*** (2.358)	−7.114** (2.282)	−5.035* (2.304)	−4.399^{+} (2.378)		−6.988** (2.283)	−6.213** (2.269)	−4.298^{+} (2.240)	−4.089^{+} (2.353)	
Attain. value (EVT)	3.667 (3.740)	3.961 (3.784)	4.140 (4.147)	4.953 (3.646)		4.428 (3.785)	4.733 (3.841)	5.044 (4.054)	5.752 (3.788)	

Table 14 (continued)

	Dependent variable: Points on end exam									
	Without spacing					With spacing				
	α = 1	α = 0.8	α = 0.6	α = 0.4	α = 0.2	α = 1	α = 0.8	α = 0.6	α = 0.4	α = 0.2
Conscient. (BF)	1.264 (2.699)	0.355 (2.330)	−0.828 (2.321)	0.025 (2.252)	−2.236 (1.505)	1.197 (2.515)	0.217 (2.178)	−0.642 (2.074)	0.140 (2.216)	−2.329 (1.473)
Neuroticism (BF)	3.000^{+} (1.707)	2.714 (1.684)	2.114 (1.806)	1.756 (1.671)	0.736 (1.656)	3.079^{+} (1.617)	2.769^{+} (1.585)	2.453 (1.643)	1.642 (1.662)	0.663 (1.646)
Self-selected goal 4	−5.046 (4.236)	−5.258 (4.226)	−6.694^{+} (4.008)	−7.775^{+} (4.574)		−6.315 (4.248)	−6.523 (4.249)	−7.891* (3.948)	−7.746^{+} (4.472)	
Self-selected goal 1	0.543 (0.615)	0.465 (0.601)	0.186 (0.583)	0.245 (0.459)		0.506 (0.582)	0.423 (0.574)	0.159 (0.558)	0.263 (0.451)	
Retaking statistics 2	14.829* (6.126)	14.090* (5.710)				12.461* (5.933)	11.702* (5.557)			
Present-bias (PBP)	−4.672^{+} (2.627)	−5.064^{+} (2.615)	−6.190* (2.478)	−6.386** (2.467)		−5.183* (2.451)	−5.597* (2.484)	−6.723** (2.365)	−6.817** (2.461)	
Openness (BF)	3.942* (1.735)	3.790* (1.774)	3.621^{+} (1.911)			5.151** (1.680)	4.968** (1.710)	5.104** (1.715)		
Perfor. appr. (AG)	2.648 (2.295)	2.396 (2.258)	0.656 (2.082)			0.966 (2.326)	0.721 (2.311)	−1.004 (2.132)		
Intrinsic value (EVT)	−2.533 (4.445)	−2.138 (4.317)	−0.111 (4.439)			−3.155 (3.931)	−2.719 (3.759)	−1.300 (3.787)		
Discount factor (PBP)	−7.190 (6.086)	−5.754 (5.844)		−0.972 (6.475)		−4.035 (6.383)	−2.537 (6.157)		0.612 (6.468)	
Self-selected goal 3	2.226 (3.397)	1.924 (3.579)	0.559 (4.117)			2.218 (3.179)	1.892 (3.389)	0.901 (3.831)		
Self-selected goal 2	12.303 (10.769)	14.685 (11.152)	12.033 (9.572)			11.768 (10.998)	14.346 (11.470)	12.198 (9.582)		

Table 14 (continued)

	Dependent variable: Points on end exam									
	Without spacing					With spacing				
	α = 1	α = 0.8	α = 0.6	α = 0.4	α = 0.2	α = 1	α = 0.8	α = 0.6	α = 0.4	α = 0.2
Agreeableness (BF)	− 1.818 (2.589)					− 1.962 (2.450)				
Intercept	20.407 (36.239)	7.754 (34.383)	7.609 (39.613)	28.519 (33.833)	17.915 (16.049)	18.668 (37.185)	5.053 (36.454)	5.227 (40.215)	26.532 (33.675)	20.600 (16.095)
N	67	67	67	67	67	67	67	67	67	67
R^2	0.749	0.746	0.716	0.687	0.618	0.765	0.762	0.742	0.699	0.630
Adj. R^2	0.606	0.610	0.583	0.578	0.566	0.622	0.625	0.612	0.586	0.571

