Received: 28 January 2020 | Revised: 25 June 2020 | Accepted: 1 August 2020

DOI: 10.1111/jcal.12491

ARTICLE

Journal of Computer Assisted Learning WILEY

Combined inner and outer loop feedback in an intelligent tutoring system for statistics in higher education

Sietske Tacoma[1] | Paul Drijvers[1] | Johan Jeuring[2,3]

[1]Freudenthal Institute, Utrecht University, Utrecht, The Netherlands

[2]Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

[3]Department of Computer Science, Open University of the Netherlands, Heerlen, The Netherlands

Correspondence
Sietske Tacoma, Freudenthal Institute, Utrecht University, P.O. Box 85170, 3508 AD Utrecht, The Netherlands.
Email: s.g.tacoma@uu.nl

Funding information
Utrecht University Education Incentive Fund

Peer Review
The peer review history for this article is available at https://publons.com/publon/10.1111/jcal.12491.

Abstract

Intelligent tutoring systems (ITSs) can provide inner loop feedback about steps within tasks, and outer loop feedback about performance on multiple tasks. While research typically addresses these feedback types separately, many ITSs offer them simultaneously. This study evaluates the effects of providing combined inner and outer loop feedback on social sciences students' learning process and performance in a first-year university statistics course. In a 2 x 2 factorial design (elaborate inner loop vs. minimal inner loop and outer loop vs. no outer loop feedback) with 521 participants, the effects of both feedback types and their combination were assessed through multiple linear regression models. Results showed mixed effects, depending on students' prior knowledge and experience, and no overall effects on course performance. Students tended to use outer loop feedback less when also receiving elaborate inner loop feedback. We therefore recommend introducing feedback types one by one and offering them for substantial periods of time.

KEYWORDS

domain reasoner, feedback, inspectable student models, intelligent tutoring systems, statistics education

1 | INTRODUCTION

Over the past decades, a huge number of computer-based learning environments have been developed that facilitate learning of many topics at all educational levels. One of their largest promises for enhancing learning is the provision of individualized and timely feedback on student work (Pardo, 2018; VanLehn, 2011). Fulfilling this promise is not straightforward, though, because there are many design choices to make when implementing feedback, regarding specificity, timing, type and complexity of information provided, and visual presentation (Shute, 2008). To better understand the consequences of such design choices, many theories have been developed about whether and how feedback contributes to student learning and motivation, both in general (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008) as well as specifically for computer-based settings (Pardo, 2018; Van der Kleij, Feskens, & Eggen, 2015). Attempts have been made to capture feedback in all its appearances in one model, such as Kluger and DeNisi's (1996) feedback intervention theory and Pardo's (2018) model for data-supported feedback. These models have in common that there is a large variety in feedback effects – positive and negative – on student learning and motivation across feedback operationalizations. Factors that influence feedback effects include not only feedback design, but also the instructional context and the learners involved (Narciss & Huth, 2004). The overall tendency in the research literature is that feedback, both in general and in computer-based learning environments, may contribute to learning (Van der Kleij et al., 2015). It is, therefore, not surprising that feedback provided by computer-based learning environments has become widespread in education (Gikandi, Morrow, & Davis, 2011).

Computer-based learning environments that provide sophisticated individualized feedback are called Intelligent tutoring systems

(ITSs). In ITSs, two general feedback types can be distinguished: inner loop feedback on stepwise solutions for single tasks, and outer loop feedback over complete tasks or multiple tasks at once (Santos & Jorge, 2013; VanLehn, 2006). Inner loop feedback typically provides information about the correctness of a (partial) solution, combined with guidance on how to resolve mistakes and how to proceed in solving the current task. According to VanLehn (2006), availability of inner loop feedback classifies a computer-based learning environment as an ITS. Outer loop feedback concerns the student's current knowledge state regarding the domain and the selection of appropriate subsequent tasks or study activities. For both types, positive effects on student learning have been reported (see, e.g., VanLehn (2011) for inner loop feedback and Bull and Kay (2016) for outer loop feedback). As a consequence, both types of feedback have been implemented in computer-based learning environments that are used in educational practice today.

Implementing research findings in educational practice, however, is not straightforward (Vanderlinde & van Braak, 2010). To enable drawing causal inferences, the research community puts great emphasis on randomized experiments and controlling context variables (Farley-Ripple, May, Karpyn, Tilley, & McDonough, 2018). Consequently, many studies focus on only one of the feedback types (inner loop or outer loop feedback) and only on specific aspects (Narciss et al., 2014). In contrast, teachers and educational designers are inclined to use multiple promising approaches for delivering rich, inspiring education. As a consequence, many ITSs provide inner and outer loop feedback simultaneously, thus offering guidance both at the level of constructing step-by-step solutions to tasks as well as at the level of task selection. Ideally, this results in optimal guidance to students during their engagement with the ITS; it might, however, also lead to an overwhelming amount of feedback information for students. To our knowledge, this question of whether combining inner and outer loop feedback influences their effects has not been studied yet.

The aim of this study is, therefore, to assess the effects of offering inner and outer loop feedback concurrently. To this end, an ITS providing both inner and outer loop feedback was offered to students in social sciences bachelor programs, in a large enrolment first-year statistics course. The topic of statistics was deemed suitable for providing ITS feedback, because statistics courses are challenging for many students (Tishkovskaya & Lancaster, 2012). Students struggle to understand the large number of complex concepts involved, such as sampling variability, probability distributions and *p*-values (Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007). A particularly important and challenging topic is the method of null hypothesis significance testing (further referred to as 'hypothesis testing'), which does not only require understanding of these complex concepts, but also an ability to follow a complex line of reasoning involving uncertainty (Falk & Greenbaum, 1995; Garfield et al., 2008). Besides this challenging character of the topic, a second reason for implementing and evaluating ITS feedback in such a course was the large group size, which makes providing individual guidance and feedback difficult for the teachers involved. The guiding research question for this evaluation was: What effects does providing both inner and outer loop feedback on online homework have on students' learning process and course performance in a university statistics course?

2 | WHAT IS FEEDBACK?

Let us first take a closer look at how existing feedback theories postulate the effects of feedback. Pardo (2018), after reviewing feedback literature, defined feedback as follows:

> A process to positively influence how students engage with their work in a learning experience so that they can improve its overall quality with respect to an appropriate reference and increase their self-evaluative capacity. (Pardo, 2018, p. 433)

Four elements of this definition are worth highlighting when considering feedback effects. First, the phrases 'positively influence' and 'improve its overall quality' emphasize the general aim of enhancing the learning process. Second, and more important, the word 'process' signifies that feedback entails much more than instantaneous information delivery to a student. This also becomes clear from Pardo's model for data-supported feedback, depicted in Figure 1: Information & Delivery comprises only one of 10 components in the feedback process. It is preceded by a phase of collecting evidence about the student's learning process (nodes 5 and 7), which is influenced by factors including the student's knowledge, goals and strategies (nodes 2, 3, 4 and 6). This evidence allows for tailoring the feedback information to the student's individual needs (Gikandi et al., 2011). The collected evidence is then analysed (nodes 8 and 9), before the feedback information is delivered to the student (node 10). Next, another important phase of the feedback process takes place: the student assimilates the information, which may result in changes in the student's knowledge, beliefs, attitudes, goals and/or strategies and tactics (Pardo, 2018; Timmers, Braber-van den Broek, & Van den Berg, 2013).

