Educational technology research and development (2022) 70:713–736
https://doi.org/10.1007/s11423-022-10105-z

RESEARCH ARTICLE

E-learning with multiple-try-feedback: Can hints foster students' achievement during the semester?

Jakob Schwerter[1] · Franz Wortha[2] · Peter Gerjets[2]

Accepted: 9 March 2022 / Published online: 17 May 2022

Abstract

E-learning opportunities have become an increasingly important component of university education. Various laboratory studies have shown that e-learning environments can meaningfully enhance learning by incorporating various interventions and design choices (e.g., providing feedback and scaffolds). However, many computer-based interventions have not yet been applied in authentic university courses, raising questions about whether and how the provision of certain forms of feedback works and scales in an applied context. In this paper, we addressed this research gap. Specifically, we investigated whether including an elaborative component (hints) in multiple-try feedback increases student learning in e-learning exercises in an undergraduate statistics course. In one exercise, after completing a statistical problem, one group received feedback that conveyed knowledge about the correct response, while the other group additionally received elaborative feedback in the form of hints. We conducted an experimental comparison of these two types of feedback with third-semester sociology students in the tutorial component of an introductory statistics course. The results show that additional feedback helps students perform better during the session and on a delayed test one week later. Implications for further research and the application of such e-learning environments in university settings are discussed.

Keywords Computer-based learning environment · Within-person randomization · Learning outcome · Delayed testing · Panel regression · Statistics in higher education

Author NoteJakob Schwerter and Franz Wortha were doctoral students at the LEAD Research Network [GSC1028], funded by the Excellence Initiative of the German federal and state governments.

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

[1] Center for Research on Education and School Development, LEAD Graduate School & Research Network, Technical University Dortmund, Campus Nord (CDI Gebäude), Vogelpothsweg 78, 44227 Dortmund, Germany

[2] Tübingen and LEAD Graduate School & Research Network, Leibniz-Institut für Wissensmedien, Tübingen, Deutschland

Introduction

E-learning environments have become essential components of modern educational settings (Bennett, 2015). Over the past decade, teaching has increasingly shifted from traditional, non-digital media (e.g., face-to-face lectures and seminars) to computer-based environments and instruction, also known as e-learning environments (Clark & Mayer, 2016; Uzunboylu, 2006). The increased prominence of e-learning instruction is driven by the unique opportunities afforded by digital technology, which lead to more efficient and effective student learning (Morrison & Anglin, 2005). One such element is the ability to provide immediate and specific feedback to multiple learners simultaneously, which exceeds the capabilities of human instructors. Thus, e-learning environments can enhance the provision of feedback as a pedagogical intervention in practical contexts, such as school or university courses.

A wealth of research has shown that feedback in digital and face-to-face interventions is one of the most effective pedagogical interventions. Recently, Wisniewski et al., (2020) summarized the results of 30 meta-analyses on the effectiveness of feedback with different foci. These reviews and meta-analyses showed that feedback interventions can have a significant positive effect on learning and learning outcomes when implemented appropriately. Four important conditions for the effectiveness and usefulness of feedback were identified by Shute (2008): (i) the learner is in a situation where they need feedback, (ii) they receive the feedback in a timely manner, and (iii) they are able and willing to use the feedback. Feedback must be appropriate for the task and the learner's disposition, and the learner must be committed to using it. In addition, according to Hattie & Gan (2011), (iv) effective feedback is related to the task itself rather than directed at the learner (Azevedo & Bernard, 1995; Kluger & DeNisi, 1996; Proske et al., 2012). Otherwise, feedback can negatively influence the learning process (Hattie & Gan, 2011; Shute, 2008). In addition, a study by Cutumisu & Schwartz (2018) showed that students engage more with feedback when they can choose whether or not to receive feedback.

Feedback plays a significant role in computer-based learning environments because it can be personalized and provided immediately in these environments (Clariana et al., 1991; Wang & Lehman, 2021). However, providing these levels of personalized and immediate feedback is infeasible for teaching assistance if the number of students is too high. Thereby, the lack of feedback in such situations might lead to disinterest or false understanding (Morrison & Anglin, 2005). Consequently, feedback in computer-based learning environments can satisfy points (i), (ii), and (iv) of the above requirements: it can be given immediately after a student encounters a problem or makes a mistake, and if implemented correctly it relates to the task itself. Accordingly, an extensive literature addresses the design and effectiveness of feedback in computer-based learning environments. A recent review has shown that the findings are consistent with previous research on feedback in other educational environments (Van der Kleij et al., 2015). Specifically, feedback interventions exhibit, on average, a significant positive effect on learning and learning outcomes, with wide variation in effect sizes. In particular, the ability to provide immediate feedback after a student has provided an incorrect answer is valuable for low-ability learners and on complex tasks (Bangert-Drowns et al., 1991).

Three general types of feedback are examined in the literature: elaborative feedback (EF), knowledge of response (KR), and knowledge of the correct response (KCR) (Clariana et al., 1991). In a computer-based environment with item-based feedback, KR only tells the learner whether their answer is correct or incorrect, while KCR also provides the correct solution when the learner provided an incorrect answer. EF might include providing a hint or a similar example to help the person solve the task. In comparisons of the three types of feedback, EF has been found to be more effective than KR and KCR (Clariana et al., 1991; Kulhavy, 1977; Van der Kleij et al., 2015) showed that combining these feedback types is more effective than any specific type alone, especially with respect to higher-order learning outcomes, such as applying one's knowledge and skills in a new environment. The literature notes that KC and KCR have only a corrective function (Hattie & Timperley, 2007), while EF is able to go beyond that and actually increase learning (Attali & van der Kleij, 2017). According to Wisniewski et al., (2020), EF (or high-information feedback) contains additional information about the task, process, and or level of self-regulation. It is generally more effective for higher-order learning tasks than less elaborate forms of feedback.

In addition to the type of feedback, timing is also essential to consider when designing feedback. We follow van der Kleij et al., (2012) in defining immediate feedback as given immediately after a specific item (exercise, question, or problem) has been answered, while delayed feedback is given after all items have been answered. Although there is some disagreement in the literature, immediate feedback appears to be more effective than delayed feedback (Van der Kleij et al., 2015). Whether feedback should be immediate or delayed seems to depend on individual ability and the task at hand (Butler & Winne, 1995; Mathan & Koedinger, 2002): for higher-order outcomes and low-ability learners, feedback should be given immediately.

One form of feedback that is particularly useful for promoting higher-order learning is multiple-try feedback. Multiple-try feedback describes a feedback process in which learners are informed that their answer is incorrect (with varying degrees of elaboration). Learners then have the opportunity to correct their errors. After learners answer the question correctly or reach the maximum number of attempts, the correct response is communicated. A review of multiple-try feedback (Clariana & Koul, 2005) found that multiple-try feedback was more effective than other forms of feedback for higher-order learning outcomes but was inferior for lower-order learning outcomes such as memorization. The authors argued that the generative effect of this feedback is exceptionally high for tasks that require students to develop a deeper understanding (rather than learning facts). In a multiple-try setting, students are explicitly allowed to make mistakes and learn from them. Being in a situation in which students are allowed to make mistakes and learn from them by allowing for additional responses promotes this so-called *error generation effect* (Kornell et al., 2009).

Studies have further indicated that positive feedback effects are particularly pronounced in STEM-related fields (e.g., mathematics Attali & van der Kleij 2017). Attali (2015) investigated the use of different types of feedback when solving mathematical problems in multiple-choice and open-ended formats: (i) multiple-try feedback with and without additional hints, (ii) knowledge of the correct response, and (iii) no feedback. Hints are intermediate steps, formulas, or additional explanations to help participants find their error independently. In this experiment, participants had to complete 15 items and received different forms of feedback depending on the experimental condition. The control group worked on similar items but received no feedback. The results showed that multiple-try feedback led to higher learning gains than learning the correct answer without being able to take multiple tries and receiving feedback after each attempt. In addition, feedback with multiple trials and hints was more effective in promoting learning than feedback without hints. Finally, learning gains were higher with open-ended questions than with multiple-choice questions. Overall, this study showed that multiple-trial feedback in multiple-choice and open-ended questions was effective for higher-order STEM-related learning outcomes, especially when the feedback incorporated additional hints. The authors explained these results with the argument that multiple-try feedback elicits more attentive and effortful problem solving.

