Received: 28 November 2024 | Accepted: 15 June 2025

DOI: 10.1111/bjet.13609

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

# Effects of AI-generated adaptive feedback on statistical skills and interest in statistics: A field experiment in higher education

Elisabeth Bauer[1] | Constanze Richters[2] | Amadeus J. Pickal[1] | Moritz Klippert[1] | Michael Sailer[1] | Matthias Stadler[2]

[1]Learning Analytics and Educational Data Mining, University of Augsburg, Augsburg, Germany

[2]Institute of Medical Education, LMU University Hospital, LMU Munich, Munich, Germany

**Correspondence**
Elisabeth Bauer, Learning Analytics and Educational Data Mining, University of Augsburg, Universitätsstr. 10, Augsburg 86159, Germany.
Email: elisabeth.bauer@uni-a.de

**Funding information**
Stiftung Innovation in der Hochschullehre

## Abstract

This study explores whether AI-generated adaptive feedback or static feedback is favourable for student interest and performance outcomes in learning statistics in a digital learning environment. Previous studies have favoured adaptive feedback over static feedback for skill acquisition, however, without investigating the outcome of students' subject-specific interest. This study randomly assigned 90 educational sciences students to four conditions in a 2×2 Solomon four-group design, with one factor *feedback type* (adaptive vs. static) and, controlling for pretest sensitisation, another factor *pretest participation* (yes vs. no). Using a large language model, the adaptive feedback provided feedback messages tailored to students' responses for several tasks on reporting statistical results according to APA style, while static feedback offered a standardised expert solution. There was no evidence of pretest sensitisation and no significant effect of the feedback type on task performance. However, a significant medium-sized effect of feedback type on interest was found, with lower interest observed in the adaptive condition than in the static condition. In highly structured learning tasks, AI-generated adaptive feedback, compared with static feedback, may be non-essential for learners' performance enhancement and less favourable for learners' interest, potentially due to its impact on learners' perceived autonomy and competence.

**KEYWORDS**
adaptivity, artificial intelligence, feedback, higher education, interest, natural language processing, statistical skills

## Practitioner notes

### What is already known about this topic

- Adaptive feedback has been shown to outperform static feedback in enhancing specific reasoning outcomes in complex, ambiguous learning tasks.
- Generative AI, particularly through large language models, has expanded opportunities for personalised learning support, such as adaptive feedback.

### What this paper adds

- This study found no significant performance differences between groups that received AI-generated adaptive feedback or static expert feedback during structured statistical tasks in a higher education field setting.
- Static feedback outperformed AI-generated adaptive feedback in fostering students' interest in statistics, potentially due to favourable effects on perceived autonomy and competence during self-assessment processes.

### Implications for practice and/or policy

- The results highlight the importance of aligning feedback approaches with task characteristics and learner needs.
- Generative AI for adaptive feedback requires careful design to enhance both cognitive and affective learner outcomes.

# INTRODUCTION

Generative artificial intelligence (generative AI), advanced through large language models (LLMs), introduces new educational opportunities by processing and generating human-like text. Advancements in LLMs have facilitated integrating generative AI into digital learning environments, offering potential for enhancing instructional practices (Kasneci et al., 2023; Yan et al., 2024). Yet, systematic research and evidence on these opportunities is still in its early stages. In particular, using generative AI for adaptive feedback offers a promising approach to support learning by delivering feedback personalised to students' needs (eg, Escalante et al., 2023; Meyer et al., 2024). However, systematic studies comparing AI-generated adaptive feedback to other feedback types are needed to clarify the specific benefits and limitations of this feedback approach and to inform future instructional design and research.

AI-generated adaptive feedback may improve educational outcomes by increasing the accessibility and specificity of feedback compared with traditional feedback approaches, such as static expert-generated feedback, thereby facilitating cognitive processing of the feedback (Chen et al., 2018; Fyfe et al., 2015). This potential may be especially relevant for addressing challenges in fostering complex cognitive skills in higher education, such as statistical skills. Statistical skills are a critical learning objective in many higher education programmes, enabling students to engage with evidence and research to make informed decisions (Gal, 2002; Sharma, 2017). However, especially students in the social sciences, such as educational sciences, often struggle with mastering statistical concepts and maintaining interest in the subject, even though statistical skills are important for understanding and contributing to issues, such as educational effectiveness (Berndt et al., 2021; Bromage et al., 2022; Williams et al., 2008). This warrants further research into the potential of opportunities for practising statistical skills with enhanced learner support. AI-generated adaptive feedback, by offering tailored and specific guidance, holds promise for mitigating these challenges, potentially fostering both statistical skills and students' subject-related interest by facilitating cognitive processing (Chen et al., 2018; Fyfe et al., 2015) and addressing basic psychological needs (Deci & Ryan, 2000; Krapp, 2005). Enhancing these outcomes is essential for equipping students with the skills necessary for academic and professional success.