The third element of Pardo's feedback definition that is worth highlighting is the phrase 'with respect to an appropriate reference'. Feedback information can only impact learning if the student knows which goals or standards to strive for. This is referred to as the feedback-standard gap (Kluger & DeNisi, 1996): feedback signals whether there is a gap between the student's current state and a desired state. If the student's current state is not up to the level of the desired state, the feedback should evoke a desire to close or reduce this gap. Hence, effective feedback should help students clarify what good performance is, in terms of desired goals or standards, and provide opportunities to close the gap between current and desired performance (Nicol & Macfarlane-Dick, 2006).

The fourth and final phrase from Pardo's feedback definition that we highlight is 'increase their self-evaluative capacity'. A characteristic of higher education, where our study is situated, is that students are expected to work independently on learning activities, especially in large enrolment courses such as first-year statistics courses. Because

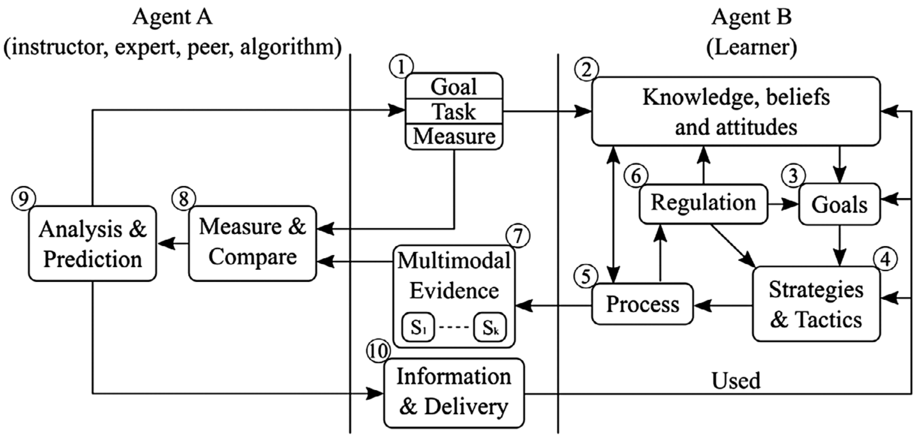


FIGURE 1 Pardo's (2018) model for data-supported feedback. Reprinted with permission of the publisher (Taylor & Francis, http://www.tandfonline.com)

of this characteristic and of the goal of higher education to prepare students for professional careers, feedback in higher education should facilitate the development of reflection on learning (Nicol & Macfarlane-Dick, 2006).

To summarize, from Pardo's feedback model and related literature we take that (a) feedback is a process, including phases of evidence collection and analysis, information delivery, and students' use of the feedback; (b) feedback should address the gap between current and desired performance and offer opportunities to close this feedback-standard gap; and (c) feedback should facilitate students' reflection on learning. In the following two sections, we outline how these feedback principles informed the two feedback implementations in this study: a domain reasoner for hypothesis testing as elaborate inner loop feedback and inspectable student models as outer loop feedback. We explicitly address these principles in the design description, which makes this study not only an evaluation of offering the combination of the two feedback types, but also of the usefulness of these guiding theory-based principles for designing feedback.

3 | INNER LOOP FEEDBACK: DOMAIN REASONERS

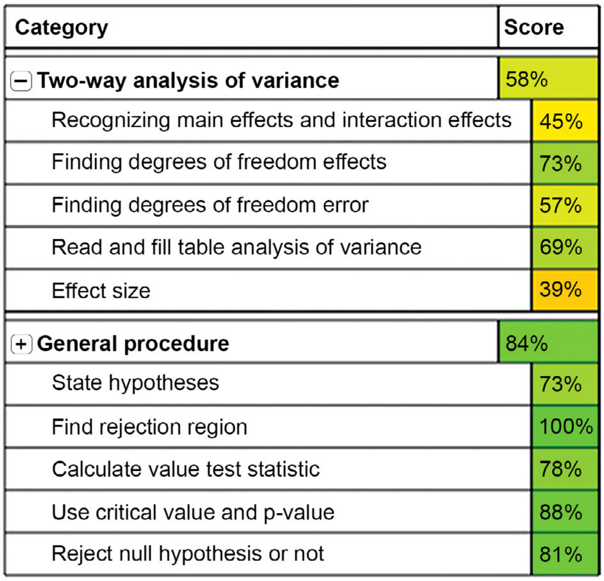
Inner loop feedback is feedback on intermediate steps in the solution to a task (VanLehn, 2006). It provides information about the correctness of the student's solution to the task so far. The evidence needed to generate this information consists of the student's steps in solving tasks. To analyse this evidence, the ITS needs to have domain knowledge: knowledge of the rules required to solve tasks in the domain at stake. In this study, the domain at stake is the topic of hypothesis testing. The component of the ITS that deals with this domain knowledge is referred to as *domain reasoner* (Goguadze, 2011). Two prevailing paradigms for the design of domain reasoners are model-tracing, in which the domain reasoner checks whether the student's solution so far follows the rules of a model solution (Anderson, Corbett, Koedinger, & Pelletier, 1995), and constraint-based modelling, in which the domain reasoner evaluates whether the student's solution violates one or more predefined constraints (Mitrovic, Martin, & Suraweera, 2007). For the topic of hypothesis testing, domain reasoners have been designed following each of the paradigms (Kodaganallur, Weitz, & Rosenthal, 2005) and even combining both (Tacoma, Heeren, Jeuring, & Drijvers, 2019).

Information delivery typically happens after each intermediate step a student takes in solving a task. If the student's solution path is correct but not complete yet, the feedback generally acknowledges this and encourages the student to continue solving the task. This encouragement implicitly signifies the incompleteness of the solution, which is a gap between current and desired performance. For incorrect solution paths, the feedback usually addresses this gap more directly, by providing information about the error and ways to repair it (Zakharov, Mitrovic, & Ohlsson, 2005). Van der Kleij et al. (2015) found that providing elaborate information is generally more effective for learning than just providing information about the correctness of the response, but noted that the type of elaborate information provided in different studies varied widely. According to Evans (2013) and Zakharov et al. (2005), such information should enable students to gain insights about underlying concepts, without explicitly dictating what the next step or these insights should be. Ideally, students would use such information to reflect on their current understanding of the concepts involved and to improve their solution to the task (VanLehn, 2011). In his review, VanLehn found that ITSs providing feedback at the level of intermediate steps are as effective as human tutors, with a mean effect size of $d = 0.76$ compared to no tutoring. Other research findings were less optimistic: for example, Narciss et al. (2014) found that students more often gave up on tasks when feedback elaborately addressed key concepts. In the domain reasoner feedback designed for the current study, therefore, we strived for providing enough, but not too much information, by briefly mentioning key concepts in hypothesis testing and relevant relations between them.

4 | OUTER LOOP FEEDBACK: INSPECTABLE STUDENT MODELS

According to VanLehn (2006), the main concern of the outer feedback loop of an ITS is the sequencing of tasks. The most rigid way of sequencing tasks is offering them in a predefined, fixed order. In this case, no actual feedback is involved. Alternatively, the ITS could analyse evidence from prior work to estimate the student's current knowledge state and use this information to select appropriate tasks. This estimation of student knowledge is called a *student model* (Brusilovsky & Millán, 2007). Such informed task selection could be interpreted as feedback, albeit rather implicitly: after analysing evidence, the ITS provides the information: 'I think that this problem is appropriate for you right now'. Thus, it provides opportunities to close the feedback-standard gap, but without explicitly indicating what this gap is.