While the research outlined above demonstrates the promise of multiple-try feedback for higher-order learning, these findings have yet to be transferred to educational practice. To our knowledge, no studies have directly examined the effectiveness of such feedback in higher education courses. Multiple-try feedback in computer-based learning environments is directly applicable to key areas of higher education, such as mathematics or statistics education. Solving mathematical and statistical problems is a core element of many fields of study and requires students to work through complex chains of calculations and revise their answers as necessary. Social science majors, for example, are less mathematical but generally still require students to take several statistics courses. Statistics in higher education is a topic that elicits anxiety even among graduates (Valle et al., 2021), which is why improving foundational statistics courses is of high relevance. Statistical education requires students to acquire sufficient knowledge and skills and apply them to different tasks. This in turn necessitates appropriately designed instruction including learning tasks with varying complexity, different kinds of auxiliary information and guidance, as well as opportunities to practice. Specifically, according to the four components of instructional design (4 C/ID) model, a well-designed educational program builds on four components: the learning task, part-task practice, supportive, and procedural information (Merrill, 2002; van Merriënboer & Kirschner, 2018). In this context, multiple-try feedback can be used to provide just-in-time information throughout different stages of the learning task (e.g., during practice) that enables learners to more efficiently and effectively acquire sophisticated schemas and mental models (Frerejean et al., 2019).

Therefore, in the present study, we aim to fill this research gap by investigating the additive value of hints in multiple-try feedback during weekly e-learning tutorials in an undergraduate statistics course for social science majors. Do university students really need extra hints, or do they compensate in other ways when no hints are provided? In our view, the results of laboratory studies might not necessarily translate to applied situations for the following reasons (among others): in the real world, students might learn in groups, help each other, read additional textbooks or watch instructional videos online, or obtain additional topic-relevant input from other sources. Students also experience more sources of disruption over the course of an entire semester. For example, they have to balance multiple courses and their personal lives and may even have to work for pay. Although multiple-try feedback with additional hints is effective in a laboratory setting, before it can be recommended in practice, its ability to help students in the regular university setting must be demonstrated, confirming the external validity of the laboratory results in more realistic settings (Morrison & Anglin, 2005; Ross & Morrison, 1989).

To this end, we aimed to test the following hypothesis derived from the literature: does multiple-try feedback with hints outperform multiple-try feedback without hints over the course of an entire semester at university? To measure the effect of additional hints, we used immediate and delayed test settings in twelve e-learning sessions and the end-of-semester exam. Thus, we distinguish between multiple-try feedback with knowledge of the correct response (MCR) and multiple-try feedback with hints (i.e., elaborative feedback) after a first incorrect answer (MCR+H). Students received alternating MCR or MCR+H in twelve e-learning sessions, one per week. We did not include a no-feedback group for ethical reasons. Following Attali (2015), we expected significantly higher learning gains in the MCR+H condition in the immediate test and in a delayed, one-week posttest. We observed 87 students with varying levels of participation in the weekly e-learning sessions. The statistics course is the second of two mandatory statistics courses for a bachelor's degree program in sociology at a large German university and is typically taken in the third semester of studies.

In a second step, we analyzed the impact of the treatment groups and individual performance in the e-learning sessions on the students' exam grades in order to examine the promise of multiple-try feedback in practice. Since the treatments alternated weekly, no general treatment group effect on exam grades was expected. However, we expected a positive correlation with performance in the e-learning sessions.

To date, research in higher education has focused on digital (or virtual) learning formats, face-to-face formats, and blended learning formats, a hybrid of the two (Alpert et al., 2016; Bettinger et al., 2017; Bowen et al., 2014; Brown & Liedholm, 2002; Coates et al., 2004; Figlio et al., 2013; Jaggars & Xu, 2016; Xu & Jaggars, 2014). However, such comparisons are crude and say nothing about which features of the blended learning format may or may not be helpful. Our experiment takes place in the context of an e-learning tutorial that accompanies a face-to-face lecture. Therefore, analyzing the effects of students' performance in the e-learning sessions on the final exam also extends the blended learning literature.

Methods

Participants

We observed sociology students from a large, public German university enrolled in a statistics course for the social sciences (Social Science Statistics 2) during winter semester 2018/2019. The course covered the topics of probability theory, random variables, discrete and continuous distributions, specific distributions, multidimensional random variables, limit theorems and sampling, point estimation, confidence interval estimation, statistical tests, and regression analysis. The course is intended for the third semester of a Bachelor of Science in Sociology degree program. Of the students who took the final exam, ten students were repeaters (individuals who took the course in a previous semester and missed or failed the exam) who did not participate in a single e-learning session and were therefore excluded from the analysis. In addition, five students were excluded from the analysis because they had not passed the exam for the precursor course (Social Science Statistics 1). This resulted in a final sample size of $N=87$.

The experiment was part of the mandatory weekly tutorial for the statistics course. However, participation in the study was voluntary. Students who did not want to participate in the experiment could continue to use the e-learning environment without having their data be collected. All students who regularly attended the tutorial gave their written consent to participate in the present study. Students were allowed to miss up to two tutorial sessions per semester without giving a reason. Otherwise, they were not allowed to take the exam (this was also the case in the years preceding the study). However, individuals who had taken the course in a previous semester and missed or failed the exam ("repeaters") were not required to attend the course lectures or tutorials, but could do so voluntarily. Students received no compensation for participating in the study. A local ethics committee approved the study.

Table 1 shows summary statistics for the 87 students for whom we obtained all necessary information, as described above. We standardized the results of the first and second final exam dates to include both in a regression. 31 of 87 students took the final exam on the second date, with 10 students who failed the exam on the first date repeating the exam as a

Table 1 Descriptive statistics: cross section data

	N	Mean	SD	Min	Max
Outcome					
Standardized points in end exam	87	0.07	0.99	-1.71	2.29
Treatment					
Treatment group 1 (of 2)	87	0.49	0.50	0.00	1.00
Exam information					
Second trial	87	0.36	0.48	0.00	1.00
Number of trials	87	0.11	0.32	0.00	1.00
Individual information					
Female	87	0.64	0.48	0.00	1.00
Age group below 20	87	0.28	0.45	0.00	1.00
Age group above 23	87	0.18	0.39	0.00	1.00
Pre-treatment ability measures					
Standardized Statistics 1 grade	87	-0.01	0.99	-2.52	1.40
Year Statistics 1 was written	87	2017.74	0.44	2017.00	2018.00
Points in pretest	87	8.63	7.21	0.00	24.00
Missed pretest	87	0.17	0.38	0.00	1.00
Posttest					
Points in posttest	87	7.02	6.29	0.00	18.67
Missed posttest	87	0.29	0.46	0.00	1.00
Global e-learning session information over 12 weeks					
Mean proportion of correct answers in the sessions over 12 weeks	87	0.58	0.27	0.00	0.95
Mean of missing exercises over each session	87	3.22	3.79	0.00	12.00
Mean of the number of mistakes per sessions	87	1.31	0.81	0.31	3.14
Global preparation counts over 12 weeks					
Number of lectures visited	87	5.39	4.06	0.00	12.00
Number of videos watched	87	4.64	3.93	0.00	12.00
Number of exercise sheets worked on	87	5.21	4.34	0.00	12.00
Number of exercise sheets solved	87	3.21	3.68	0.00	12.00

Notes: Only the students who took the exam are included in this table. Further, if students did not participate in the pre- or posttest, we set their points to zero second attempt, i.e., sitting for the final exam twice within the semester. The aforementioned repeaters, who were excluded from the analyses, are students who had enrolled in the course the year before.

About 64% of the students in the class were female; 24 students are under 20 years old, while 16 are over 23 years old. Some students took the exam of the precursor course in 2017, while the majority took that exam in the previous semester (summer semester 2018). We further asked students to solve a pretest at the beginning of the semester, but 14 students missed this pretest. The average pretest score was 8.63 out of 24. Even more students missed the posttest at the end of the semester one week before the exam (25). Therefore, we do not focus on the pretest and posttest in our analysis.

Learning Outcomes and Performance Measures

We measured three phases of learning within the learning environment. The first is performance in the exercises during the learning phase itself. The second is an example exam question from past semesters that was presented immediately after the learning phase. The third is the same question presented again at the beginning of the following week's session to analyze learning after a one-week delay.

In addition, we also analyzed students' final exam grades, for which we should not find a treatment effect, since the e-learning environment with all feedback hints was made accessible to all students at the end of the semester. Students usually study intensively in the one to two weeks before the exam and should therefore have been able to catch up on what they missed. There was no way to measure whether students reviewed the topics for which they received the treatment more quickly.

No one scored 100% on each item of the exercises over the twelve weeks of the e-learning tutorial. The highest score was 95% of correct items for one e-learning sessions, and the mean was 58%, just over half the potential points available. On average, students failed to complete 3.22 items per session because they ran out of time.