This study examines the effects of AI-generated adaptive feedback versus static expert feedback on educational sciences students' statistical skills and interest in statistics within a digital learning environment, as systematic comparisons of AI-generated adaptive feedback and investigations of the effects of adaptive feedback on learner interest are scarce. Therefore, the study aims to provide insights into the potential benefits and limitations of generative AI for enhancing cognitive and affective outcomes in learning statistical skills.

## Statistical skills in social sciences education

Statistical skills are crucial in many professional fields taught in higher education, including educational sciences and other social sciences, where statistical results are used to inform recommendations on societal and political issues, such as educational effectiveness. Statistical skills are knowledge elements (eg, statistical knowledge and critical thinking skills) that, together with dispositional elements (eg, beliefs and attitudes), contribute to statistical literacy, which more broadly encompasses the ability to critically evaluate findings and apply statistical reasoning in decision making (Gal, 2002; Sharma, 2017). Because of their role in statistical literacy, statistical skills do not only contribute to professional skill sets but also facilitate scientific reasoning and statistical citizenship (Rumsey, 2002), enabling individuals to critically engage with data in public discourse, make informed decisions based on statistical evidence, and actively participate in a data-driven society (Hetmanek et al., 2018; Watson & Callingham, 2003). Despite its importance, social sciences students often struggle with statistical concepts and their application. Berndt et al. (2021) found that educational sciences, sociology and psychology students showed lower performance in a test on risk literacy, statistical concepts and the interpretation of statistical data than peers in medicine and economics. Similarly, Haller and Kraus (2002) found that even psychology students, who typically receive more statistical training than many other social sciences students, had significant deficiencies in understanding and interpreting significance tests. These challenges highlight the need for research on targeted instructional strategies to help social sciences students learn statistical skills.

Learning statistical skills is a progressive process, where students must first acquire basic knowledge and foundational skills before advancing to higher-order skills, such as critical evaluation and reasoning about statistical procedures (Ben-Zvi & Garfield, 2004; Franklin et al., 2007). However, traditional statistics instruction often focuses on conveying knowledge about mathematical computation (Nikiforidou et al., 2010). To advance the application of statistical knowledge (eg, understanding tabular outputs on descriptive and inferential statistics), students need to practice foundational skills, such as the core skills of interpreting and communicating statistical results (Gal, 2002), which can be done through practice tasks, such as reporting statistical results. Such active learning approaches were shown to improve engagement and mastery of statistical concepts and foundational statistical skills (Lloyd & Robertson, 2012; Stark et al., 2009). For example, Stark et al. (2009) showed that a problem-based digital environment with additional learning support significantly improved students' performance in statistical tasks. Digital learning environments generally offer scalable solutions for teaching statistical skills to large student groups, especially for practising skills, such as interpreting and communicating statistical results. However, the effectiveness of these teaching approaches depends on students' engagement, which is partially influenced by subject-specific interest (Hui et al., 2019).

## Social sciences students' interest in statistics

Social science students often have ambivalent views on statistics, approaching courses with limited interest and perceiving the subject as overly technical and disconnected from their disciplines (Bromage et al., 2022; Williams et al., 2008). Interest is broadly defined by positive cognitive and emotional engagement with a specific object or domain (Hidi et al., 2004; Krapp & Prenzel, 2011). Within self-determination theory, it is viewed as a component of intrinsic motivation, which describes the tendency to engage in activities out of interest and enjoyment (Deci & Ryan, 2000; Krapp, 2002). Many students lack specific interest in statistics, even if acknowledging its utility for academic and professional goals, encouraging pragmatic engagement (Kulacki & Aikens, 2024). However, learners' subject interest enhances intrinsic motivation and task engagement, and sustained engagement can, in turn, facilitate future interest development (Hui et al., 2019; Rotgans & Schmidt, 2011a; Schiefele, 1991).