A third alternative for task sequencing is to offer a set of tasks and let the student decide. In this case, the student model that informed task selection in the previous alternative can be used as explicit feedback to aid students in the task selection process. When the student model is visualized and delivered to the student, it is called an *inspectable student model* (Bull & Kay, 2016). In its most popular form, the student model represents the student's knowledge state as a subset of expert-level knowledge of the domain (Brusilovsky & Millán, 2007). It allows the student to quickly identify a feedback-standard gap as those *knowledge components* – elements of knowledge in the domain under study – for which the current knowledge state is below the expert-level. An example of an inspectable student model that was used in this study is shown in Figure 2. Especially the low percentages for the elements 'Recognizing main effects and interaction effects' and 'Effect size' indicate a feedback-standard gap. Based on this information, a student could decide to work on more tasks addressing these topics. Besides this function of informing task selection, the information can also influence the student's knowledge, beliefs and attitudes – for example, when the student believes to understand an element well, but finds a low percentage in the student model. In this way, inspectable student models may promote reflection (Bull & Kay, 2016; Long & Aleven, 2011). Finally, inspectable student models could provide opportunities to close the feedback-standard gap, for example by suggesting appropriate follow-up tasks (e.g., Sosnovsky & Brusilovsky, 2015).



Category	Score
⊟ Two-way analysis of variance	58%
Recognizing main effects and interaction effects	45%
Finding degrees of freedom effects	73%
Finding degrees of freedom error	57%
Read and fill table analysis of variance	69%
Effect size	39%
⊞ General procedure	84%
State hypotheses	73%
Find rejection region	100%
Calculate value test statistic	78%
Use critical value and p-value	88%
Reject null hypothesis or not	81%

FIGURE 2 Inspectable student model evaluated in this study [Colour figure can be viewed at wileyonlinelibrary.com]

Although the ideas behind inspectable student models have become common (Bull & Kay, 2016), questions such as how to collect evidence and which information to provide to which students still receive considerable research attention. Sosnovsky and Brusilovsky (2015) found that using broader topics for collecting evidence, rather than very detailed ones, results in less accurate models, but still provides a basis for successful personalization. For educational practice this is a promising result, given time constraints for teachers to develop detailed student models. Regarding information to provide, Al-Shanfari, Epp, and Baber (2017) found that students who were presented with more details and information about uncertainty in their student models, viewed their student models more often and worked on more tasks. Finally, several studies suggest that low-achieving students benefit more from availability of student models than high-achieving students, especially when student models include a social component (Brusilovsky, Somyürek, Guerra, Hosseini, & Zadorozhny, 2015) or when they are combined with support for task selection (Mitrovic & Martin, 2007). In the current study, we focus on the effects of offering student models without such additional support, this being the most common implementation in educational practice.

5 | MATERIALS AND METHODS

5.1 | Participants

The study was carried out within a first-year statistics course at a Dutch research university. This course was mandatory for all students who enrolled in any social science bachelor program. Participants in this study were 521 out of the cohort of 1,294 students who met all inclusion criteria as described in Section 5.3. The majority of students, 82%, were female, and their ages varied between 17 and 28 years (*M* = 19 years and 11 months, *SD* = 1 year and 5 months).

Students took the 8-week Methods and Statistics course after an Introductory Methods and Statistics course. The course was offered in three variants, for three different groups of students:

- Educational sciences (EDU, 94 students included out of the 186 enrolled);
- General social sciences, Cultural anthropology and Sociology (GCS, 150 students included out of the 390 enrolled);
- Psychology (PSY, 277 students included out of the 718 enrolled).

5.2 | Educational setting

The three course variants covered the same content. In five weeks of the course, students received online homework sets on statistical topics: correlation, regression, one-way ANOVA, two-way ANOVA and Chi-square tests. These homework sets contained 5–10 tasks, each with 1–8 sub-tasks. The homework sets were delivered through the Digital Mathematics Environment, a computer-based learning environment developed by the Freudenthal Institute (Drijvers, Boon, Doorman, Bokhove, & Tacoma, 2013). The DME is a widespread learning environment for mathematics education in the Netherlands, used by approximately 80 secondary schools and higher education institutions. Participants already had experience with the DME from their introductory Methods and Statistics course. The five homework sets had been used in the previous academic year and were only slightly adjusted. One of the DME's standard features, common in computer-based learning environments, is immediate verification feedback. For all tasks, students received feedback on the correctness of their solution immediately after answering. For incorrect solutions, no information was given about the correct solution, but students could attempt answering tasks until their solution was correct.

The course concluded with an exam that lasted for 2 hr and consisted of 30 four-option multiple-choice items: 15 on methods and 15 on statistics. In this study, only the students' results on the statistics items were used. The exams for the GCS and PSY groups were identical. The EDU exam was offered later in the academic year and hence was different. Typical exam items, for example, provided partial SPSS-output of a statistical test for a given context, and students had to choose the correct null hypothesis, or the correct estimation of the *p*-value, given the test value and degrees of freedom.

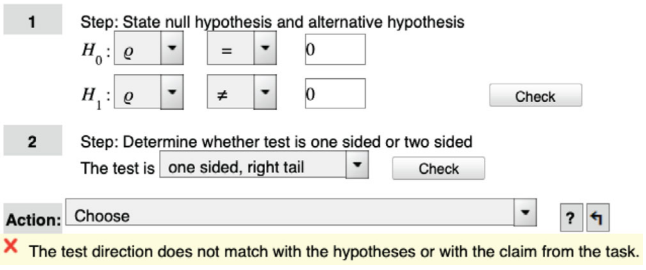
5.3 | Feedback design

In this study, two additional feedback types were implemented in the DME and evaluated: inner loop feedback on three hypothesis-testing tasks by means of a domain reasoner and outer loop feedback in the form of inspectable student models. By adding these two feedback types, the DME became an ITS as defined by VanLehn (2006). To facilitate evaluating the effects of both feedback types separately and in combination, four versions of the homework sets were designed: providing both domain reasoner feedback and student model feedback, domain reasoner feedback only, student model feedback only and none of the designed feedback types, respectively. As mentioned in Section 5.2, all four versions provided immediate verification feedback on the correctness of responses to tasks.

5.3.1 | Design of inner loop feedback

Because hypothesis testing plays an important role in statistics, the homework sets contained many tasks that involved hypothesis testing. Most of these tasks were quite structured, paving the way for students to smoothly solve them and become familiar with the many abstract concepts that play a role in hypothesis testing. In particular, most tasks asked students to carry out a single step or a predefined sequence of steps in the hypothesis-testing procedure. As argued before, however, offering only such highly structured tasks may reduce the need for students to think for themselves (Evans, 2013). Therefore, in three of the five homework sets one highly structured hypothesis-testing task was replaced by an open-ended version, in which students could stepwise set up a hypothesis test for a given research context. To do so, they could select general steps ('State hypotheses', 'Calculate the test statistic', etc.) from a drop-down menu. After selecting a step, the student could complete it with specific details for the current task. For these three hypothesis-testing tasks, domain reasoner feedback was available. In the versions of the homework sets with domain reasoner feedback, students received feedback on the correctness of their solution so far, each time they completed a step. If the partial solution contained errors, the feedback provided information on how to resolve these errors. Furthermore, students could ask for a hint suggesting an appropriate next step. The versions of the homework sets without domain reasoner feedback did not offer such elaborate services: they only provided verification feedback on each single step, without further elaboration and without considering previous steps the student had taken. As such, the versions without domain reasoner feedback provided minimal inner loop feedback, while the domain reasoner provided elaborate inner loop feedback.