Also, students averaged a few more errors than items per session. Note that students could make up to three attempts per exercise if their first two answers were incorrect. It was rare for students to have everything right on the first try. This suggests that the intervention took place in an environment where students need help.

Moreover, at the beginning of the e-learning sessions, students self-reported how well they had prepared for the tutorial in terms of watching instructional videos and solving exercises sheets preparing for the e-learning session (see 2.3 Course structure). The last four entries in Table 1 show that students had attended the week's lecture 60% of the time (lecture attendance was not mandatory). In about half of cases, they had watched the tutorial videos before the e-learning session; slightly more often, they had worked on the exercise sheet before the e-learning session, but only in 36% of cases did they think they had solved it. These variables were used to control for different preparation behavior prior to the e-learning session and as a general measure of students' engagement with the study material when analyzing their exam performance.

In addition, Table 2 provides information on each weekly session, some of which is summarized across all weeks at the end of Table 1. We measure session performance in seven different ways: (i) proportion of correct answers in the learning phase (LP), (ii) effective proportion of correct answer in the learning phase, i.e., only including finished exercises

Table 2 Descriptive statistics: panel data

	N	Mean	SD	Min	Max
Weekly e-learning session achievements					
Proportion of correct answers in the learning phase (LP)	745	0.66	0.28	0.0	1.00
Effective proportion of correct answer in the learning phase, i.e., only including finished exercises (ELP)	745	0.70	0.27	0.0	1.00
Mean number of mistakes per try in the learning phase (MIST)	745	1.09	0.89	0.0	6.42
Relative number of exercises missed to solve until the end of the learning phase (MISS)	745	0.08	0.17	0.0	1.00
Bias-score (BIAS)	745	-25.01	22.33	-94.8	50.51
Proportion of correct answers in the immediate testing (IT)	745	0.28	0.31	0.0	1.00
Proportion of correct answers in the delayed testing (DT)	578	0.35	0.33	0.0	1.00
Treatment condition					
Treatment condition	745	0.50	0.50	0.0	1.00
Self-reported weekly offline preparation					
Visited the lecture	745	0.61	0.49	0.0	1.00
Watched the tutorial video	745	0.53	0.50	0.0	1.00
Worked on the exercise sheet	745	0.60	0.49	0.0	1.00
Solved the exercise sheet	745	0.36	0.48	0.0	1.00

Note: The table shows the variables for over each e-learning session

(ELP), (iii) mean number of mistakes per try in the learning phase (MIST), (iv) relative number of exercises missed to solve until the end of the learning phase (MISS), (v) a bias score (BIAS; metacognitive bias based on the last attempt for each task, see Schraw 2009), (vi) proportion of correct answers in the immediate testing (without hints; IT), and (vii) its repetition one week later (proportion of correct answers in the delayed testing; DT). Note that students had three attempts per exercise and could have up to three times the number of errors as exercises per session. In addition, the tables show that the treatment conditions are balanced evenly and the student' self-reported level of preparation each week.

Course structure

The experiment spanned 15 tutorial sessions. The first and last sessions were reserved for the pretest and posttest, and one session before the posttest was reserved for an introduction to the statistical software STATA. The remaining 12 weeks were mandatory e-learning sessions in which students were asked to complete assignments (see below). At the beginning of the semester, students enrolled in one of six tutorial groups. These groups determined the day and time of the weekly tutorial. The tutorial was held in a PC lab, and students could not attend from home. The face-to-face lectures were not affected by the experiment.

Lecture slides were uploaded at the beginning of the semester, and there were (some) pre-recorded lecture videos from previous semesters. The tutorial had a different format than in the previous semester. The original tutorial was a traditional face-to-face tutorial where tutors explained how to solve the assignment sheets. Now, the tutorial took place in computer labs and consisted of e-learning sessions where students had to solve assignments. With slight (numerical) modifications, these problems represented a subset of the previously uploaded problem sheets from earlier semesters. This allowed students to prepare for the e-learning session by working on the problem sheets. Because students might have felt overwhelmed trying to solve the problems using the lecture slides alone, we also pre-recorded solution videos for each problem sheet. The videos and practice sheets tackled the same (example) problems, while the e-learning session exercises had different numerical examples/solutions.

The e-learning sessions and all materials were integrated into the university's online learning management system. The university uses the open-source online learning management system ILIAS (www.ilias.de/en/). The learning environment had a homepage that was accessible after logging in with one's student ID. All content was made available on this page. Students had access to the e-learning sessions only during their respective tutorial time.

Structure of the e-learning session

The experimental sessions followed the same structure throughout the semester, consisting of three main parts: (1) an example exam question on the previous week's topics for delayed testing, (2) e-learning phase with experimental manipulation, and (3) an example exam question on the e-learning phase topics for immediate testing. Specifically:

(1) Students had 15 min to answer an example exam question on the previous week's topics that was identical to the example exam question at the end of the previous week's session. The exam questions consisted of either one or two main questions with corresponding sub-questions, whose length was tailored to be solvable within 15 min. In the first session, this question covered the contents from the pretest. In all other sessions, the exam question consisted of several open-ended statistical problems. Students received no feedback while working on these problems. In addition, to mimic traditional pen-and-paper exam situations as closely as possible, these questions were displayed on a single page (with the ability to scroll). At the end of the 15 min, participants automatically moved on to the learning phase.

(2) The learning phase included the experimental manipulation. For up to 50 min, participants completed statistical exercises. The number of exercises depended on the length of each task. Each exercise was displayed on a single page, and participants had to submit a response to continue on to the next exercise. When a response was submitted, students received feedback with multiple trials. Specifically, if the answer was correct, students were informed that their answer was correct (i.e., that they knew the answer), and the button allowing them to continue was activated. In addition, in the experimental condition, the elaborative feedback message was displayed. If their answer was incorrect, students were informed of this and asked to try again. In addition, in the experimental condition, an elabo-

rative hint was displayed. If students still did not provide the correct answer on their third attempt, the correct response was displayed, and students could proceed to the next question. In addition, if students provided a response in an inappropriate format (e.g., a word when a number was required), they received feedback that their response was invalid. This did not count as a solution attempt. Participants were further asked to indicate their confidence that their answer was correct or incorrect. After 50 min, students automatically proceeded to the final phase even if they were not finished with this section.

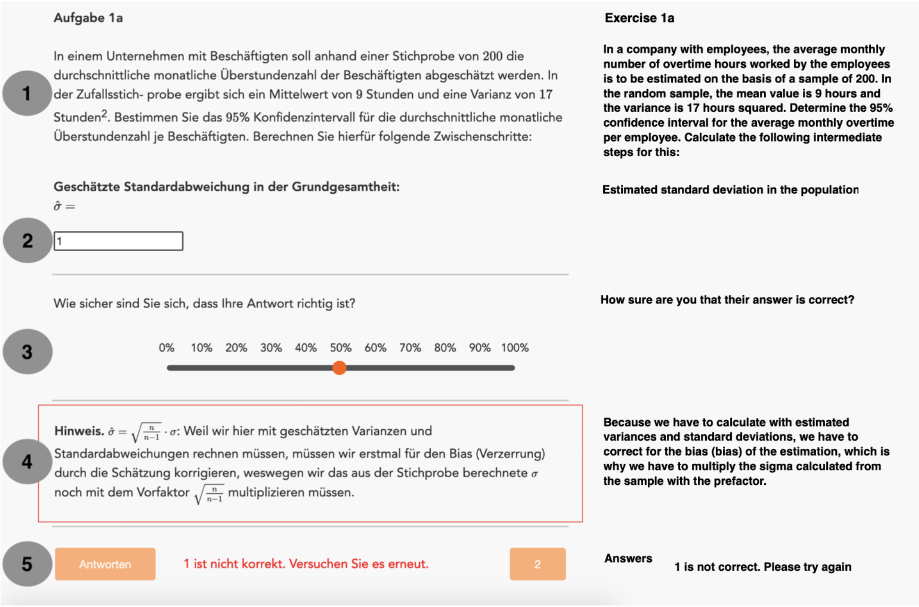
(3) In the last part of the session, students had 15 min to complete an example exam question that covered the learning phase content. This question was structured identically to the one in the first part of the session. The same question was then presented at the beginning of the following week's tutorial.