This tables shows the full regression estimates from Table 4X. Heteroskedastic robust standard errors in parenthesis. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Table 15 Listwise deletion depending on selected control variables without spacing

	Dependent variable: Points on the end exam				
	Points				
	Lasso	EN (0.8)	EN (0.6)	EN (0.4)	EN (0.2)
	(1)	(2)	(3)	(4)	(5)
Retrieval practice attempts	0.891*	0.855*	1.229***	1.229***	1.695***
	(0.486)	(0.459)	(0.354)	(0.354)	(0.421)
Mean FTF tutorial preparation	6.618***	5.913**	3.001	3.001	6.868***
	(2.460)	(2.491)	(1.956)	(1.956)	(2.213)
Standardised points in statistics 1	11.438***	11.159***	13.635***	13.635***	
	(3.907)	(3.750)	(2.642)	(2.642)	
Self-concept (EVT)	5.946**	4.899			
	(2.958)	(3.403)			
Utility value (EVT)	5.488	5.118			
	(4.111)	(4.206)			
Present-bias (PBP)	−3.905**				
	(1.733)				
Mastery approach (AG)	−3.494				
	(2.257)				
Female (Char)	−0.212	−0.509			
	(4.479)	(3.795)			
Semesters (Char)	0.646	0.855			
	(0.723)	(0.755)			
Retaking statistics 2	7.480	5.735			
	(6.090)	(5.512)			
Cost (EVT)	2.025	−0.727			
	(4.012)	(4.527)			
Constant	18.427	5.057	38.513***	38.513***	32.648***
	(19.983)	(18.914)	(3.737)	(3.737)	(4.204)
Observations	48	50	67	67	67
R^2	0.621	0.563	0.547	0.547	0.306

Table 15 (continued)

	Dependent variable: Points on the end exam				
	Points				
	Lasso	EN (0.8)	EN (0.6)	EN (0.4)	EN (0.2)
	(1)	(2)	(3)	(4)	(5)
Adjusted R^2	0.505	0.465	0.526	0.526	0.285

The table shows regressions results without multiple imputation, including the maximum number of individuals depending on the selected variables. Heteroskedastic robust standard errors in parenthesis. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Table 16 Listwise deletion depending on selected control variables with spacing

	Dependent variable: Point on end exam				
	Points				
	Lasso	EN (0.8)	EN (0.6)	EN (0.4)	EN (0.2)
	(1)	(2)	(3)	(4)	(5)
Retrieval practice	0.326	0.802	0.821	0.780**	1.221**
	(0.498)	(0.522)	(0.541)	(0.372)	(0.498)
Mean FTF tutorial preparation	4.263*	4.783*	6.533***	2.262	6.071***
	(2.567)	(2.559)	(2.520)	(1.975)	(2.290)
Spaced learning	0.449*	0.453**	0.315	0.506**	0.532**
	(0.254)	(0.230)	(0.235)	(0.200)	(0.263)
Standardised points in statistics 1	11.416***	15.579***	13.838***	13.568***	
	(4.267)	(3.648)	(3.376)	(2.645)	
Self-concept (EVT)	6.908**		6.230**		
	(2.998)		(3.091)		
Utility value (EVT)	8.387*				
	(4.415)				
Mastery approach (AG)	−1.182				
	(2.132)				
Performance avoidance (AG)	0.125				
	(1.117)				
Conscientiousness (BF)	−0.954				
	(1.788)				
Neuroticism (BF)	2.065				
	(1.587)				
Female (Char)		−4.314	−2.570		
		(4.031)	(4.104)		
Attainment value (EVT)	6.195	0.130			
	(4.012)	(2.844)			