The evidence that the domain reasoner collected consisted of all steps the student had taken in the hypothesis-testing procedure so far. In the analysis phase, the domain reasoner checked whether the student was on a correct solution path and, if not, diagnosed which parts of the solution were inconsistent, incomplete or incorrect. For full details of the domain reasoner design, we refer to Tacoma et al. (2019). Figure 3 shows an example of information that was provided by the domain reasoner. For comparison, Figure 4 shows the feedback that was given for the same partial solution in the versions of the homework sets that provided minimal feedback. The domain reasoner feedback was intended to facilitate improvement of the current solution, hence providing opportunities to reduce the feedback-standard gap. Also, we tried to provide just enough information, allowing the students to think of some of the required steps by themselves and hence to critically reflect on and expand their current knowledge of hypothesis testing. In other words, the feedback pinpointed inconsistencies in student solutions and mentioned key concepts related to these inconsistencies, but did not explicitly describe how these inconsistencies could be resolved. By considering



1 Step: State null hypothesis and alternative hypothesis

H_0 : ϱ = 0

H_1 : ϱ ≠ 0

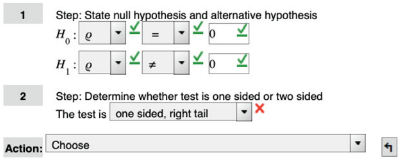
2 Step: Determine whether test is one sided or two sided

The test is one sided, right tail

Action: Choose

✗ The test direction does not match with the hypotheses or with the claim from the task.

FIGURE 3 Domain reasoner feedback for an incorrect test direction [Colour figure can be viewed at wileyonlinelibrary.com]



1 Step: State null hypothesis and alternative hypothesis

H_0 : ϱ ✓ = ✓ 0 ✓

H_1 : ϱ ✓ ≠ ✓ 0 ✓

2 Step: Determine whether test is one sided or two sided

The test is one sided, right tail ✗

Action: Choose

FIGURE 4 Verification feedback for an incorrect test direction [Colour figure can be viewed at wileyonlinelibrary.com]

the student's entire solution so far, the domain reasoner forced students to add several essential steps for hypothesis-testing, and to add these steps in a logical line of reasoning. This was not the case in the versions of the homework sets with minimal feedback: here students could enter steps in any order they liked and could omit as many steps as they liked. Besides checking the correctness of partial solutions, the domain reasoner could also provide hints for appropriate next steps. Students could ask for a hint at any time while solving the tasks. Hints were formulated in general terms ('State hypotheses' or 'Calculate the value of the test statistic'), so that the students still needed to fill in the details for the current situation.

Before the study, the domain reasoner had been implemented and evaluated in a course for psychology students (Tacoma et al., 2019). As a consequence, students in the PSY course variant already had experience with the domain reasoner feedback, while students in the EDU and GCS variants did not. For this study, improvements were made based on the previous evaluation and the domain reasoner software was extended to support tests for correlation, ANOVA and Chi-square.

5.3.2 | Design of outer loop feedback

Each homework set had a student model, containing 9–14 knowledge components divided into two or three categories. When students opened the student model, all their attempts at answering tasks in the homework set so far were used to generate the student model. An example of a student model that was used is shown in Figure 2. The colouring served to help students identify where the feedback-standard gap was the largest. No explicit standards were communicated to the students (e.g., 'keep practicing until you reached 80% for all knowledge components').

Students could access the student models at all times while working on the homework tasks, and as many times as they liked. Furthermore, the final page of the homework sets explicitly mentioned the student models. Hence, after students had worked on all tasks of that week's homework set, they were encouraged to use the student model to decide which topics they still needed to work on. To promote reflection, this page also contained more detailed descriptions of the knowledge components in the student models. After viewing the student model, students were free to choose whether and how to proceed their practice session. All homework tasks were optional and students could attempt tasks as many times as they liked. Allowing such resubmission offers students important opportunities to use the feedback from the student models in their learning process (Nicol & Macfarlane-Dick, 2006). Although tasks were not annotated with the knowledge components they contributed to, from the context it was often clear to which knowledge components they belonged. Students could also choose to (re)read topics in the textbook and to work on textbook tasks for topics that needed their attention.

The student model design was evaluated in two earlier courses (Tacoma, Sosnovsky, Boon, Jeuring, & Drijvers, 2018). Participants in the current study had been enrolled in one of these earlier courses and, consequently, were already familiar with inspectable student models.

5.4 | Study design

Students were randomly assigned to one of the four conditions (domain reasoner and student models available, only domain reasoner available, minimal feedback and student models available, and only minimal feedback available). Randomization took place within the three course variants, ensuring approximately equal group sizes within all three variants. Students were only included as participants if they met all of the following criteria:

- They gave active consent for use of their DME work and exam results for this study (618 students excluded)
- They worked on the homework sets for at least 1 hr, including breaks of up to 5 min (55 students excluded)
- They worked on at least one task with stepwise construction of hypothesis tests (three students excluded)
- Their exam results for both the current course and the previous course were available (94 students excluded)
- They worked in only one condition (three students excluded)

These inclusion criteria resulted in 521 students being included. Table 1 summarizes the number of students in each course variant and each experimental condition. Age and gender distribution did not differ significantly between conditions (F(3,517) = 0.24, p = .868 for age and χ^2(3) = 0.72, p = .869 for gender).

Data consisted of the students' work in the DME, exam results in the course and exam results in the previous course. DME work included all attempts at all tasks, as well as information about student model viewing. After anonymization, these DME log data were used to calculate the students' time-on-task in the DME, the number of open-ended hypothesis-testing tasks students correctly solved and the number of student model views, for those students who had student models available. From both exams, only the results on the statistics items were used. For the exam of the current course, values of Cronbach's α for the 15 statistics items were .61 for the GCS and PSY variant and .56 for the EDU variant. For the regular exam of the previous course, Cronbach's α for the 16 statistics items was similar, namely .59. Although these values are not high, they seem reasonable for exams that assess a wide variety of topics within the domain of statistics (e.g., choice of statistical test, stating hypotheses,

TABLE 1 Numbers of students in the three course variants and four experimental conditions

	SM			No SM			All		
Course variant	DR	No DR	All	DR	No DR	All	DR	No DR	All
EDU	25	30	55	20	19	39	45	49	94
GCS	41	34	75	38	37	75	79	71	150
PSY	80	81	161	55	61	116	135	142	277
Total	146	145	291	113	117	230	259	262	521

Abbreviations: DR, domain reasoner feedback available; SM, student model feedback available.

TABLE 2 Outcome variables and potential predictor variables for the four multiple regression models

Model no.	Outcome variable	Description outcome variable	Extra predictor variables[a]
1	Time-on-task	Total time (in hours) students worked in the DME, including breaks of up to 5 min	Prior performance Age Gender
2	Student model views	Number of times students viewed their student models	Prior performance Age Gender Time-on-task[b]
3	Hypothesis-testing score	Number of hypothesis-testing tasks (out of 3) in the DME for which students gave a complete correct solution	Course variant Prior performance Age Gender
4	Exam result	Score on statistics items in the exam (out of 15 items)	Course variant Prior performance Age Gender Time-on-task[b]

[a] Variables domain reasoner and student model and their interaction were always used as predictor variables, except for Model 2. In Model 2, only students in the student model conditions were taken into account and the variable domain reasoner was included.

[b] Time-on-task was only included as predictor variable if no strong relationship between time-on-task and condition variables would be found in Model 1.

interpreting test output) with relatively few items (Taber, 2018). The exam results of the previous course were used as a measure of prior performance. The average of this prior performance score over all students was 11.3 points (*SD* = 2.5) and prior performance did not differ significantly between the four experimental conditions (*F* (3, 517) = 0.28, *p* = .839).

5.5 | Data analysis

Four outcome variables were used to describe the students' learning processes and performance: time-on-task, student model views, hypothesis-testing score and exam result. Descriptions of these outcome variables are given in Table 2. For each of these four outcome variables, a multiple linear regression model was created, using the experimental conditions and several other variables as predictors.