Interest in statistics involves cognitive and emotional engagement, driven by personal relevance and positive experiences that foster sustained interaction and knowledge acquisition (Hidi et al., 2004; Krapp & Prenzel, 2011). Accordingly, interest serves as both a condition and outcome of learning (Krapp, 2002; Prenzel, 1992). Its development begins with triggered situational interest, where external stimuli or novelty spark temporary engagement, progressing to maintained situational interest, characterised by a more lasting but still context-dependent involvement (Hidi & Renninger, 2006). Repeated, personally relevant interactions can evolve situational into stable individual interest, reflecting a relatively enduring predisposition to engage with the subject. Evidence from a latent-state trait analysis provides support for the role of situation-specific effects on situational interest, as substantial variance remained independent of pre-existing individual interest (Knogler et al., 2015). Facilitators of situational interest involve cognitive and affective components, shaped by personal relevance and supportive social or educational contexts (Hidi & Renninger, 2006; Krapp, 2002). Specifically, building on self-determination theory (Deci & Ryan, 2000), the fulfilment of basic psychological needs for competence, autonomy and relatedness is considered to facilitate subject-specific interest due to the positive emotional and cognitive engagement during relevant person-object interactions (Krapp, 2005). Competence entails feeling effective and capable, autonomy reflects the need for self-direction aligned with personal values, and relatedness involves feeling connected and supported. When these needs are fulfilled, they generate positive experiences crucial for maintaining situational interest (Benlahcene et al., 2021; Durik et al., 2015; Linnenbrink-Garcia et al., 2013).

Digital learning environments offer scalable solutions for teaching, including courses on statistical skills, but their impact on students' interest varies. Depending on their degree of flexibility, these environments can hinder or foster autonomy, while insufficient feedback and interactivity often fail to meet competence and relatedness needs (Chiu, 2023; Griendling et al., 2022). This highlights the need for further research on how to design digital environments that not only support the learning of statistical skills but also foster interest in statistics, for example, through suitable feedback.

## Effects of feedback in digital learning environments

Feedback is essential for fostering both cognitive and affective outcomes in digital learning environments. It provides information to help learners modify their thinking or behaviour and improve learning (Shute, 2008). Meta-analyses indicate that detailed feedback on adequate task processing is particularly effective for enhancing cognitive outcomes, such as task performance, but its benefits for affective outcomes are less consistent (Hattie & Timperley, 2007; Wisniewski et al., 2020).

Digital learning environments that use activating learning formats, such as writing tasks for enhancing reasoning skills, often employ *static feedback*, such as expert solutions. Static feedback can provide elaborated information on optimal task processing but does not explicitly address learners' task solutions or adapt to their individual needs in any way (ie, knowledge of correct response; Narciss et al., 2014). Therefore, static feedback requires learners to compare their performance against the indicated desired outcomes, engaging in resource-intensive self-assessment processes (Black & Wiliam, 2009; Nicol, 2021). While advanced learners might benefit from this additional engagement, especially novice learners might experience increased cognitive load, which can contribute to cognitive resource depletion and diminished learning effectiveness (Chen et al., 2018; Fyfe et al., 2015). In contrast, *adaptive feedback* personalises feedback to highlight areas for improvement, thereby increasing accessibility and specificity (Deeva et al., 2021; Maier & Klotz, 2022; Plass & Pawar, 2020), potentially making the feedback easier to process (Sailer et al., 2023). Grounded in automated analytics for formative assessment (Bauer et al., 2025), adaptive feedback can compare learners' current performance (feed back) with the target performance or learning goal (feed up) and provide guidance on next steps for improvement (feed forward; Hattie & Timperley, 2007).

Studies demonstrate adaptive feedback's effectiveness in contexts, such as essay writing, exam performance and self-regulated learning (Butterfuss et al., 2022; Horbach et al., 2022; Lim et al., 2021), though comparisons often focus on prior performance (ie, without control condition) or no-feedback control groups. However, such study methods limit the ability to derive specific implications for optimising the design of feedback, prompting further research to compare different feedback types to gain more targeted insights. For example, immediate adaptive feedback and delayed peer feedback yielded similar effects on intrinsic motivation and writing performance while outperforming a no-feedback condition (Fidan & Gencel, 2022). Additionally, two studies compared adaptive and static feedback in digital simulation-based learning environments, finding that while both feedback types resulted in similar judgement accuracy, adaptive feedback significantly enhanced justification quality (Bauer et al., 2025; Sailer et al., 2023). While these studies employed unsupported posttestand found persisting positive effects, other studies suggest that adaptive feedback benefits during learning do not always extend to posttests without feedback (Ahmed et al., 2020). These mixed findings suggest that adaptive feedback can enhance certain cognitive learning outcomes, such as reasoning in terms of justification skills. However, further research is needed to determine the conditions under which personalised learning support through adaptive feedback adds value or has limited effects.