Table 16 (continued)

	Dependent variable: Point on end exam				
	Points				
	Lasso	EN (0.8)	EN (0.6)	EN (0.4)	EN (0.2)
	(1)	(2)	(3)	(4)	(5)
Semesters (Char)			1.207		
			(0.841)		
Cost (EVT)			2.152		
			(4.299)		
Constant	− 15.443	36.519***	8.701	39.331***	33.538***
	(28.772)	(8.509)	(18.091)	(3.695)	(4.224)
Observations	47	54	54	67	67
Adjusted R^2	0.494	0.556	0.583	0.543	0.300

The table shows regressions results without multiple imputation, including the maximum number of individuals depending on the selected variables. Heteroskedastic robust standard errors in parenthesis. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the LEAD Graduate School and Research Network [GSC1028] and *From Prediction to Agile Interventions in the Social Sciences (FAIR) [PROFILNRW-2020-068z]. The project From Prediction to Agile Interventions in the Social Sciences (FAIR)* is receiving funding from the programme "Profilbildung 2020", an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia. The sole responsibility for the content of this publication lies with the authors.

Data availability The data analyzed in the current study are not publicly available due to privacy and confidentiality restrictions. The data are available from the corresponding author upon reasonable request.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial 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.

Conflict of interest The authors have no conflicts of interest to declare relevant to this article's content.

Ethical Approval Approval was granted by the Ethics Committee of the University of Tübingen.

References

Akande, O., Li, F., & Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *American Statistician, 71*(2), 162–170. https://doi.org/10.1080/00031305.2016.1277158

Alexander, P. A., Dinsmore, D. L., Parkinson, M. M., & Winters, F. I. (2011). Self-regulated learning in academic domains. In B. Z. & D. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 393–407). Routledge.

Anthony, B., Kamaludin, A., Romli, A., Raffei, A. F. M., Phon, D. N. A. L. E., Abdullah, A., & Ming, G. L. (2020). Blended learning adoption and implementation in higher education: A theoretical and systematic review. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-020-09477-z

Attali, Y. (2015). Effects of multiple-try feedback and question type during mathematics problem solving on performance in similar problems. *Computers and Education, 86*, 260–267. https://doi.org/10.1016/j.compedu.2015.08.011

Attali, Y., & van der Kleij, F. (2017). Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers and Education, 110*, 154–169. https://doi.org/10.1016/j.compedu.2017.03.012

Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning, 4*(1), 87–95. https://doi.org/10.1007/s11409-009-9035-7

Bailey, T. H., & Phillips, L. J. (2016). The influence of motivation and adaptation on students' subjective well-being, meaning in life and academic performance. *Higher Education Research and Development, 35*(2), 201–216.

Baker, R., Evans, B., Li, Q., & Cung, B. (2019). Does inducing students to schedule lecture watching in online classes improve their academic performance? An experimental analysis of a time management intervention. *Research in Higher Education, 60*(4), 521–552. https://doi.org/10.1007/S11162-018-9521-3

Baker, R., Xu, D., Park, J., Yu, R., Li, Q., Cung, B., Fischer, C., Rodriguez, F., Warschauer, M., & Smyth, P. (2020). The benefits and caveats of using clickstream data to understand student self-regulatory

behaviors: Opening the black box of learning processes. *International Journal of Educational Technology in Higher Education, 17*(1), 1–24. https://doi.org/10.1186/s41239-020-00187-1

Baranik, L. E., Stanley, L. J., Bynum, B. H., & Lance, C. E. (2010). Examining the construct validity of mastery-avoidance achievement goals: A meta-analysis. *Human Performance, 23*(3), 265–282. https://doi.org/10.1080/08959285.2010.488463