Most of the questions were numerical in nature: for example, students had to calculate a probability or variance. In addition, multiple-choice questions with four answer options requiring students to select an explanation or statistical distribution of a random variable

Table 3 Study Design

Session 1	Session 2	Session 3	Session 4	Session 5	…	Session 12
MCR+H	MCR+H	MCR	MCR+H	MCR	…	MCR+H
MCR	MCR	MCR+H	MCR	MCR+H	…	MCR

Note: The treatment varied weekly between the two groups. Unfortunately, there was a problem at the beginning, which is why the weekly within change started after the second session



Aufgabe 1a

In einem Unternehmen mit Beschäftigten soll anhand einer Stichprobe von 200 die durchschnittliche monatliche Überstundenzahl der Beschäftigten abgeschätzt werden. In der Zufallsstich- probe ergibt sich ein Mittelwert von 9 Stunden und eine Varianz von 17 Stunden^2. Bestimmen Sie das 95% Konfidenzintervall für die durchschnittliche monatliche Überstundenzahl je Beschäftigten. Berechnen Sie hierfür folgende Zwischenschritte:

Geschätzte Standardabweichung in der Grundgesamtheit:

$\hat{\sigma} =$ 1

Wie sicher sind Sie sich, dass Ihre Antwort richtig ist?

Hinweis. $\hat{\sigma} = \sqrt{\frac{n}{n-1}} \cdot \sigma$: Weil wir hier mit geschätzten Varianzen und Standardabweichungen rechnen müssen, müssen wir erstmal für den Bias (Verzerrung) durch die Schätzung korrigieren, weswegen wir das aus der Stichprobe berechnete σ noch mit dem Vorfaktor $\sqrt{\frac{n}{n-1}}$ multiplizieren müssen.

1 ist nicht korrekt. Versuchen Sie es erneut.

Exercise 1a

In a company with employees, the average monthly number of overtime hours worked by the employees is to be estimated on the basis of a sample of 200. In the random sample, the mean value is 9 hours and the variance is 17 hours squared. Determine the 95% confidence interval for the average monthly overtime per employee. Calculate the following intermediate steps for this:

Estimated standard deviation in the population

How sure are you that their answer is correct?

Because we have to calculate with estimated variances and standard deviations, we have to correct for the bias (bias) of the estimation, which is why we have to multiply the sigma calculated from the sample with the prefactor.

Answers **1 is not correct. Please try again**

Fig. 1 Example screen of the e-learning environment with English translation on the right side

were employed to check students' general understanding. At the very beginning and end of each session, we also asked students about their self-reported emotions, which are not analyzed in this paper because we received feedback that students perceived these questions as intrusive in initial sessions and altered their response behavior with regards to these questions for subsequent sessions (e.g., indicating the same value for all emotion items).

Self-report surveys were also administered in the pretest and posttest sessions (e.g., general mental ability; Raven & Raven 2003), but cannot be examined in-depth due to a high number of missing values. This likely occurred because students were allowed to miss two sessions, and many students did not attend the first and last sessions. Table 3 provides an overview of the course structure.

Experimental manipulation

In this study, two types of feedback were selected as experimental manipulations: MCR and MCR+H, as described above. Participants had up to three attempts to answer the open-ended or multiple-choice question in both conditions. After an incorrect answer in the MCR condition, participants were informed that their answer was incorrect and asked to try again. After the third incorrect attempt, the correct answer was displayed. For the MCR+H group, the only difference was that the participants additionally got a hint after the first incorrect answer (see Fig. 1 below for an example). For each question, only one hint was displayed. In the MCR+H condition, students who provided a correct answer on the first attempt were still given the additional hint to ensure that all students in the treatment group received the same information.

The hint generally provided students with information about crucial steps to solving the statistical problem at hand. For example, in one task, students were instructed to calculate the variance of a population (see Fig. 1). The note explained that in such cases, the standard deviation must be calculated using the estimated variance and showed how the estimated variance fits into the formula for calculating the standard deviation.

The experimental manipulation was randomized at the individual level. Specifically, half of the students started with a MCR session and the other half with a MCR+H session; both groups then alternated between MCR+H and MCR sessions in opposite order, as shown in Table 3. The distance between students in the lab in which they completed the e-learning sessions was large enough that it required significant effort to look at another student's screen. In addition, the tutor overseeing the session ensured that students did not collaborate.

Due to a technical problem, the MCR+H vs. MCR designation was not calibrated correctly for the second session, resulting in half of the participants having five MCR+H sessions and seven MCR sessions, and the other half vice versa. To ensure that this technical problem did not affect the results, all analyses in the present study were repeated without the first and second sessions. The reported results did not change. Additionally, the robustness of results with and without including the first two sessions indicated that familiarity with the new e-learning environment and differences in technology skills did not alter the results (Clarke et al., 2005).

All materials including the additional hints were made available to all students after they had completed the posttest to provide all students with equal study opportunities before the final exam. Thus, all students eventually had access to the elaborated feedback for all sessions.

The e-learning environment

The e-learning environment included all content from the tutorial sessions, including self-reported preparation and lecture attendance. The e-learning environment was implemented with Qualtrics® survey software (Qualtrics, 2020) using JavaScript modifications and displayed in a web browser during the sessions. MathJax (Cervone, 2012) was used to display mathematical formulas. An example is shown in Fig. 1.

The e-learning environment had a linear structure, with backward navigation disabled. Before each phase of the tutorial session (see Sect. 2.3), instructions for the upcoming phase and the corresponding time limit were displayed. The timer for the phase started as soon as participants advanced from that page. For the delayed testing at the beginning of the session for previous week's topics (i.e., old exemplary exam questions), and immediate testing at the end of the session (i.e., new exemplary exams questions), all subquestions and their corresponding response confidence were displayed on a single page. During the learning phase, each task was displayed on a separate page. Such pages consisted of (1) the framing and question, (2) response field(s), (3) response confidence slider, (4) elaborated feedback message, and (5) control buttons with the performance feedback field. Participants had to enter an answer in the corresponding field (2) and indicate their answer confidence (3) before submitting the answer via the answer button (5). Depending on the experimental conditions, participants might receive performance feedback (5) and elaboration (4) after submitting their answer. If the answer was incorrect, the feedback message had red font and borders. Participants were prompted to try again in the performance feedback field, and the response confidence slider was reset (i.e., centered at 50%). In addition, the counter on the next button indicating the number of attempts remaining decreased by one. If the third attempt was still incorrect, the answer button was activated, and the student was prompted to proceed to the next exercise. If the student submitted a correct answer, the feedback font and border color was displayed in green, and the next button (5) was activated. For multiple-choice questions, the procedure was identical, with students needing to select an answer and indicate their confidence before submitting a response. The procedure for questions with multiple response fields (e.g., requiring participants to calculate an estimated confidence interval for a point estimate) was analogous. Here, participants could submit only one response at a time (e.g., they had to fill in the lower bound of the interval before filling in the upper bound). Participants had three attempts per response field (e.g., three trials for the lower bound and three trials for the upper bound of the confidence interval). All submitted answers, and the corresponding response confidence values were recorded in logfiles.

Analysis

To evaluate the treatment group effect on outcomes within the e-learning sessions, we apply random- and fixed-effects methods with clustered standard errors at the individual level:

$$Session_{it} = \rho \text{Treatment-Group}_{it} + X'_{it} \cdot \beta + \mu_i + \varepsilon_{it} \tag{1}$$

where the index *i* remains the same and *t* contains the time dimension representing the twelve sessions. The outcome variable $Session_{it}$ is a placeholder for the weekly e-learning sessions achievements presents in Table 2: LP, ELP, MIST, MISS, BIAS, IT, and DT. For the fixed-effects method, the general intercept and all constants are included in the individual fixed effect μ_i. Thus, X'_{it} includes only observed variables that change from session to session, such as the weekly information on preparation (see Table 2 for the variables that change weekly). Thereby, ε_{it} includes only nonconstant unobserved characteristics. In the random-effects method, X'_{it} also includes the constant control variables shown in Table 1. For the fixed-effects method, we subtracted the group mean from each variable to demean Eq. (1), which then cancels out μ_i since this parameter is constant over time. For the random-effects method, the mean is weighted by a ratio of the variance within and between groups. Therefore, μ_i is not canceled out, and we must assume that no constant unobserved variables bias the estimate. The regression results of the fixed- and random-effects methods in our study are very close, suggesting that the random-effects results are unbiased. Since the treatment was randomly assigned at the beginning of the semester, we did not expect much difference between the two methods. Therefore, we only report the random-effects results in this paper. Interested readers can find the fixed-effects results in the online appendix.