While the findings mainly focus on cognitive outcomes, the effects of adaptive versus static feedback on interest remain underexplored. Feedback variations might differ in their impact on factors, such as perceived autonomy (Wisniewski et al., 2020), as static feedback, by allowing learners to engage in self-assessment at their own pace and chosen depth, might better support learners' perceived autonomy, fostering a sense of control over the learning process. In contrast, adaptive feedback, with its interactivity and personalised support, might enhance learners' perceived relatedness by simulating a sense of individualised attention, which could be particularly valuable in digital learning environments with limited social interaction (Salikhova et al., 2020). While digital settings cannot replace interactions with peers and teachers, feedback that is not only immediate but also adaptive to learners' needs might still foster a sense of interactivity and engagement (Jeon, 2022). Adaptive feedback's impact on perceived competence likely varies depending on whether it provides positive or negative evaluations. Feedback reinforcing correct responses can enhance subject interest by increasing perceived competence (Jansen et al., 2025). Simultaneously, corrective feedback on errors can reduce perceived competence and affective outcomes although this does not necessarily diminish cognitive benefits (Gombert et al., 2024; Kuklick et al., 2023), as effort-intensive cognitive gains do not always align with positive affect (Guggemos et al., 2022). Thus, feedback effectiveness depends on its design and can vary in its impact on different outcomes, highlighting the need for further research on how adaptive and static feedback influence not only task performance but also learners' interest.

## Adaptive feedback using generative AI

Advancements in AI, particularly in natural language processing (NLP), have driven research on using AI to deliver adaptive feedback. Most prior research relied on analytical AI approaches, using algorithms to automatically assess student responses and adaptively select predefined feedback. For example, gradient boosting classifiers with decision trees have been applied to provide simple adaptive feedback (eg, correct/incorrect) for several categories, alongside recommendations for writing quality (Horbach et al., 2022). Neural networks, such as bidirectional long short-term memory networks with a conditional random fields output layer, have been used to analyse responses and adaptively combine predefined feedback paragraphs (Bauer et al., 2025; Sailer et al., 2023). However, research comparing algorithms for automated essay rating shows that more recent NLP algorithms—specifically, transformer-based LLMs, such as bidirectional encoder representations from transformers (BERT)—outperform traditional methods like logistic regression, random forest and gradient boosting in automated text analysis (Gombert et al., 2024).

While analytical AI approaches dominate past research, the new transformer-based LLMs like BERT and generative pre-trained transformers (GPT) are also known for their capacity to generate human-like text, expanding the potential of generative AI for adaptive feedback. Evaluations of AI-generated feedback show mixed results compared with expert feedback, with findings ranging from student preference for expert feedback to evidence supporting AI-generated feedback as a high-quality, efficient alternative (W. Dai et al., 2024; Jacobsen & Weber, 2025; Jansen et al., 2024; Steiss et al., 2024). Such findings also highlight that high-quality feedback from generative AI systems depends on proper prompt engineering and well-defined (eg, expert) task solutions (Heston & Khun, 2023; Jacobsen & Weber, 2025; Stamper et al., 2024).

Empirical studies have started testing the effects of AI-generated adaptive feedback. For example, AI-generated feedback for English language learners' writing performance was equally effective to tutor feedback, with comparable learner performance and perceptions of both feedback types (Escalante et al., 2023). Another study found that, compared with no feedback, AI-generated feedback enhanced secondary students' text revision, motivation and positive emotions (Meyer et al., 2024). Some studies explored rather generic chatbot applications that deliver feedback alongside further learner support functions, such as answering questions or providing hints (eg, Ma et al., 2024). However, research specifically comparing AI-generated adaptive feedback to other feedback types remains limited. Thus, further studies need to assess its effects on outcomes like student performance and interest, particularly in comparison to other feedback types, such as static feedback.