Becker, A., Deckers, T., Dohmen, T., Falk, A., & Kosse, F. (2012). The relationship between economic preferences and psychological personality measures. *Annual Review of Economics, 4*(1), 453–478. https://doi.org/10.1146/annurev-economics-080511-110922

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies, 81*(2), 608–650. https://doi.org/10.1093/restud/rdt044

Benden, D. K., & Lauermann, F. (2022). Students' motivational trajectories and academic success in math-intensive study programs: Why short-term motivational assessments matter. *Journal of Educational Psychology, 114*(5), 1062–1085. https://doi.org/10.1037/edu0000708

Bisin, A., & Hyndman, K. (2020). Present-bias, procrastination and deadlines in a field experiment. *Games and Economic Behavior, 119*, 339–357. https://doi.org/10.1016/j.geb.2019.11.010

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Brenner, P. S., & DeLamater, J. (2017). Lies, damned lies, and survey self-reports? identity as a cause of measurement bias. *Social Psychology Quarterly*, *176*(5), 139–148. https://doi.org/10.1177/0190272516628298.

Brisson, B. M., Dicke, A.-L., Gaspard, H., Häfner, I., Flunger, B., Nagengast, B., & Trautwein, U. (2017). Short intervention, sustained effects: Promoting students' math competence beliefs, effort, and achievement. *American Educational Research Journal, 54*(6), 1048–1078. https://doi.org/10.3102/0002831217716084

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology, 36*(5), 1118–1133. https://doi.org/10.1037/a0019902

Butler, A. C., Black-Maier, A. C., Raley, N. D., & Marsh, E. J. (2017). Retrieving and applying knowledge to different examples promotes transfer of learning. *Journal of Experimental Psychology: Applied, 23*(4), 433–446. https://doi.org/10.1037/xap0000142

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bulletin and Review, 18*(6), 1238–1244. https://doi.org/10.3758/s13423-011-0173-y

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory and Cognition, 36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245. https://doi.org/10.2307/1170684

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279–283. https://doi.org/10.1177/0963721412452728

Carpenter, S. K. (2014). Improving student learning in low-maintenance and cost-effective ways. *Journal of Applied Research in Memory and Cognition, 3*(3), 121–123. https://doi.org/10.1016/j.jarmac.2014.07.004

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review, 28*(2), 353–375. https://doi.org/10.1007/s10648-015-9311-9

Carvalho, P. F., McLaughlin, E. A., & Koedinger, K. R. (2022). Varied practice testing is associated with better learning outcomes in self-regulated online learning. *Journal of Educational Psychology, 114*(8), 1723–1742. https://doi.org/10.1037/edu0000754

Castro, M. D. B., & Tumibay, G. M. (2021). A literature review: Efficacy of online learning courses for higher education institution using meta-analysis. *Education and Information Technologies, 26*(2), 1367–1385. https://doi.org/10.1007/s10639-019-10027-z

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553

Collins, D. P., Rasco, D., & Benassi, V. A. (2018). Test-enhanced learning: Does deeper processing on quizzes benefit exam performance? *Teaching of Psychology, 45*(3), 235–238. https://doi.org/10.1177/0098628318779262

Condron, D. J., Becker, J. H., & Bzhetaj, L. (2018). Sources of students' anxiety in a multidisciplinary social statistics course. *Teaching Sociology, 46*(4), 346–355. https://doi.org/10.1177/0092055X18780501

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: reconsidered once again. *Review of Educational Research, 71*(1), 1–27. https://doi.org/10.3102/00346543071001001

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review, 1*(4), 309–330. https://doi.org/10.1007/BF01320097

Denny, P., McDonald, F., Empson, R., Kelly, P., & Petersen, A. (2018). Empirical support for a causal relationship between gamification and learning outcomes. *CHI, 311*, 1–13. https://doi.org/10.1007/978-3-030-37386-3_30

Digman, J. M. (1990). Personality structure: Emergence if the five-factor model. *Annual Review of Psychology, 41*, 417–440.