Next, we apply an ordinary least squares (OLS) with heteroskedastic robust standard errors to analyze possible effects of treatment group on exam scores at the end of the semester. The model is as follows:

$$Exam_i = \alpha + \lambda_1 performance_i + \lambda_2 preparation_i + \rho\text{Treatment-Group}_i + X'_i \cdot \beta + \varepsilon_i \quad (2)$$

where index *i* is the student, α is the intercept, and ε_i is the idiosyncratic error term. The outcome Exam_i is the standardized final exam score at either the first or second date. In this regression, we are primarily interested in the performance and preparation variables, which measure student performance during the e-learning session and behavior before the e-learning session, respectively. For performance, we use the mean percentage of correct answers during the twelve weeks as well as the mean number of tasks not answered and the mean number of errors per session over the twelve weeks. For student preparation, we use the mean number of lectures attended, number of learning videos viewed, number of exercise sheets worked on and completed, and a sum of all four variables. A positive coefficient for performance would (i) indicate that effort is conducive to exam performance and (ii) demonstrate the importance of improving student performance in these e-learning sessions. The coefficient ρ of the treatment group variable, when statistically significantly different from zero, indicates that neither treatment group benefited more from the additional hints than the other. Because our treatment changed weekly and all students received access to the hints two weeks before the exam, we did not expect to find a treatment group effect on exam scores. We include this variable to ensure that our treatment did not have a random negative effect.X'_i is a vector of all control variables, and β is the corresponding vector of coefficients. The control variables are shown in Table 1.

Results

E-Learning sessions

First, we analyzed whether additional hints within the session contributed to students performing better on the immediate and delayed testing (i.e., the exemplary exam questions) at the end of the session and at the beginning one week later. Table 4 shows the results of the random-effects model for different session outcomes. The dependent variable in column (1) of Table 4 is the proportion of correct answers in the learning phase (LP); in column (2) the

Table 4 Several session outcomes (Panel: Random Effects)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Dependent variables:						
	LP	ELP	MIST	MISS	BIAS	IT	DT
Treatment condition	0.037^{*}	0.047^{**}	-0.143^{**}	0.009	-2.026^{+}	0.011	0.046^{*}
	(0.015)	(0.015)	(0.052)	(0.011)	(1.216)	(0.016)	(0.023)
Lecture visited	0.013	0.019	-0.126^{+}	0.003	3.861^{*}	0.075^{**}	0.125^{***}
	(0.021)	(0.020)	(0.066)	(0.018)	(1.706)	(0.028)	(0.024)
Video watched	0.109^{***}	0.103^{***}	-0.280^{**}	-0.043^{+}	-3.655	0.044	0.082^{*}
	(0.030)	(0.026)	(0.089)	(0.026)	(2.653)	(0.035)	(0.041)
Exercise sheet worked on	0.015	0.029	-0.271^{*}	0.025	4.016^{+}	0.058	-0.007
	(0.031)	(0.030)	(0.118)	(0.028)	(2.423)	(0.039)	(0.043)
Exercise sheet solved	0.084^{***}	0.069^{***}	-0.208^{***}	-0.032^{+}	-4.592^{***}	0.071^{*}	0.088^{**}
	(0.021)	(0.020)	(0.062)	(0.017)	(1.385)	(0.033)	(0.030)
Statistics 1	0.045^{*}	0.049^{**}	-0.113^{*}	0.001	0.398	0.042^{*}	0.033^{*}
	(0.019)	(0.018)	(0.050)	(0.009)	(1.828)	(0.020)	(0.015)
Year of Statistics 1	-0.090^{**}	-0.112^{***}	0.314^{**}	-0.014	1.083	-0.055	-0.066
	(0.034)	(0.034)	(0.120)	(0.027)	(4.077)	(0.036)	(0.042)
Female	0.003	0.0002	0.042	0.012	-9.521^{**}	-0.023	-0.014
	(0.031)	(0.032)	(0.104)	(0.023)	(3.684)	(0.033)	(0.032)
Age group below 20	0.004	-0.001	0.050	-0.010	5.058	0.052	0.077^{*}
	(0.035)	(0.035)	(0.099)	(0.023)	(4.786)	(0.040)	(0.034)
Age group above 23	-0.022	-0.026	0.023	0.042	6.771	0.051	0.011
	(0.065)	(0.060)	(0.173)	(0.036)	(4.441)	(0.045)	(0.062)
Pretest points	0.010^{***}	0.008^{**}	-0.026^{**}	-0.003	-0.167	0.012^{***}	0.014^{***}
	(0.003)	(0.003)	(0.008)	(0.002)	(0.367)	(0.003)	(0.003)
Missed pretest	0.059	0.019	-0.224	-0.044	3.514	0.126	0.029
	(0.059)	(0.057)	(0.178)	(0.039)	(6.378)	(0.080)	(0.075)
Observations	718	718	718	718	718	718	557
R^2	0.189	0.194	0.192	0.038	0.034	0.170	0.285

Note: LP: Proportion of correct answers in the learning phase. ELP: Effective ratio of correct answer in the learning phase, i.e., only including exercises finished by the students. MIST: Number of mistakes per trial in the learning phase. MISS: Number of exercises missed to solve until the end of the learning phase. BIAS: metacognitive bias based on the last try of each task (see Schraw, 2009). IT: Proportion of correct answers in the immediate testing based on example exam questions. DT: Proportion of correct answers in the delayed testing. Standard errors are clustered at the individual level. The coefficients of the fixed effects model, which can we found in a supplemental online appendix, are very close to the corresponding random effects coefficients, which should rule out omitted variable bias due to constant variables. $^{+}p<0.10$, $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$

effective proportion of correct answer in the learning phase, i.e., only including exercises finished by the students (ELP); column (3) the mean number of mistakes per trial in the learning phase (MIST), column (4) the relative number of exercises missed to solve until the end of the learning phase (MISS); column (5) the metacognitive bias score based on the last try of each task (Schraw, 2009); column (6) immediate testing using the proportion of

Table 5 New and old exam question including mediators (Panel: Random Effects)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A	*Dependent variable*: Proportion of correct answer in **immediate** testing						
Treatment condition	0.001	0.001	0.005	0.013	0.012	0.004	0.005
	(0.016)	(0.016)	(0.017)	(0.016)	(0.016)	(0.014)	(0.014)
Mean correct answers	0.246***					0.399***	
	(0.038)					(0.072)	
Effective mean correct answers		0.200***					0.278***
		(0.037)					(0.064)
Mean mistakes			-0.040***			-0.017^{+}	-0.027*
			(0.010)			(0.010)	(0.011)
Sum of missings				-0.280***		-0.136*	-0.311***
				(0.045)		(0.055)	(0.048)
Bias-Score					0.001	0.004***	0.004***
					(0.001)	(0.001)	(0.001)
Additional controls as in Table 4	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	718	718	718	718	718	718	718
R^2	0.235	0.209	0.187	0.209	0.172	0.277	0.263
Panel B	*Dependent variable*: Proportion of correct answers in **delayed** testing						
Treatment condition	0.038^{+}	0.037	0.042^{+}	0.048*	0.048*	0.044*	0.039^{+}
	(0.023)	(0.023)	(0.024)	(0.023)	(0.023)	(0.021)	(0.020)
Mean correct answers	0.184***						0.263***
	(0.049)						(0.067)
Effective mean correct answers		0.167**					
		(0.052)					
Mean mistakes			-0.022^{+}				0.019
			(0.012)				(0.012)
Sum of missings				-0.114^{+}			0.133^{+}
				(0.066)			(0.074)
Bias-Score					0.0004		0.002*
					(0.001)		(0.001)
Previous exam question results						0.472***	0.450***
						(0.049)	(0.046)
Additional controls as in Table 4	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	557	557	557	557	557	557	557
R^2	0.337	0.333	0.322	0.289	0.283	0.497	0.540

Note: Panel A shows the results for the new exam question of the same week, while panel B shows the results for the old exam question in the thereafter. Heteroskedastic robust standard errors in parentheses. $^{+}p<0.10$, $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$

correct answers in the new (exemplary) exam question (IT); and column (7) delayed testing using proportion of correct answers in the old (exemplary) exam question in the following week (DT).

Column (1) shows that treatment, i.e., receiving additional hints, increases performance by 3.7%. When we consider in column (2) only the tasks students actually reached (they might have been pressed for time and unable to solve everything), the treatment coefficient increases to 4.7%. However, the results also show that other variables, such as whether the preparation videos were watched, had a more substantial effect on performance (e.g., watching the videos prior to the session led to a 10.9% increase in performance). Nonetheless, the treatment variable had significant additional explanatory value. Column (3) shows that students with additional feedback are 14.3% less likely to make errors. For the number of missing exercises in column (4), we find no significant result. The negative coefficient in the model explaining the bias score (column (5)) is significant at the 10% level and corresponds to a 2% reduction in bias score.