## The present field study

This experimental field study examines the effects of AI-generated adaptive feedback versus static expert feedback in a digital learning environment used during a regular statistics lecture for educational sciences students. The study focuses on two key outcomes: (1) students' task performance when reporting statistical results as an indicator for their statistical skills and (2) students' interest in statistics. Previous research indicates that adaptive feedback can outperform static feedback in enhancing reasoning performance both during a learning phase with feedback and in a posttest without feedback (Bauer et al., 2025; Sailer et al., 2023). However, this previous research did not yet employ generative AI, but was based on analytic AI approaches to adaptively combine predefined feedback paragraphs. While AI-generated adaptive feedback compared with peer feedback has shown positive effects on the related outcome of intrinsic motivation (Fidan & Gencel, 2022), its effects relative to static feedback on subject interest remain unknown.

> **H1.** Based on prior research (Bauer et al., 2025; eg, Sailer et al., 2023), we hypothesise that adaptive feedback enhances performance in statistical tasks more effectively than static feedback, observable both (1a) during a learning phase with feedback and (1b) in a posttest phase without feedback.

> **H2.** We hypothesise that adaptive and static feedback differ in their effects on students' interest in statistics. However, as these feedback types may have varying effects on perceived autonomy, competence and relatedness, the direction of the feedback difference is unclear.

# METHODS

## Sample and design

The sample consisted of first-semester bachelor's students in educational sciences at a German university. Data collection occurred under field conditions, in a lecture on empirical research methods. Of the initial 118 participants, those with incomplete posttest data were excluded, resulting in a final sample of $N=90$ students. Participants had an average age of $M=21.6$ years ($SD=6.3$).

Participants were randomly assigned to one of the four groups in our field experiment with a 2×2 Solomon four-group design, including the factors *feedback type* (adaptive vs. static) and *pretest participation* (yes vs. no). This design, developed by Solomon (1949), extends the classic pre-post two-group design to control for potential training and sensitisation effects from pretests, especially when pretest tasks resemble intervention tasks (as is the case in our study; see Procedure section). Comparing pretested and non-pretested groups allows us to differentiate effects of the feedback treatment from pretest-induced changes. In the present study, this design was employed to account for potential training effects or testing fatigue resulting from the pretest tasks, as these would influence students' engagement with the feedback types and the outcome variables. A main effect of pretest participation would indicate that the pretest either facilitated learning (training effect) or caused cognitive strain (testing fatigue). An interaction effect with feedback type would suggest that these effects were either amplified or compensated for by adaptive or static feedback. This design ensures that observed differences in the outcome variables can be attributed to the feedback intervention rather than to confounding influences related to the pretest.

## Procedure

The study was conducted under field conditions during a regular first-semester lecture session on empirical research methods in an educational sciences bachelor's programme. Students were informed in advance about a training session on *t*-test interpretation, involving voluntary, anonymous data collection for a study. However, before and during the experiment, students were not specifically informed about the experimental feedback variation to avoid expectancy effects that could bias their engagement with the feedback. Using their personal laptops or tablets, students participated in the lecture room.