Dobson, J., Linderholm, T., & Perez, J. (2018). Retrieval practice enhances the ability to evaluate complex physiology information. *Medical Education, 52*(5), 513–525. https://doi.org/10.1111/medu.13503

Donoghue, G. M., & Hattie, J. A. C. (2021). A meta-analysis of ten learning techniques. *Frontiers in Education, 6*(March), 1–9. https://doi.org/10.3389/feduc.2021.581216

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Dunn, T. J., & Kennedy, M. (2019). Technology enhanced learning in higher education; motivations, engagement and academic achievement. *Computers and Education, 137*(March), 104–113. https://doi.org/10.1016/j.compedu.2019.04.004

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). Freeman.

Elliot, A. J., & McGregor, H. (2001). A 2×2 achievement goal framework. *Journal of Personality and Social Psychology, 80*, 501–519. https://doi.org/10.1037//0022-3514.80.3.501

Elliot, A. J., McGregor, H. A., & Gable, S. (1999). Achievement goals, study strategies, and exam performance: A mediational analysis. *Journal of Educational Psychology, 91*(3), 549–563. https://doi.org/10.1037/0022-0663.91.3.549

Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology, 100*(3), 613–628. https://doi.org/10.1037/0022-0663.100.3.613

Förster, M., Maur, A., Weiser, C., & Winkel, K. (2022). Pre-class video watching fosters achievement and knowledge retention in a flipped classroom. *Computers & Education, 179*, 104399. https://doi.org/10.1016/j.compedu.2021.104399

Förster, M., Weiser, C., & Maur, A. (2018). How feedback provided by voluntary electronic quizzes affects learning outcomes of university students in large classes. *Computers and Education, 121*, 100–114. https://doi.org/10.1016/j.compedu.2018.02.012

Frederick, S., & Loewenstein, G. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature, 40*, 351–401. https://doi.org/10.1257/002205102320161311

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer Series in Statistics.

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review, 75*(3), 372–396. https://doi.org/10.1111/j.1751-5823.2007.00029.x

Gaspard, H., Häfner, I., Parrisius, C., Trautwein, U., & Nagengast, B. (2017). Assessing task values in five subjects during secondary school: Measurement structure and mean level differences across grade level, gender, and academic subject. *Contemporary Educational Psychology, 48*, 67–84. https://doi.org/10.1016/j.cedpsych.2016.09.003

Graham, C. R., Woodfield, W., & Harrison, J. B. (2013). A framework for institutional adoption and implementation of blended learning in higher education. *Internet and Higher Education, 18*, 4–14. https://doi.org/10.1016/j.iheduc.2012.09.003

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance

from freshman year through graduation. *Journal of Educational Psychology, 94*(3), 562–575. https://doi.org/10.1037/0022-0663.94.3.562

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin and Review, 19*(1), 126–134. https://doi.org/10.3758/s13423-011-0181-y

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Linear methods for regression. In T. Hastie, R. Tibshirani, & J. Friedman (Eds.), *The elements of statistical learning* (pp. 43–94). Springer. https://doi.org/10.1007/b94608

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning Memory and Cognition, 39*(1), 290–296. https://doi.org/10.1037/a0028468

Hellings, J., & Haelermans, C. (2020). The effect of providing learning analytics on student behaviour and performance in programming: A randomised controlled experiment. *Higher Education*. https://doi.org/10.1007/s10734-020-00560-z

Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology, 102*(4), 880–895. https://doi.org/10.1037/a0019506

Ifenthaler, D., Schumacher, C., & Kuzilek, J. (2023). Investigating students' use of self-assessments in higher education using learning analytics. *Journal of Computer Assisted Learning, 39*(1), 255–268. https://doi.org/10.1111/jcal.12744

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test… or testing to teach: exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26*(2), 307–329. https://doi.org/10.1007/s10648-013-9248-9