Next, we analyze performance on the immediate testing after the learning period and delayed testing one week later. We find no statistically significant increase in correct answers for the immediate testing (column (6)), but for the delayed testing (column (7)). One possibility is that students became accustomed to the feedback, and not receiving it led to a worse outcome. Another possibility could be the relatively long duration of the tutorial session. Students may have been fatigued from solving the tasks and therefore performed worse overall. Nevertheless, column (7) shows that students in the MCR+H group learned more than the MCR group one week later.

Panel A in Table 5 further shows that none of the e-learning results presented in Table 4 from column (1) to (5) affect the lack of treatment effect on the immediate testing at the end of the session. The coefficient is still small and not statistically significant, confirming the results in Table 4 column (6). However, providing reassurance that the model is plausible in general, we see that better performance in the e-learning session is associated with better performance on the immediate testing.

For the delayed learning outcome, i.e., the delayed testing at the beginning of the next session, Panel B in Table 5 shows a relatively stable treatment condition estimate, confirming the result in Table 4 column (7). The different e-learning outcomes only slightly reduce the treatment coefficient (columns 1, 2, 3, 6, and 7). In general, a reduction would only indicate that improved performance in the session accounts for the positive main effect of the treatment on the delayed testing. However, the reduction is small in magnitude. This could indicate a treatment effect one week later that is not fully captured by performance in the e-learning session. It seems that students who receive additional hints benefit from this without necessarily exhibiting better performance in the session itself.

Exam

Next, we analyzed whether performance in the mandatory online sessions explained final exam scores. Therefore, in Table 6, we include several e-learning session variables in a regression explaining exam scores: the mean percentage correct across all sessions in column (1), the mean number of missing tasks per session in column (2), and the mean number of errors in column (3). We then include all three performance measures from the e-learning sessions in column (4) to see which variables remain important. Column (5) adds overall

Table 6 Exam Results including Session Information

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Dependent variable: Exam points						
Treatment condition	0.049	-0.046	0.0003	0.053	0.034	0.064	0.051
	(0.200)	(0.217)	(0.204)	(0.200)	(0.194)	(0.168)	(0.163)
Exam trial	0.376	0.082	0.339	0.383	0.460*	0.455**	0.509**
	(0.267)	(0.264)	(0.266)	(0.269)	(0.271)	(0.232)	(0.240)
Number of exam trials	-0.088	0.066	-0.127	-0.087	-0.101	-0.029	-0.040
	(0.338)	(0.332)	(0.345)	(0.350)	(0.356)	(0.261)	(0.266)
Mean correct answers	1.903****			2.271*	1.889	1.891*	1.625
	(0.407)			(1.289)	(1.309)	(1.118)	(1.113)
Mean number of errors		-0.069***		0.005	0.027	-0.005	0.011
		(0.024)		(0.042)	(0.037)	(0.037)	(0.034)
Sum of missings			-0.564****	0.112	0.132	0.263	0.272
			(0.125)	(0.372)	(0.383)	(0.316)	(0.322)
General preparation					0.069*		0.050*
					(0.037)		(0.030)
Statistics 1						0.501****	0.486****
						(0.083)	(0.088)
Year of statistics 1						-0.208	-0.197
						(0.226)	(0.222)
Trials	Both	Both	Both	Both	Both	Both	Both
Observations	87	87	87	87	87	87	87
R^2	0.244	0.073	0.203	0.245	0.274	0.446	0.461
Adjusted R^2	0.207	0.028	0.164	0.188	0.210	0.389	0.398

Note: The variable mean correct answers is short for the mean proportion of correct answers for all e-learning sessions. The variable general preparation is the simple sum of the variables visited lectures, videos watched, exercise sheets worked on and exercise sheets solved. A regression estimation with the four variables is very similar. Heteroskedastic robust standard errors in parentheses. + $p<0.10$, *$p<0.05$, **$p<0.01$, ***$p<0.001$

preparation for the e-learning sessions calculated from the (self-reported) attended lectures, tutorial videos watched before the session, exercise sheets worked on prior to the session, and exercise sheets completed prior to the session. Column (6) adds in the student's standardized exam score from the precursor statistics course and the year the student took that exam. Finally, column (7) contains both the preparation and prior exam variables.

Performance (mean correct answers in column (1), (4) to (7); mean number of errors in column (2); sum of missing's in column (3)) in the sessions is statistically significant in all models except columns (5) and (7). Although the coefficient for the variable *mean correct answers* in these two columns is not significant, it robustly remains in the range of 1.625 in column (7) to 2.271 in column (5); here, the number of observations may simply be too small to efficiently estimate all variables. Moreover, adding the standardized exam score from the precursor statistics course and total preparation across all weeks reduces the explanatory power of performance in the sessions. Students who are good at statistics, try harder during the semester, and do better in the e-learning sessions also do better on the exam.

As expected, the first row in all columns shows that treatment group assignment does not affect exam grades. The coefficient is small and insignificant.

Discussion

This study analyzed multiple-try feedback within a university statistics course. Specifically, we examined the incremental value of hints in multiple-try feedback messages on several outcomes over the course of a semester. Improving statistics courses is of general relevance in higher education in light of its importance for fields of study from STEM to the social sciences, and because students report issues specific to statistics (e.g., statistics anxiety; Valle et al., 2021).

First, regarding the additional effect of hints in multiple-try feedback during tutorial sessions, we found that students in sessions with additional hints performed significantly better during the session and on the delayed testing one week later, i.e., the example exam question. However, they did not score higher on the immediate testing at the end of the learning session. At first glance, these results seem to contradict previous studies. Attali (2015) found no effect of multiple-try feedback with or without additional hints during the initial learning phase (equivalent to the learning phase in this study), but significant differences in the exam immediately following. However, this study compared multiple feedback types (across two types of items) in a small experimental study. Our study had a more extensive learning session (approximately 50 min), and some of the tasks in this phase were interrelated. For example, in one session, participants had to proceed through the different steps of a hypothesis test. Therefore, additional hints in early exercises might have positively affected subsequent exercises, which was also reflected in a lower probability of errors in sessions with additional hints in multiple-try feedback (see below).

From an instructional design perspective, the hints during these sessions provided (procedural) information that aided students in acquiring schemas (e.g., Frerejean et al., 2019) that could be applied to subsequent tasks. Furthermore, the instructional design framework provides a potential explanation for the lack of increased performance in tasks directly following the learning phase. Instructional theories emphasize the importance of gradually fading out support as learners apply newly acquired skills (van Merriënboer et al., 2003). The present study did not incorporate fading; thus, learners might not have sufficiently internalized the newly acquired skills and schemas to the extent required to solve the immediate testing questions without support within the time limit. However, the additional hints might have advanced students' understanding sufficiently for them to benefit more from learning activities in-between sessions (e.g., attending the lecture, solving additional statistics problems), which led to increased performance on the delayed test (i.e., one-week follow-up test).

With the delayed tests, we directly addressed a shortcoming of previous research (i.e., the longevity of performance improvement, Attali 2015). Participants performed significantly better on the one-week follow-up tests after receiving feedback with hints (even when controlling for prior knowledge, performance on the same immediate test in the previous week, and lecture attendance in between). This improvement suggests that additional hints have a medium-term effect in applied educational contexts.

Consistent with findings that feedback with multiple trials and hints is particularly effective after initial errors (Attali, 2015), we also found that the likelihood of making errors during the learning phase was reduced in sessions with additional hints. The additional guidance provided by a hint promoting reassessment (e.g., Corbett & Anderson 2001; Lepper & Woolverton, 2002) may be particularly relevant in our study, because the problems within a learning session were often interrelated (e.g., one session involved performing all steps of a hypothesis test). Therefore, the additional hints alerting students to the underlying statistical calculations/formulas may have been helpful for avoiding errors in future steps. Our results further translates to instructional design approaches for courses to acquire complex skills (e.g., the four components of instructional design −4 C/ID; van Merriënboer 1997; van Merriënboer & Kirschner, 2018 and principles of instruction; Merrill, 2002). From this perspective, the additional hints in the feedback messages served as supportive information during practice, which is particularly effective for novice learners who are still acquiring schemata to solve the statistical problems at hand (e.g., by preventing cognitive overload; van Merriënboer et al., 2002; van Merriënboer et al., 2003).