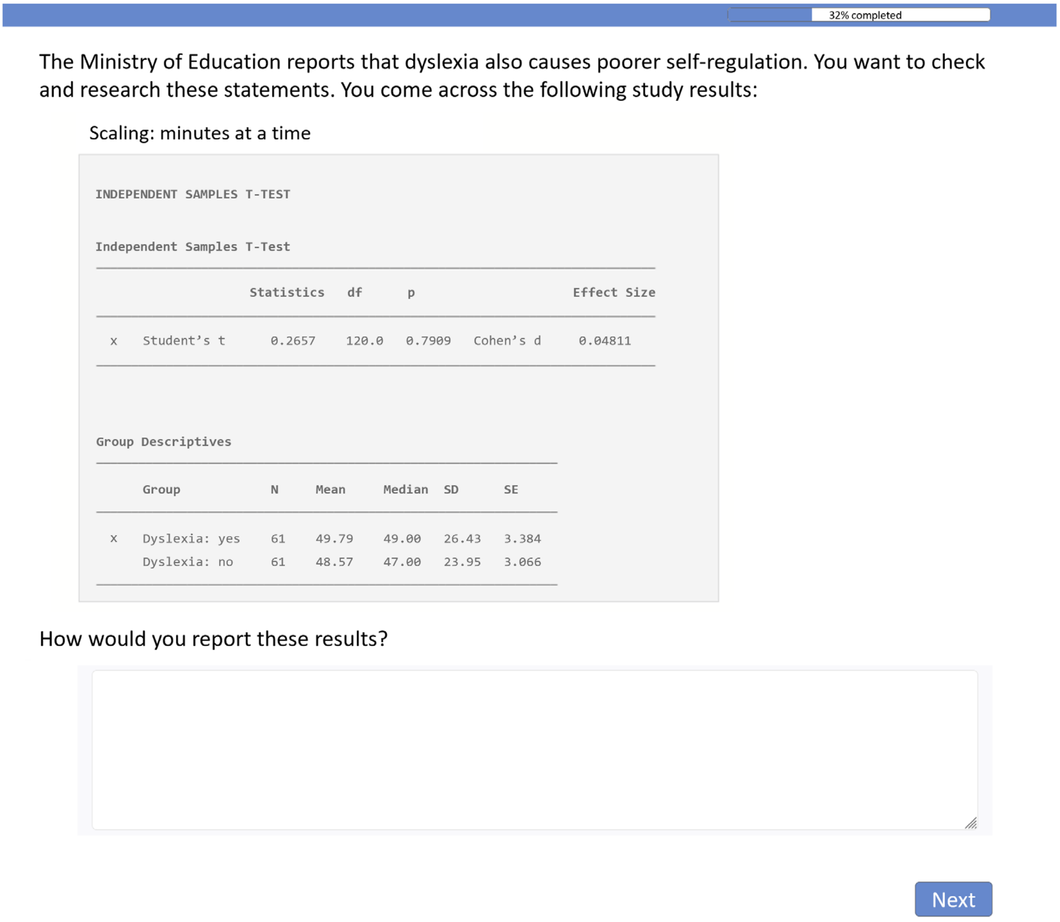
The session began with a briefing explaining the study procedure and aims, followed by a recapitulation of *t*-test interpretation and reporting. Students accessed a testing and learning environment implemented on the SoSci Survey platform. Participants first completed a questionnaire on prior interest in statistics (see *Measurements*) and then worked on several tasks requiring them to interpret and report *t*-test results. Each task included a brief, fictional study description, a tabular *t*-test output and an open-ended question for reporting findings in APA format (see Figure 1). The *t*-test tasks, which were randomly assigned to measurement points and experimental groups, had an identical structure and varied only in variable names and numerical outputs to ensure high task similarity and equal task difficulty. The pretest groups completed three *t*-test tasks without feedback as a performance pretest. All groups participated in the learning phase, completing five *t*-test tasks with feedback that varied by group (AI-generated adaptive feedback or static expert feedback; see *Feedback intervention*). Students did not revise their responses based on the feedback; however, due to the high task similarity, we assumed that they could easily transfer insights from the feedback on one *t*-test task to the next. Finally, all participants completed a performance posttest of three additional *t*-test tasks without feedback and a questionnaire on posttest interest in statistics (see Table 1 for an overview of the study procedure). A debriefing at the end disclosed that the purpose of the study was to compare the effects of AI-generated adaptive feedback and static expert feedback.

Students worked at their own pace, with the measurements and tasks lasting about 45 minutes for those without performance pretest and 60 minutes for those with performance pretest (see Table 1). As compensation, all students were given access to an online tool with similar *t*-test tasks with static expert solutions for the rest of the semester.

## Feedback intervention

### Static feedback

The static feedback consisted of a standardised expert solution for each task that reported the *t*-test results in APA format (see Figure 2). The expert solutions for each task were developed collaboratively by the two senior lecturers of the empirical research methods course,



The Ministry of Education reports that dyslexia also causes poorer self-regulation. You want to check and research these statements. You come across the following study results:

Scaling: minutes at a time

```
INDEPENDENT SAMPLES T-TEST

Independent Samples T-Test
─────────────────────────────────────────────────────────────────────
                  Statistics   df      p                  Effect Size
─────────────────────────────────────────────────────────────────────
  x   Student’s t      0.2657     120.0   0.7909   Cohen’s d    0.04811
─────────────────────────────────────────────────────────────────────


Group Descriptives
─────────────────────────────────────────────────────────
      Group            N     Mean     Median  SD       SE
─────────────────────────────────────────────────────────
  x   Dyslexia: yes    61    49.79    49.00   26.43    3.384
      Dyslexia: no     61    48.57    47.00   23.95    3.066
─────────────────────────────────────────────────────────
```

How would you report these results?