The outcome variable in Model 1, time-on-task, was measured as the number of hours that students worked in the DME and was expected to possibly be influenced by student model availability: students with student models available were expected to reflect more on their learning and hence, possibly, to choose to work on more tasks. At the same time, however, time-on-task was also expected to be largely determined by student characteristics, such as diligence and motivation, which were not directly measured in this experiment. To enable taking these student characteristics into account in later models, it was, therefore, desirable to include time-on-task as independent variable in these models. Evidently, this could only be done if no strong relationship between experimental conditions and time-on-task would be found in Model 1, because otherwise time-on-task would not be an independent variable. Hence, Model 1 served to shed light on the relation between time-on-task and the experimental conditions, with the extra goal to assess whether time-on-task could be used as independent variable in later models.

Model 2, concerning the number of student model views, only included the students who had student models available. It allowed us to explore whether students with and without domain reasoner feedback used the student models differently. Model 3 concerned hypothesis-testing score: the number of hypothesis-testing tasks in the DME that students solved completely, out of the three stepwise hypothesis-testing tasks. This score was expected to be mainly influenced by domain reasoner availability. Finally, the outcome variable in Model 4, exam result, concerned the students' score on the 15 statistics items in the exam and was expected to be influenced by both feedback conditions. Besides the two experimental conditions, five other variables were deemed important: prior course performance (score between 0 and 16), course variant (EDU, GCS, PSY), age (in years), gender, and time-on-task (if Model 1 would not yield strong dependence of time-on-task on experimental conditions). The most widespread method to take such covariates into account when evaluating experimental conditions is an ANCOVA, but this method does not account for potential effects of interactions between experimental conditions and covariates on the outcome variables. Since we were especially interested in such potential interaction effects – for example, to investigate whether the effect of student model availability on exam results was different for students with different prior performance scores – we opted for the more general approach of multiple linear regression.

For each of the four outcome variables, we started with a model including all predictor variables and interactions that were judged to be relevant. Table 2 summarizes the variables used at the start of creating each model. As argued above, time-on-task would only be included as predictor variable in Models 2 and 4 if no strong relationship between time-on-task and the experimental conditions would be found in Model 1. The predictor variables age, prior performance and time-on-task were always added centred to their mean. Next, step-by-step, non-significant interactions and predictors were removed from the model. In this phase, the experimental conditions and their interaction were retained in the models, regardless of their significance. Once a model was obtained in which all predictors apart from the experimental conditions were significant, this model was regarded as the complete model for that outcome variable. Because of the large influence that outliers can have on model parameters, outliers were removed and the model was fitted again, until no more outliers were found. The normality assumption of residual distribution and the assumption of homoscedasticity were checked as well. Next, to assess the influence of the experimental conditions and their interactions with predictor variables and with each other, these were removed from the complete model one by one. After each removal, the value of R^2 for the new model was calculated as measure of effect size, as well as an *F*-ratio for the comparison of the models before and after the removal.

6 | RESULTS

Table 3 summarizes the means and standard deviations of the outcome variables for all experimental conditions. In addition to the information in the table, it is worth mentioning that the students attempted on average 31 out of 34 tasks (*SD* = 5) and that the students with student models available viewed their student models for on average 48 s (*SD* = 63 s). Furthermore, students with domain reasoner feedback available on average requested approximately one hint per task. The results of the regression analyses are presented in the following four sections.

6.1 | Time-on-task

The parameter estimates of the regression model predicting time-on-task are summarized in Table 4. One outlier was removed. The complete model explained 8% of the variation in students' time-on-task. It showed no significant effects of either of the experimental conditions, nor of their interaction, on time-on-task in the DME. It did show a significant interaction effect between age and student model availability, suggesting that older students with student models available tended to work longer than younger students with student models available. This interaction effect accounted for 1.6% of the variance, a small but significant contribution to the model (*F*[1, 512] = 9.15, *p* = .003). After removing the interaction between domain reasoner and student model availability, the availability of student models had a significant positive effect on students' time-on-task. This effect is also reflected in the average time-on-task reported in Table 3, which was 6.1 hr for students with student models and 5.6 hr for students without. Removing student model availability from the model showed that it explained 0.9% of the variability in time-on-task, a small but significant contribution (*F*[1, 514] = 4.62, *p* = .032). Since this relationship was only weak, it was deemed justifiable to include time-on-task as predictor variable in later regression models: these results imply that time-on-task seemed to be mainly determined by other factors than student model or domain reasoner availability.

TABLE 3 Overview of outcome variables for students in the four experimental conditions

		SM			No SM			All		
		DR	No DR	All	DR	No DR	All	DR	No DR	All
Outcome variable	*n*	146	145	291	113	117	230	259	262	521
Time-on-task	Mean	6.0	6.1	6.1	5.6	5.7	5.6	5.8	5.9	5.9
	SD	2.5	2.5	2.5	2.3	2.4	2.4	2.5	2.4	2.4
Student model views	Mean	5.4	6.3	5.9	–	–	–	–	–	–
	SD	4.4	5.3	4.9	–	–	–	–	–	–
Hypothesis-testing score	Mean	1.1	1.1	1.1	1.2	1.0	1.1	1.2	1.1	1.1
	SD	1.1	1.0	1.1	1.0	1.0	1.0	1.0	1.0	1.0
Exam result	Mean	11.0	11.2	11.1	11.3	11.1	11.2	11.1	11.1	11.1
	SD	2.5	2.3	2.4	2.5	2.4	2.4	2.5	2.3	2.4

Abbreviations: DR, domain reasoner feedback available; SM, student model feedback available.

TABLE 4 Parameter estimates and model fits for Model 1: time-on-task predicted by experimental conditions, prior performance, age, gender and interactions

	Complete model	Interaction student model age removed	Interaction domain reasoner student model removed	Student model removed	Domain reasoner removed
Intercept	4.87***	4.83***	4.87***	5.13***	5.10***
Prior performance	0.13**	0.12**	0.12**	0.12**	0.12**
Age	0.08	0.32***	0.32***	0.32***	0.32***
Gender female	0.91**	0.94***	0.94***	0.93***	0.93***
Domain reasoner available	0.00	0.02	−0.06	−0.05	
Student model available	0.49	0.51	0.44*		
Student model × age	0.43**				
Domain reasoner × student model	−0.10	−0.13			
R^2	.080	.064	.064	.056	.055
R^2 change		.016	.000	.008	.000
F change		9.15**	0.11	4.62*	0.06

Note: *p < .05; **p < .01; ***p < .001.

TABLE 5 Parameter estimates and model fits for Model 2: number of student model views predicted by domain reasoner availability, time-on-task and their interaction

	Complete model	Interaction removed	Domain reasoner removed
Intercept	5.84***	5.94***	5.58***
Time-on-task	1.22***	0.89***	0.89***
Domain reasoner available	−0.56	−0.73	
Domain reasoner × time-on-task	−0.65**		
R^2	.246	.212	.212
R^2 change		.034	.000
F change		10.63**	2.17

Note: *p < .05; **p < .01; ***p < .001.

6.2 | Student model views

The model predicting the number of student model views included the predictor time-on-task and its interaction with domain reasoner availability. Prior performance, age and gender were found not to be significant predictors. The parameter estimates and model fits for the complete model are given in Table 5. One outlier was removed and the model showed some heteroscedasticity: the number of student model views varied widely for high values of time-on-task and far less for low values of time-on-task.