In terms of correcting potential misconceptions, feedback with multiple trials and hints may be more beneficial than elaboration aimed at teaching appropriate methods/approaches to solving a statistical problem. Research has shown that misconceptions are common among students, especially in the social sciences, even with respect to less advanced statistical concepts (Mevarech, 1983). Feedback research has shown that feedback can be particularly effective in correcting errors or misconceptions that were made with high confidence (Kulhavy, 1977). In line with these finding, the additional hints given in our study may have addressed misconceptions more effectively, as reflected by the slightly reduced bias in these sessions.

In addition, when we included the session outcomes to explain the delayed testing (Table 5), the estimation results showed that the improvements in the delayed testing were not just driven by the performance in the e-learning sessions, but the hint added additional, independent explanatory value. Therefore, the effect of the manipulation does not depend on performance in the e-learning session. These results are in line with Clark & Bjork (2014), who showed that students can learn even in the absence of a direct performance increase. Another explanation could be that the *error generation effect* (taking challenging tests in which mistakes are made can promote effective learning; Kornell et al., 2009).

Next, we found that randomizing additional hints in multiple-try feedback within a semester did not affect either group's exam performance. This result was in line with our expectations. It showed that each student had equal learning opportunities despite the experimental manipulation during the tutorial sessions (e.g., students were given access to all materials, including the e-learning environment with hints for all sessions, before the exam). More importantly, performance during the e-learning sessions was a robust predictor of exam grades across multiple models beyond the explanatory value of prior knowledge (i.e., performance on the final exam in the precursor statistics course). This result indicated that good performance in the e-learning tutorial implemented in this study significantly positively affected undergraduate students' statistics performance, which is in line with previous research (Förster et al., 2018; Schwerter et al., 2022). This suggests that multiple-trial feedback is an appropriate tool to guide and promote learning in STEM-related problem-based learning tasks (Attali, 2015; Clariana & Koul, 2005). It is important to note, that these findings do not necessarily imply a causal relationship. Generally, one would assume that students with higher ability would perform better in the e-learning sessions as well as on the final exam.

However, since we control for the subject-specific ability measure of final exam scores for the precursor statistics course, ability bias seems unlikely. Two other important explanations for students' performance are motivation and personality. We did not collect variables for these factors. However, we can infer who generally put more effort into learning statistics from the preparation variables. More motivated students would have been more likely to watch the videos, attend the lectures, and (try to) solve the assignment sheets before the e-learning sessions. Thus, these preparation variables could be a proxy for other individual differences. With the present data, we cannot confirm a causal relationship that better performance leads to better grades. Still, it seems unlikely that the causal effect is far from the relationship we estimated. The statistically significant relationship between performance in the e-learning environment during the semester and final exam performance points to several avenues for further research. Similar to our results, a previous study showed that preparatory e-learning courses do indeed predict performance during the semester (Fischer et al., 2019). Harnessing this relationship between learning in e-learning environments and performance in related courses (including exams) seems particularly promising for the development of timely, individualized interventions. In particular, the ability to track individuals' learning and performance in real time before or during a semester can be used to design digital interventions that can promote learning as it unfolds, potentially circumventing problems before they arise.

The great potential of interventions such as prompts and feedback from pedagogical agents (e.g., Azevedo 2005) has been repeatedly demonstrated in laboratory settings. However, systematic translation to applied educational settings (e.g., university courses) is still needed (Reeves & Lin, 2020). As discussed above, many potential functional mechanisms (including interactions between different factors) may explain the positive effect of hints in multiple-try feedback in a university statistics course. However, the present study cannot shed light on the specific (cognitive) processes that cause these effects. Instead, we focused on transferring previous findings (e.g., Attali 2015) to an applied context. Although limited by contextual constraints, such as the lack of a classical control group without feedback due to ethical concerns, this study showed that the added value of hints in multiple-trial feedback is robust enough to yield a significant effect in "noisy" applied settings. Moreover, in light of the well-established problems in replicating the results of experimental research, studies such as this one demonstrates the robustness of effects across different contexts. Future studies should extend this approach to other interventions and designs to bridge the gap between experimental research and educational practice.

In the university context, alongside the positive effects of e-learning environments (Uzunboylu, 2006), the decreased costs are particularly promising. While the initial design and implementation of such environments may be costly, ongoing costs are low and scalability is high. Ideally, e-learning environments should not replace instructors (or teaching assistant in higher education) but rather be used in conjunction with them, freeing up instructors to focus on more complex educational problems.

Supplemental Online Appendix

Table A1 Several session outcomes (Panel: Fixed-Effects)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Dependent variables:						
	LP	ELP	MIST	MISS	BIAS	IT	DT
Treatment condition	0.038*	0.045**	-0.132*	0.005	-2.159^{+}	0.009	0.048*

Table A1 Several session outcomes (Panel: Fixed-Effects)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	(0.015)	(0.015)	(0.052)	(0.011)	(1.222)	(0.015)	(0.022)
Lecture visited	0.002	0.012	-0.133^{+}	0.010	3.943*	0.070*	0.096***
	(0.021)	(0.020)	(0.070)	(0.016)	(1.747)	(0.028)	(0.029)
Video watched	0.109***	0.095***	-0.230*	-0.066*	-3.659	0.051	0.087^{+}
	(0.032)	(0.028)	(0.100)	(0.029)	(2.615)	(0.037)	(0.044)
Exercise sheet worked on	-0.001	0.014	-0.261*	0.034	4.539^{+}	0.057	-0.032
	(0.032)	(0.029)	(0.123)	(0.028)	(2.461)	(0.042)	(0.050)
Exercise sheet solved	0.077***	0.067***	-0.232***	-0.016	-4.470***	0.038	0.051
	(0.021)	(0.020)	(0.064)	(0.019)	(1.353)	(0.036)	(0.036)
Observations	745	745	745	745	745	745	578
R^2	0.080	0.089	0.103	0.020	0.024	0.047	0.053

Note: LP: Proportion of correct answers in the learning phase. ELP: Effective ratio of correct answer in the learning phase, i.e. only including exercises finished by the students. MIST: Number of mistakes per trial in the learning phase. MISS: Number of exercises missed to solve till the end of the learning phase. IT: Proportion of correct answers in the immediate testing based on example exam questions. DT: Proportion of correct answers in the delayed testing. Standard errors are clustered at the individual level. $^{+}p<0.10$, $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$

Author contribution **Jakob Schwerter**: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing – Original Draft, Writing – Review & Editing, Visualization, Project administration, Funding acquisition. **Franz Wortha**: Conceptualization, Methodology, Software, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing, Funding acquisition. **Peter Gerjets**: Conceptualization, Writing – Review & Editing, Funding acquisition, Supervision.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations This work was supported by an intramural research grant of the LEAD Graduate School and Research Network [GSC1028]. The authors have no conflicts of interest to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

Ethics Approval was granted by the Ethics Committee of the University of Tübingen. All of the students who regularly attended the tutorial gave their written consent to participate in the present study.

References

Alpert, W. T., Couch, K. A., Harmon, O. R., & Gpa, P. (2016). A Randomized Assessment of Online Learning. *American Economic Review: Papers & Proceedings*, *106*(5), 378–382. https://doi.org/10.1257/aer.p20161057

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. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist*, 40(4), 199–209. https://doi.org/10.4324/9781315866239-2

Azevedo, R., & Bernard, R. M. (1995). A Meta-Analysis of the Effects of Feedback in Computer-Based Instruction. *Journal of Educational Computing Research*, 13(2), 111–127. https://doi.org/10.2190/9lmd-3u28-3a0g-ftqt

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research*, 61(2), 213–238. https://doi.org/10.3102/00346543061002213

Bennett, R. E. (2015). The Changing Nature of Educational Assessment. *Review of Research in Education*, 39(1), 370–407. https://doi.org/10.3102/0091732X14554179

Bettinger, E. P., Fox, L., Loeb, S., & Taylor, E. S. (2017). Virtual Classrooms: How Online College Courses Affect Student Success. *American Economic Review*, 107(9), 2855–2875. https://doi.org/10.1257/aer.20151193

Bowen, W. G., Chingos, M. M., Lack, K. A., & Nygen, T. I. (2014). Interactive Learning Online at Public Universities: Evidence from a Six-Campus Randomized Trial. *Journal of Policy Analysis and Management*, 33(4), 1047–1049. https://doi.org/10.1002/pam

Brown, B. W., & Liedholm, C. E. (2002). Can Web Courses Replace the Classroom in Principles of Microeconomics? *American Economic Review*, 92(2), 444–448. https://doi.org/10.1257/000282802320191778

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

Cervone, D. (2012). MathJax: a platform for mathematics on the Web. *Notices of the AMS*, 59(2), 312–316

Clariana, R. B., & Koul, R. (2005). Multiple-try feedback and higher-order learning outcomes. *International Journal of Instructional Media*, 32(3), 239–245

Clariana, R. B., Ross, S. M., & Morrison, G. R. (1991). The effects of different feedback strategies using computer-administered multiple-choice questions as instruction. *Educational Technology Research and Development*, 39(2), 5–17. https://doi.org/10.1007/BF02298149

Clark, C. M., & Bjork, R. A. (2014). When and why introducing difficulties and errors can enhance instruction. In A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying science of learning in education: Infusing psychological science into the curriculum* (pp. 20–30). Society for the Teaching of Psychology. https://doi.org/10.4324/9780203817421

Clark, R. C., & Mayer, R. E. (2016). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning* (4th ed.). John Wiley & Sons, Ltd.