Jonides, J. (2004). How does practice makes perfect? *Nature Neuroscience, 7*(1), 10–11. https://doi.org/10.1038/nn0104-10

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*(4), 469–486. https://doi.org/10.1037/a0017341

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. T. Wixted (Ed.), *Cognitive psychology of memory, of learning and memory: A comprehensive reference* (2nd ed., pp. 487–514). Academic Press. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 139*, 772–774. https://doi.org/10.1126/science.1199327

Karpicke, J. D., Blunt, J. R., Smith, M. A., & Karpicke, S. S. (2014). Retrieval-based learning: The need for guided retrieval in elementary school children. *Journal of Applied Research in Memory and Cognition, 3*(3), 198–206. https://doi.org/10.1016/j.jarmac.2014.07.008

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*(1), 17–29. https://doi.org/10.1016/j.jml.2012.02.004

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102

Komarraju, M., Karau, S. J., & Schmeck, R. R. (2009). Role of the big five personality traits in predicting college students' academic motivation and achievement. *Learning and Individual Differences, 19*(1), 47–52. https://doi.org/10.1016/j.lindif.2008.07.001

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning Memory and Cognition, 35*(4), 989–998. https://doi.org/10.1037/a0015729

Krause, A., Rinne, U., & Zimmermann, K. F. (2012). Anonymous job applications of fresh Ph.D. economists. *Economics Letters, 117*, 441–444. https://doi.org/10.1016/j.econlet.2012.06.029

Landrum, R. E. (2007). Introductory psychology student performance: Weekly quizzes followed by a cumulative final exam. *Teaching of Psychology, 34*(3), 177–180. https://doi.org/10.1080/00986280701498566

Lechuga, M. T., Ortega-Tudela, J. M., & Gómez-Ariza, C. J. (2015). Further evidence that concept mapping is not better than repeated retrieval as a tool for learning from texts. *Learning and Instruction, 40*, 61–68. https://doi.org/10.1016/j.learninstruc.2015.08.002

Lim, S. W. H., Ng, G. J. P., & Wong, G. Q. H. (2015). Learning psychological research and statistical concepts using retrieval-based practice. *Frontiers in Psychology, 6*, 1484. https://doi.org/10.3389/fpsyg.2015.01484

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods, 22*(1), 141–165. https://doi.org/10.1037/met0000096

Macher, D., Papousek, I., Ruggeri, K., & Paechter, M. (2015). Statistics anxiety and performance: Blessings in disguise. *Frontiers in Psychology, 6*, 1116.

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology, 110*, 63–73. https://doi.org/10.1016/j.jclinepi.2019.02.016

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology, 81*(1), 59–77. https://doi.org/10.1348/000709910X503501

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable: Research article. *Psychological Science, 20*(4), 516–522. https://doi.org/10.1111/j.1467-9280.2009.02325.x

McKenzie, K., & Schweitzer, R. (2001). Who succeeds at university? Factors predicting academic performance in first year Australian university students. *Higher Education Research & Development, 20*(1), 21–33.

Morrison, G. R., & Anglin, G. J. (2005). Research on cognitive load theory: Application to e-learning. *Educational Technology Research and Development, 53*(3), 94–104. https://doi.org/10.1007/BF02504801

Mundt, D., Abel, R., Hänze, M., Abel, R., & Hänze, M. (2020). Exploring the effect of testing on forgetting in vocabulary learning: An examination of the bifurcation model examination of the bifurcation model. *Journal of Cognitive Psychology, 32*(2), 214–228. https://doi.org/10.1080/20445911.2020.1733584

Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science, 33*(2), 142–159. https://doi.org/10.1214/18-STS644

Murre, J. M. J., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLoS ONE, 10*(7), 1–23. https://doi.org/10.1371/journal.pone.0120644

O'Brien, M., & Verma, R. (2019). How do first year students utilize different lecture resources? *Higher Education, 77*(1), 155–172. https://doi.org/10.1007/s10734-018-0250-5

Paechter, M., Maier, B., & Macher, D. (2010). Students' expectations of, and experiences in e-learning: Their relation to learning achievements and course satisfaction. *Computers and Education, 54*(1), 222–229.