The complete model explained 25% of the variance in number of student model views. It should be noted that this might be a slight overestimation, because of the above-mentioned heteroscedasticity. The model shows, not surprisingly, that students who worked more in the DME also tended to view their student models more often. Removing the interaction between domain reasoner and time-on-task from the model showed that this interaction term accounted for 3.4% of the variance and contributed significantly to the model (F[1, 287] = 10.63, p = .001). This means that while there was no main effect of domain reasoner availability, there was an interaction effect, which is illustrated in Figure 5. Of the students who used the DME intensively, those with domain reasoner feedback available viewed their student models less often than their peers who did not receive domain reasoner feedback.

6.3 | Hypothesis-testing score

Since students in the PSY variant and domain reasoner conditions already had previous experience with the domain reasoner, whereas students in the other two course variants had not, an interaction effect between the domain reasoner condition and course variant was included in the regression model predicting students' hypothesis-testing score. The model's parameter estimates are summarized in Table 6. No outliers were detected, but the normality assumption of residual distribution was violated. This was not regarded a problem, because of the large sample size of 521 students.

The complete model explained 9% of the variance in hypothesis-testing score. Consecutively removing interactions and the experimental conditions revealed that the interaction between domain reasoner availability and course variant explained 3.4% of the variance, which was a significant contribution to the model ($F[2, 513] = 9.61$, $p < .001$). The significant negative effect for domain reasoner availability in the complete model indicates that students in the baseline course variant, GCS, seemed to perform better without than with domain reasoner feedback. Meanwhile, the significant effect of the interaction between domain reasoner availability and course variant reveals that the domain reasoner's effectiveness was different for the different groups of students. To further investigate this interaction, Table 7 summarizes hypothesis-testing scores for students with and without domain reasoner feedback within the three course variants. The *t* test results confirm that domain reasoner feedback hindered student performance in the GCS variant, while students in the PSY course variant performed significantly better when domain reasoner feedback was available to them. No significant effect was found for the EDU group.

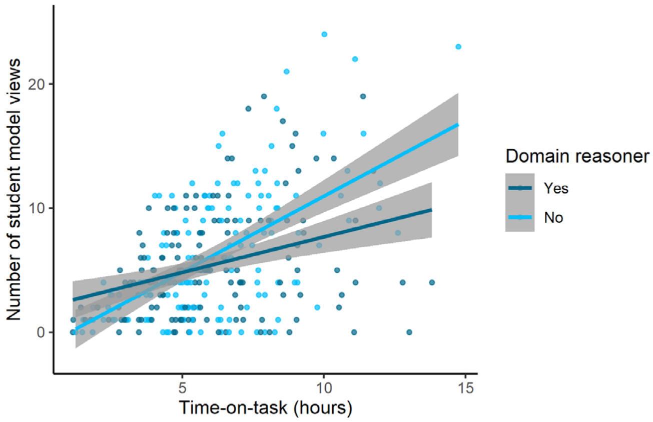


FIGURE 5 Effect of the interaction between time-on-task and domain reasoner availability on number of student model views [Colour figure can be viewed at wileyonlinelibrary.com]

6.4 | Exam result

The parameter estimates of the model predicting exam result from the experimental conditions, prior performance and time-on-task are given in Table 8. Course variant, age and gender were no significant predictors of exam result and one outlier was removed.

The complete model explained 24% of the variance and showed that prior performance and time-on-task significantly affected exam result. Furthermore, it revealed an interaction effect between student model availability and prior performance. This interaction effect explained 0.7% of the variance in exam result and contributed significantly to the model ($F[1, 514] = 4.63$, $p = .032$). Student model availability, domain reasoner availability and their interaction did not contribute significantly to the model. Figure 6 illustrates the interaction effect between student model availability and prior performance. For low prior performance scores, the regression line for students with student models is higher than the one for students without student models, meaning that students with low prior performance

TABLE 6 Parameter estimates and model fits for Model 3: hypothesis-testing score predicted by experimental conditions, course variant, their interactions and prior performance

	Complete model	Interaction DR course variants removed	Interaction SM DR removed	DR removed	SM removed
Intercept	0.80***	0.52***	0.56***	0.60***	0.59***
Prior performance	0.04*	0.03	0.03	0.03	0.03
Gender female	0.31**	0.30*	0.30*	0.30**	0.30**
Variant EDU	0.05	0.37**	0.37**	0.37**	0.37**
Variant PSY	−0.04	0.41***	0.42***	0.41***	0.41***
Domain reasoner available	−0.38*	0.17	0.08		
Student model available	0.10	0.07	−0.02	−0.02	
Domain reasoner × EDU	0.61*				
Domain reasoner × PSY	0.89***				
Domain reasoner × student model	−0.23	−0.17			
R^2	.091	.057	.055	.054	.054
R^2 change		.034	.002	.001	.000
F change		9.61***	0.98	0.75	0.04

Note: *$p < .05$; **$p < .01$; ***$p < .001$.
Abbreviations: DR, domain reasoner feedback available; SM, student model feedback available.

TABLE 7 Hypothesis-testing score by course variant

Course variant	DR	No DR	*t*	*df*	*p*	Cohen's *d*
EDU (*n* = 94)	1.3 (0.9)	1.2 (1.1)	0.42	92	.678	—
GCS (*n* = 150)	0.6 (0.8)	1.1 (1.0)	−3.26	148	.001*	0.53
PSY (*n* = 277)	1.4 (1.1)	1.1 (1.0)	2.95	275	.003*	0.35

Abbreviation: DR, domain reasoner feedback available.

TABLE 8 Parameter estimates and model fits for Model 4: exam result predicted by experimental conditions, prior performance, time-on-task and interactions

	Complete model	Interaction SM prior performance removed	Interaction SM DR removed	SM removed	DR removed
Intercept	11.17***	11.16***	11.31***	11.17***	11.13***
Prior performance	0.47***	0.40***	0.39***	0.39***	0.39***
Time-on-task	0.19***	0.19***	0.19***	0.19***	0.19***
Domain reasoner available	0.22	0.22	−0.09	−0.09	
Student model available	0.01	0.01	−0.26		
Student model × prior performance	−0.16*				
Domain reasoner × student model	−0.54	−0.54			
R^2	.244	.237	.234	.231	.231
R^2 change		.007	.003	.003	.000
F change		4.63*	2.18	1.94	0.24

Note: *$p < .05$; **$p < .01$; ***$p < .001$.

benefited from the student models. The opposite holds for students with high prior performance: for these students, the availability of student models had a negative effect on course performance.

7 | DISCUSSION

While many research studies address only inner loop or only outer loop feedback, many ITSs used in educational practice offer these two feedback types simultaneously. Therefore, the aim of this study was to assess the effects of combined inner and outer loop feedback. The guiding research question was: what effects does providing both inner and outer loop feedback on online homework have on students' learning process and course performance in a university statistics course? To answer this question, in the following section we first discuss the effects we found of both feedback types separately and next reflect on their combination.

7.1 | Effects of the feedback types and their combination

Elaborate inner loop feedback was provided in the form of a domain reasoner for tasks in which students set up hypothesis tests. Students with prior experience with the domain reasoner benefited from its feedback for solving hypothesis-testing tasks, but students without prior experience did not. This corroborates earlier findings that students may need some time to familiarize themselves with elaborate inner loop feedback (Tacoma et al., 2019). More specifically, students needed to learn which steps are essential in the hypothesis-testing procedure, because partial solutions omitting essential steps were regarded as incorrect by the domain reasoner, while they were marked correct in the conditions with minimal feedback. The effect size for students already familiar with the feedback was *d* = 0.35, which is slightly smaller than the average effect size of 0.50 that Van der Kleij and colleagues found for elaborate feedback (Van der Kleij et al., 2015). Furthermore, no direct effects were found of domain reasoner availability on time-on-task and exam result. Given the mixed effects of domain reasoner availability on students' hypothesis-testing score and the different nature of exam items (multiple choice tasks as opposed to stepwise hypothesis testing tasks), this lack of a positive effect on exam result is not surprising (Shute, 2008).