Clarke, T., Ayres, P., & Sweller, J. (2005). The impact of sequencing and prior knowledge on learning mathematics through spreadsheet applications. *Educational Technology Research and Development*, 53(3), 15–24. https://doi.org/10.1007/BF02504794

Coates, D., Humphreys, B. R., Kane, J., & Vachris, M. A. (2004). ``No significant distance'' between face-to-face and online instruction: Evidence from principles of economics. *Economics of Education Review*, 23(5), 533–546. https://doi.org/10.1016/j.econedurev.2004.02.002

Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: impact on learning rate, achievement and attitudes. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 245–252. https://doi.org/10.1145/365024.365111

Cutumisu, M., & Schwartz, D. L. (2018). The impact of critical feedback choice on students' revision, performance, learning, and memory. *Computers in Human Behavior*, 78(June), 351–367. https://doi.org/10.1016/j.chb.2017.06.029

Figlio, D., Rush, M., & Yin, L. (2013). Is It Live or Is It Internet? Experimental Estimates of the Effects of Online Instruction on Student Learning. *Journal of Labor Economics*, 31(4), 763–784. https://doi.org/10.1086/669930

Fischer, C., Zhou, N., Rodriguez, F., Warschauer, M., & King, S. (2019). Improving College Student Success in Organic Chemistry: Impact of an Online Preparatory Course. *Journal of Chemical Education*, 96(5), 857–864. https://doi.org/10.1021/acs.jchemed.8b01008

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

Frerejean, J., van Merriënboer, J. J. G., Kirschner, P. A., Roex, A., Aertgeerts, B., & Marcellis, M. (2019). Designing instruction for complex learning: 4 C/ID in higher education. *European Journal of Education*, 54(4), 513–524. https://doi.org/10.1111/ejed.12363

Hattie, J., & Gan, M. (2011). Instruction Based on Feedback. In R. E. Mayer, & P. A. Alexander (Eds.), *Handbook of Research on Learning and Instruction*. Routledge

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

Jaggars, S. S., & Xu, D. (2016). How do online course design features influence student performance? *Computers and Education*, 95, 270–284. https://doi.org/10.1016/j.compedu.2016.01.014

Kluger, A. N., & DeNisi, A. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. https://doi.org/10.1037//0033-2909.119.2.254

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

Kulhavy, R. W. (1977). Feedback in Written Instruction. *Review of Educational Research*, 47(2), 211–232. https://doi.org/10.3102/00346543047002211

Lepper, M., & Woolverton, M. (2002). The Wisdom of Practice: Lessons Learned from the Study of Highly Effective Tutors. In J. Aronson (Ed.), *Improving Academic Achievement: Impact of Psychological Factors on Education* (pp. 135–158). Academic Press

Mathan, S. A., & Koedinger, K. R. (2002). In S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *An Empirical Assessment of Comprehension Fostering Features in an Intelligent Tutoring System BT - Intelligent Tutoring Systems* (pp. 330–343). Berlin Heidelberg: Springer

Merrill, M. (2002). First principles of instruction. *Educational Technology Research and Development*, 50(3), 43–59. https://doi.org/10.1007/BF02505024

Mevarech, Z. R. (1983). A Deep Structure Model of Students' Statistical Misconceptions. *Educational Studies in Mathematics*, 14(4), 415–429. https://doi.org/10.1007/BF00368237

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

Proske, A., Körndle, H., & Narciss, S. (2012). Interactive Learning Tasks. *Encyclopedia of the Sciences of Learning*, 1606–1610. https://doi.org/10.1007/978-1-4419-1428-6_1100

Qualtrics. (2020). *Qualtrics®*. Qualtrics

Raven, J., & Raven, J. (2003). *Raven Progressive Matrices BT - Handbook of Nonverbal Assessment* (R. S. McCallum (ed.); pp. 223–237). Springer US. https://doi.org/10.1007/978-1-4615-0153-4_11

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

Ross, S. M., & Morrison, G. R. (1989). In search of a happy medium in instructional technology research: Issues concerning external validity, media replications, and learner control. *Educational Technology Research and Development*, 37(1), 19–33. https://doi.org/10.1007/BF02299043

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, 4(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3

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

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189. https://doi.org/10.3102/0034654307313795

Uzunboylu, H. (2006). A Review of Two Mainline E-Learning Projects in the European Union. *Educational Technology Research and Development*, 54(2), 201–209. https://doi.org/10.1007/s11423-006-8255-7

Valle, N., Antonenko, P., Valle, D., Sommer, M., Huggins-Manley, A. C., Dawson, K. … Baiser, B. (2021). Predict or describe? How learning analytics dashboard design influences motivation and statistics anxiety in an online statistics course. *Educational Technology Research and Development*, *0123456789*. https://doi.org/10.1007/s11423-021-09998-z

van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers and Education*, 58(1), 263–272. https://doi.org/10.1016/j.compedu.2011.07.020

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. *Review of Educational Research*, 85(4), 475–511. https://doi.org/10.3102/0034654314564881

van Merriënboer, J. J. G. (1997). *Training complex cognitive skills: A four-component instructional design model for technical training*. Educational Technology Publications, Inc

van Merriënboer, J. J. G., Clark, R. E., & de Croock, M. B. M. (2002). Blueprints for complex learning: The 4 C/ID-model. *Educational Technology Research and Development*, 50(2), 39–64. https://doi.org/10.1007/bf02504993 van Merriënboer, J. J. G., & Kirschner, P. A. (2018). 4 C/ID in the Context of Instructional Design and the Learning Sciences. In P. R. Frank Fischer, Cindy E. Hmelo-Silver, Susan R. Goldman (Ed.), *International Handbook of the Learning Sciences* (pp. 169–179). Routledge. https://doi.org/10.4324/9781315617572-17

van Merriënboer, J. J. G., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational Psychologist*, 38(1), 5–13. https://doi.org/10.1207/S15326985EP3801_2

Wang, H., & Lehman, J. D. (2021). Using achievement goal-based personalized motivational feedback to enhance online learning. *Educational Technology Research and Development*, 69(2), 553–581. https://doi.org/10.1007/s11423-021-09940-3

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

Xu, D., & Jaggars, S. S. (2014). Performance Gaps Between Online and Face-to-Face Courses: Differences Across Types of Students and Academic Subject Areas. *Journal of Higher Education*, 85(5), 633–659. https://doi.org/10.1080/00221546.2014.11777343

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dr. Jakob Schwerter is a PostDoc at the TU Dortmund, Center for Research on education and School Development, and an associated member of the LEAD graduate school and research network. For his Ph.D. he worked at the University of Tübingen, Chair for Econometrics, Statistics and Quantitative Methods. The title of his dissertation is Econometric Analysis of the Effects of Educational Decisions on Labor Market Outcomes and the Influence of Self-Testing on Learning Outcomes. His main research focus lies in the field of education economics, e-learning in (higher) education, educational data science, as well as policy analysis.

Franz Wortha is a PostDoc at the Department of Psychology at the University of Greifswald, and an associated member of the LEAD graduate school and research network. He was a PhD student at the LEAD Graduate School and Research Network at the University of Tübingen investigating facets of self-regulation and their interaction in educational settings. He is particularly interested in analyzing metacognitive and affective processes when learning with computer-based learning environments using process and interaction data.

Prof. Dr. Peter Gerjets is head of the multimodal interaction lab at the Leibniz-Institut für Wissensmedien in Tübingen. His broad spectrum of research interests involves the role of learner strategies and learner characteristics in digital environments, learning with hypermedia and dynamic visualizations, embodied interaction with multi touch devices, and the relation between working-memory limitations and comprehension proces ses.