Panus, P. C., Stewart, D. W., Hagemeier, N. E., Thigpen, J. C., & Brooks, L. (2014). A subgroup analysis of the impact of self-testing frequency on examination scores in a pathophysiology course. *American Journal of Pharmaceutical Education, 78*(9), 165. https://doi.org/10.5688/ajpe789165

Park, J., Yu, R., Rodriguez, F., Baker, R., Smyth, P., & Warschauer, M. (2018). Understanding student procrastination via mixture models. In *Proceedings of the 11th international conference on educational data mining (EDM)* (pp. 187–197).

Payne, S. C., Youngcourt, S. S., & Beaubien, J. M. (2007). A meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology, 92*(1), 128–150. https://doi.org/10.1037/0021-9010.92.1.128

Plante, I., O'Keefe, P. A., & Théorêt, M. (2013). The relation between achievement goal and expectancy-value theories in predicting achievement-related outcomes: A test of four theoretical conceptions. *Motivation and Emotion, 37*(1), 65–78. https://doi.org/10.1007/s11031-012-9282-9

Racsmány, M., Szőllősi, Á., & Marián, M. (2020). Reversing the testing effect by feedback is a matter of performance criterion at practice. *Memory and Cognition, 48*(7), 1161–1170. https://doi.org/10.3758/s13421-020-01041-5

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory and Cognition, 43*(4), 619–633. https://doi.org/10.3758/s13421-014-0477-z

Reeves, T. C., & Lin, L. (2020). The research we have is not the research we need. *Educational Technology Research and Development, 68*(4), 1991–2001. https://doi.org/10.1007/s11423-020-09811-3

Rimfeld, K., Kovas, Y., Dale, P. S., & Plomin, R. (2016). True grit and genetics: Predicting academic achievement from personality. *Journal of Personality and Social Psychology, 111*(5), 780–789.

Rodriguez, F., Fischer, C., Zhou, N., Warschauer, M., & Massimelli, J. (2021a). Student spacing and self-testing strategies and their associations with learning in an upper division microbiology course. *SN Social Sciences, 1*(38), 1–21. https://doi.org/10.1007/s43545-020-00013-5

Rodriguez, F., Kataoka, S., Janet Rivas, M., Kadandale, P., Nili, A., & Warschauer, M. (2021b). Do spacing and self-testing predict learning outcomes? *Active Learning in Higher Education, 22*(1), 77–91. https://doi.org/10.1177/1469787418774185

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382–395. https://doi.org/10.1037/a0026252

Roediger, H. L., III., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Schupp, J., & Gerlitz, J. (2014). Big five inventory-SOEP (BFI-S). *Zusammenstellung Sozialwissenschaftlicher Items Und Skalen (ZIS)*.

Schwerter, J., Dimpfl, T., Bleher, J., & Murayama, K. (2022). Benefits of additional online practice opportunities in higher education. *Internet and Higher Education, 53*, 100834. https://doi.org/10.1016/j.iheduc.2021.100834

Schwerter, J., Wortha, F., & Gerjets, P. (2022). E-learning with multiple-try-feedback: Can hints foster students' achievement during the semester? *Educational Technology Research and Development, 70*, 713–736. https://doi.org/10.1007/s11423-022-10105-z

Sorić, I., Penezić, Z., & Burić, I. (2017). The big five personality traits, goal orientations, and academic achievement. *Learning and Individual Differences, 54*, 126–134. https://doi.org/10.1016/j.lindif.2017.01.024