Outer loop feedback was implemented in the form of inspectable student models. Student model availability slightly influenced the time students worked in the ITS: students with student models tended to work slightly longer than students without. This effect was small: student model availability only explained 0.9% of the variance in time-on-task. This is similar to results reported by Sosnovsky and Brusilovsky (2015), who found a correlation coefficient of *r* = .13, which, squared, yields 1.7% explained variance. Student model availability did not affect hypothesis-testing scores, but did seem to affect exam results. The exam results revealed that students with low prior performance slightly benefited from the student models, while students with high prior performance were slightly hindered by them. This finding is similar to findings from studies in which extra

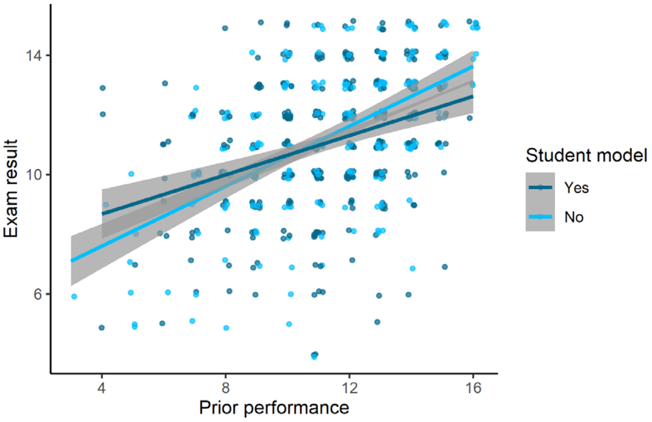


FIGURE 6 Effect of the interaction between student model availability and prior performance on exam result. For visualization purposes, some random noise was added to prior performance and exam result [Colour figure can be viewed at wileyonlinelibrary.com]

components – a social component or support for appropriate task selection – were added to the student models (Brusilovsky et al., 2015; Mitrovic & Martin, 2007).

Having established the effects of both feedback types separately, we now turn to their combination to answer our research question completely. Our findings showed no interaction effects between the feedback types concerning performance outcome variables. Hence, in this study the two feedback types did not amplify or attenuate each other's effects on students' course performance. Regarding the students' learning process, however, domain reasoner availability did influence the students' use of the student models: students who used the ITS intensively tended to view the student models less often if they also received domain reasoner feedback. This means that for the students' learning process, domain reasoner availability did attenuate students' student model use, but this did not affect the students' course performance.

7.2 | Revisiting feedback principles

How can these findings be interpreted in the light of the feedback principles we identified in Section 2? The first principle states that feedback is a process, including phases of evidence collection and analysis, information delivery and students' use of the feedback (Pardo, 2018; Timmers et al., 2013). Our results illustrate that these phases may influence each other in different ways for high- and low-achieving students. The finding that high-achieving students did not benefit from the student models could be a consequence of design choices in the evidence analysis phase: our procedure may have resulted in too optimistic estimations of high-achieving students' domain knowledge. This, in turn, may have given these students the impression that they were well prepared for the exam and did not need further practice, resulting in suboptimal use of the feedback. Meanwhile, for lower-achieving students the calibration of the estimations seems to have been more appropriate, given that student model availability had a positive, though small, effect on these students' exam results. Other subtle design choices may have influenced the feedback process as well, such as a page notifying students of the student models, the setting that students could attempt tasks as often as they liked and the exact wording of the feedback messages that the domain reasoner provided. Even when adhering to general guidelines for effective feedback, such design details can considerably influence its actual effectiveness (Kluger & DeNisi, 1996; Zakharov et al., 2005).

The second feedback principle states that feedback should address the feedback-standard gap and offer opportunities to close it. While domain reasoner feedback quite explicitly provided guidance in closing this gap by mentioning key concepts, our implementation of student models did not provide explicit suggestions about how to reduce the gap, in terms of appropriate tasks or reading material. This could explain why students with both feedback types available tended to use the student models less: the more explicit suggestions by the domain reasoner feedback may have been easier to follow than the implicit messages the student models gave them. Earlier research on student models has shown that explicit suggestions can contribute to keeping students engaged with learning material (Arroyo et al., 2007) and to help students to allocate their attention to appropriate tasks given their current knowledge (Sosnovsky & Brusilovsky, 2015). It should be noted, however, that implementing such explicit suggestions puts higher demands on course design than the approach opted for in this study.

The third feedback principle states that feedback should facilitate students' reflection on learning. Inspectable student models are valued for the opportunities for planning and reflection that they offer (Bull & Kay, 2016), but in this study only the weaker students benefited to some extent from these opportunities. Furthermore, the finding that outer loop feedback was used less by students who also received elaborate inner loop feedback suggests that the students' need or capacity for reflection is limited. Students may not have had the cognitive capacity to process both feedback types together optimally. In other words, our results may indicate a feedback ceiling effect: a maximum amount of feedback that students can process at once.

7.3 | Limitations

While revisiting the feedback principles in Section 7.2, we discussed two aspects of this study that could be regarded as limitations: the exact calibration of estimations in the student models and the absence of explicit suggestions in the student models to close the feedback-standard gap. A third limitation of this study is that both feedback interventions were rather small. On average, students worked almost 6 hr on homework tasks in the ITS, but typically spent less than 1 min viewing their student models. Likewise, only three out of the 34 tasks in the homework sets were stepwise hypothesis-testing tasks in which students could receive domain reasoner feedback. Hence, opportunities to learn from the student models and domain reasoner feedback were rather sparse. For the domain reasoner feedback this was especially true for students who had no prior experience with it and hence, presumably, needed time to familiarize themselves with its feedback (Tacoma et al., 2019). Arguably, providing the evaluated feedback types for longer time periods (one academic year, or throughout a complete study program) could lead to larger learning effects (Evans, 2013) and would be a promising direction for further research.

7.4 | Implications and recommendations

This study has a number of implications for theory and practice. First of all, the interaction effects found between feedback conditions, prior performance and prior experience illustrate that the same feedback may have different effects for different learners. This finding supports theory reflected in Pardo's model for data-supported feedback (Pardo, 2018): the learner's knowledge, beliefs and attitudes influence how feedback changes a student's strategies and learning process. For educational practice, this implies that providing multiple types of feedback, such as both inner loop and outer loop feedback, may result in more students receiving feedback that is helpful for them.

In this study, we also found indications, though, that introducing multiple feedback types at once may result in suboptimal use of the feedback, a feedback ceiling effect. Introducing feedback types one by one would therefore be recommendable. Taking this a step further, based on our results we suggest that lower-achieving students could first be provided with inspectable student models, since they seem to benefit from this outer loop feedback. Higher-achieving students could be expected to familiarize themselves more quickly with a new feedback type and, hence, could benefit more quickly from detailed inner loop feedback. We encourage further research investigating this hypothesis.

Finally, regardless of the exact implementation, our findings imply that students need time to get used to new feedback and to know how the feedback can help them learn. From a theoretical perspective, this suggests that the three feedback principles could be supplemented with a fourth regarding the amount of feedback: students should be given enough time and opportunities to familiarize themselves with and to learn from feedback. For educational practice, this implies that new feedback implementations should be offered for substantial periods of time (i.e., preferably longer than one semester) and students should be offered sufficient guidance in interpreting feedback information.