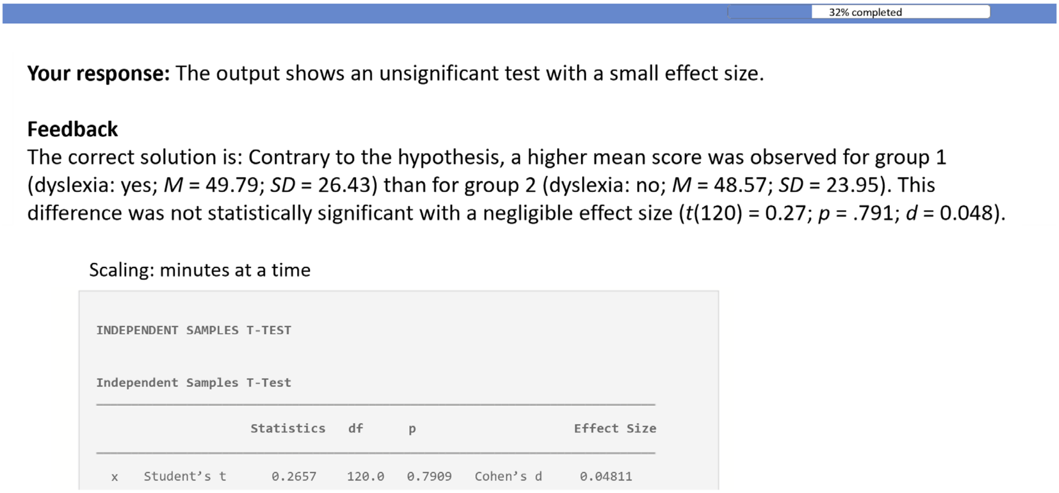
**FIGURE 1** Screenshot of a *t*-test task, including a brief information about a fictional study, a fictional *t*-test output and an open response question for writing a report of the findings. Translated from German.

**TABLE 1** 2×2 Solomon four-group design (Solomon, 1949) with the factor feedback type (adaptive vs. static) and the factor pretest participation (yes vs. no), additionally including questionnaire measurements of students' interest in statistics.

| Experimental group (*n*) | Pre-interest in statistics (2 min) | Performance pretest (12 min) | Learning phase with different feedback types (30 min) | Performance posttest (12 min) | Post-interest in statistics (2 min) |
|---|---|---|---|---|---|
| 1 (18) | Q | T | $TF_1$ | T | Q |
| 2 (24) | Q | T | $TF_2$ | T | Q |
| 3 (24) | Q | | $TF_1$ | T | Q |
| 4 (24) | Q | | $TF_2$ | T | Q |

Abbreviations: T, measurement tasks; TF, learning tasks with feedback ($TF_1$ = with adaptive feedback; $TF_2$ = with static feedback); Q, questionnaire measurement.

ensuring consistency and high quality through expert agreement. The static feedback was shown immediately after participants completed a *t*-test task, regardless of the quality of their response.