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Su, N., Buchin, Z. L., & Mulligan, N. W. (2020). Levels of retrieval and the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(4), 652–670. https://doi.org/10.1037/xlm0000962

Susser, J. A., & McCabe, J. (2013). From the lab to the dorm room: Metacognitive awareness and use of spaced study. *Instructional Science, 41*(2), 345–363. https://doi.org/10.1007/s11251-012-9231-8

Tempel, T., Kaufmann, K., Kranz, J., & Möller, A. (2020). Retrieval-based skill learning: Testing promotes the acquisition of scientific experimentation skills. *Psychological Research Psychologische Forschung, 84*(3), 660–666. https://doi.org/10.1007/s00426-018-1088-2

Tullis, J. G., & Maddox, G. B. (2020). Self-reported use of retrieval practice varies across age and domain. *Metacognition and Learning, 15*(2), 129–154. https://doi.org/10.1007/s11409-020-09223-x

Vaessen, B. E., van den Beemt, A., van de Watering, G., van Meeuwen, L. W., Lemmens, L., & den Brok, P. (2017). Students' perception of frequent assessments and its relation to motivation and grades in a statistics course: A pilot study. *Assessment and Evaluation in Higher Education, 42*(6), 872–886. https://doi.org/10.1080/02602938.2016.1204532

van der Velde, R., Blignaut–van Westrhenen, N., Labrie, N. H. M., & Zweekhorst, M. B. M. (2021). 'The idea is nice… but not for me': First-year students' readiness for large-scale 'flipped lectures'—what (de)motivates them? *Higher Education, 81*(6), 1157–1175. https://doi.org/10.1007/s10734-020-00604-4

van Lent, M., & Souverijn, M. (2020). Goal setting and raising the bar: A field experiment. *Journal of Behavioral and Experimental Economics, 87*(May), 101570. https://doi.org/10.1016/j.socec.2020.101570

Van Yperen, N. W., Blaga, M., & Postmes, T. (2014). A meta-analysis of self-reported achievement goals and nonself-report performance across three achievement domains (work, sports, and education). *PLoS ONE, 9*(4), e93594. https://doi.org/10.1371/journal.pone.0093594

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68–81. https://doi.org/10.1006/ceps.1999.1015

Wigfield, A., & Eccles, J. S. (2020). 35 years of research on students' subjective task values and motivation: A look back and a look forward. In A. J. Elliot (Ed.), *Advances in motivation science* (Vol. 7, pp. 161–198). Elsevier Inc. https://doi.org/10.1016/bs.adms.2019.05.002

Wilson, M. L., & Zietz, J. (2004). Systematic Bias in Student Self-Reported Data. *Journal for Economic Educators*, *4*(4), 13–19.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*(3087), 1–14. https://doi.org/10.3389/fpsyg.2019.03087

Wong, S. S. H., & Lim, S. W. H. (2022). Deliberate errors promote meaningful learning. *Journal of Educational Psychology, 114*(8), 1817–1831. https://doi.org/10.1037/edu0000720

Wong, S. S. H., Ng, G. J. P., Tempel, T., & Lim, S. W. H. (2019). Retrieval practice enhances analogical problem solving. *Journal of Experimental Education, 87*(1), 128–138. https://doi.org/10.1080/00220973.2017.1409185

Yan, V. X., Thai, K. P., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: Do they vary with mindset. *Journal of Applied Research in Memory and Cognition, 3*(3), 140–152. https://doi.org/10.1016/j.jarmac.2014.04.003

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A Systematic and meta-analytic review. *Psychological Bulletin, 147*(4), 399. https://doi.org/10.1037/bul0000309

Ziegler, M., Schroeter, T. A., Lüdtke, O., & Roemer, L. (2018). The enriching interplay between openness and interest: A theoretical elaboration of the OFCI model and a first empirical test. *Journal of Intelligence, 6*(3), 1–22. https://doi.org/10.3390/jintelligence6030035