ACKNOWLEDGEMENTS

Utrecht University's Education Incentive Fund supported the design and implementation of the educational materials used in this study. The authors wish to thank Jeltje Wassenberg-Severijnen, teacher of the course, for the valuable collaboration and Wouter van Joolingen and Nathalie Kuijpers for their comments on earlier versions of this article.

CONFLICT OF INTEREST

The authors declare there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Sietske Tacoma: Conceptualization, Formal analysis, Writing – original draft. Paul Drijvers: Writing – review & editing, Supervision. Johan Jeuring: Writing – review & editing, Supervision.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Sietske Tacoma https://orcid.org/0000-0002-9662-8489

REFERENCES

Al-Shanfari, L., Epp, C. D., & Baber, C. (2017). Evaluating the effect of uncertainty visualisation in open learner models on students' metacognitive skills. In E. André, R. Baker, X. Hu, M. Rodrigo, & B. du Boulay (Eds.), *Artificial intelligence in education, AIED 2017, LNCS 10331* (pp. 15–27). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-61425-0_2

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, *4*, 167–207. https://doi.org/10.1207/s15327809jls0402_2

Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., … Woolf, B. P. (2007). Repairing disengagement with non-invasive interventions. In R. Luckin, K. Koedinger, & J. Greer (Eds.), *Proceedings of the 2007 conference of artificial intelligence in education: Building technology-rich learning contexts that work* (pp. 195–202). Amsterdam, the Netherlands: ISO Press.

Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web, LNCS 4321* (pp. 3–53). Berlin Heidelberg, Germany: Springer-Verlag.

Brusilovsky, P., Somyürek, S., Guerra, J., Hosseini, R., & Zadorozhny, V. (2015). The value of social: Comparing open student modeling and open social student modeling. In F. Ricci, K. Bontcheva, O. Conlan, & S. Lawless (Eds.), *User modeling, adaptation and personalization* (pp. 44–55). Cham, Switzerland: Springer International Publishing.

Bull, S., & Kay, J. (2016). SMILI: A framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education*, *26*, 293–331. https://doi.org/10.1007/s40593-015-0090-8

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*, 98–113. https://doi.org/10.1016/j.edurev.2007.04.001

Drijvers, P., Boon, P., Doorman, M., Bokhove, C., & Tacoma, S. (2013). Digital design: RME principles for designing online tasks. In C. Margolinas (Ed.), *Proceedings of ICMI study 22 task Design in Mathematics Education* (pp. 52–62). Clermont-Ferrand, France: ICMI.

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research*, *83*(1), 70–120. https://doi.org/10.3102/0034654312474350

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*, 75–98. https://doi.org/10.1177/0959354395051004

Farley-Ripple, E., May, H., Karpyn, A., Tilley, K., & McDonough, K. (2018). Rethinking connections between research and practice in education: A conceptual framework. *Educational Researcher*, *47*, 235–245. https://doi.org/10.3102/0013189X18761042

Garfield, J. B., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). *Learning to reason about statistical inference. Developing students' statistical reasoning* (pp. 261–288). Dordrecht, the Netherlands: Springer.

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, *57*, 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Goguadze, G. (2011). *ActiveMath - generation and reuse of interactive exercises using domain reasoners and automated tutorial strategies* (Doctoral dissertation), Saarland University, Saarbrücken, Germany. Retrieved from https://publikationen.sulb.uni-saarland.de/bitstream/20.500.11880/26153/1/goguadzeDiss2011.pdf

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

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance. *Psychological Bulletin*, *119*, 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Kodaganallur, V., Weitz, R. R., & Rosenthal, D. (2005). A comparison of model-tracing and constraint-based intelligent tutoring paradigms. *International Journal of Artificial Intelligence in Education*, *15*, 117–144.

Long, Y., & Aleven, V. (2011). Students' understanding of their student model. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education: 15th international conference* (pp. 179–186). Berlin Heidelberg, Germany: Springer-Verlag. https://doi.org/10.1007/978-3-642-21869-9_25

Mitrovic, A., & Martin, B. (2007). Evaluating the effect of open student models on self-assessment. *International Journal of Artificial Intelligence in Education*, *17*, 121–144.

Mitrovic, A., Martin, B., & Suraweera, P. (2007). Intelligent tutors for all: The constraint-based approach. *IEEE Intelligent Systems*, *22*, (4), 38–45.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. In H. M. Niegemann, D. Leutner, & R. Brünken (Eds.), *Instructional Design for Multimedia Learning* (pp. 181–195). Münster, Germany: Waxmann.

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education*, *71*, 56–76. https://doi.org/10.1016/j.compedu.2013.09.011

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*, 199–218. https://doi.org/10.1080/03075070600572090

Pardo, A. (2018). A feedback model for data-rich learning experiences. *Assessment & Evaluation in Higher Education*, *43*, 428–438. https://doi.org/10.1080/02602938.2017.1356905

Santos, G. S., & Jorge, J. (2013). Interoperable intelligent tutoring systems as open educational resources. *IEEE Transactions on Learning Technologies*, *6*, 271–282. https://doi.org/10.1109/TLT.2013.17

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

Sosnovsky, S., & Brusilovsky, P. (2015). Evaluation of topic-based adaptation and student modeling in QuizGuide. *User Modeling and User-Adapted Interaction*, *25*, 371–424. https://doi.org/10.1007/s11257-015-9164-4

Taber, K. S. (2018). The use of cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*, 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tacoma, S., Heeren, B., Jeuring, J., & Drijvers, P. (2019). Automated feedback on the structure of hypothesis tests. In U. T. Jankvist, M. van den Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Proceedings of the eleventh congress of the European Society for Research in mathematics education* (pp. 2969–2976). Utrecht, the Netherlands: Freudenthal Group & Freudenthal Institute, Utrecht University and ERME.

Tacoma, S., Sosnovsky, S., Boon, P., Jeuring, J., & Drijvers, P. (2018). The interplay between inspectable student models and didactics of statistics. *Digital Experiences in Mathematics Education*, *4*(2–3), 139–162. https://doi.org/10.1007/s40751-018-0040-9

Timmers, C. F., Braber-van den Broek, J., & Van den Berg, S. (2013). Motivational beliefs, student effort, and feedback behaviour in computer-based formative assessment. *Computers & Education*, *60*, 25–31. https://doi.org/10.1016/j.compedu.2012.07.007

Tishkovskaya, S., & Lancaster, G. A. (2012). Statistical education in the 21st century: A review of challenges, teaching innovations and strategies for reform. *Journal of Statistics Education*, *20*(2), 1–55. https://doi.org/10.1080/10691898.2012.11889641

Van der Kleij, F., Feskens, R., & Eggen, T. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes. A meta-analysis. *Review of Educational Research*, *85*, 475–511. https://doi.org/10.3102/0034654314564881

Vanderlinde, R., & van Braak, J. (2010). The gap between educational research and practice: Views of teachers, school leaders, intermediaries and researchers. *British Educational Research Journal*, *36*, 299–316. https://doi.org/10.1080/01411920902919257

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*, 227–265.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*, 197–221. https://doi.org/10.1080/00461520.2011.611369

Zakharov, K., Mitrovic, A., & Ohlsson, S. (2005). Feedback micro-engineering in EER-tutor. In C.-K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology* (pp. 718–725). Amsterdam, the Netherlands: IOS Press.

How to cite this article: Tacoma S, Drijvers P, Jeuring J. Combined inner and outer loop feedback in an intelligent tutoring system for statistics in higher education. *J Comput Assist Learn*. 2021;37:319–332. https://doi.org/10.1111/jcal.12491