**Your response:** The output shows an unsignificant test with a small effect size.

**Feedback**

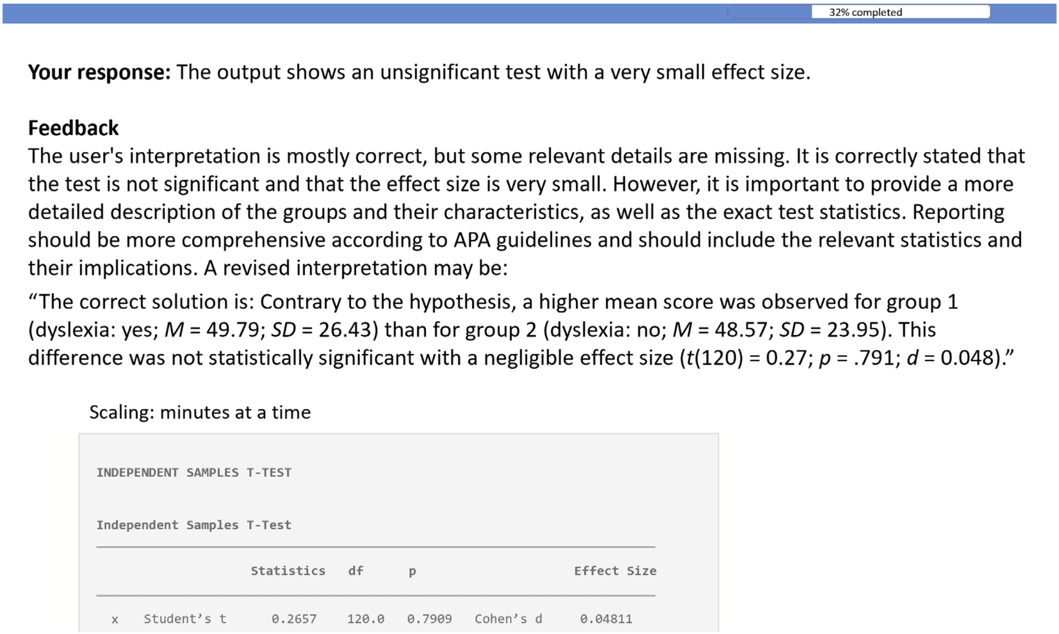
The correct solution is: Contrary to the hypothesis, a higher mean score was observed for group 1 (dyslexia: yes; *M* = 49.79; *SD* = 26.43) than for group 2 (dyslexia: no; *M* = 48.57; *SD* = 23.95). This difference was not statistically significant with a negligible effect size (*t*(120) = 0.27; *p* = .791; *d* = 0.048).

Scaling: minutes at a time

```
INDEPENDENT SAMPLES T-TEST


Independent Samples T-Test
─────────────────────────────────────────────────────────────────────────
                         Statistics    df       p                     Effect Size
─────────────────────────────────────────────────────────────────────────
  x    Student’s t        0.2657     120.0    0.7909    Cohen’s d       0.04811
```

FIGURE 2 Screenshot of a static expert feedback for one of the *t*-test tasks. Translated from German.



**Your response:** The output shows an unsignificant test with a very small effect size.

**Feedback**

The user's interpretation is mostly correct, but some relevant details are missing. It is correctly stated that the test is not significant and that the effect size is very small. However, it is important to provide a more detailed description of the groups and their characteristics, as well as the exact test statistics. Reporting should be more comprehensive according to APA guidelines and should include the relevant statistics and their implications. A revised interpretation may be:

“The correct solution is: Contrary to the hypothesis, a higher mean score was observed for group 1 (dyslexia: yes; *M* = 49.79; *SD* = 26.43) than for group 2 (dyslexia: no; *M* = 48.57; *SD* = 23.95). This difference was not statistically significant with a negligible effect size (*t*(120) = 0.27; *p* = .791; *d* = 0.048).”

Scaling: minutes at a time

```
INDEPENDENT SAMPLES T-TEST


Independent Samples T-Test
─────────────────────────────────────────────────────────────────────────
                         Statistics    df       p                     Effect Size
─────────────────────────────────────────────────────────────────────────
  x    Student’s t        0.2657     120.0    0.7909    Cohen’s d       0.04811
```

FIGURE 3 Screenshot of an artificial intelligence (AI)-based adaptive feedback for one of the *t*-test tasks. Translated from German.

## Adaptive feedback

The adaptive feedback was generated by ChatGPT, which delivered personalised messages tailored to students' responses (see Figure 3). This feedback was integrated into the digital learning environment via API calls linking the SoSci Survey platform with GPT-3.5 Turbo (OpenAI, 2022).

The feedback was created using a standardised prompt based on best practices from Heston and Khun (2023). ChatGPT was instructed to act as a tutor specialising in empirical research methods, particularly in reporting *t*-test results in APA format. It compared students' responses to the expert solution (also used as static feedback) and provided constructive guidance, highlighting areas for improvement. The prompt remained consistent across tasks and was adapted only for each task's specific expert solution. The prompt had two components:

For the *context and role* definition, ChatGPT was instructed to adopt the role of a tutor and compare the student's interpretation with the expert solution: 'You are acting as a tutor for empirical research methods, specialising in reporting *t*-test results in APA format. Compare the user's interpretation with the correct APA-style solution: [Expert solution]. Below is the user's interpretation: [Student's answer inserted here]'.

For specifying the *required action and output format*, ChatGPT was instructed to evaluate the student's response, identify discrepancies and provide actionable suggestions for improvement: 'Evaluate the user's interpretation by comparing it to the correct reporting solution and provide feedback on how it can be improved'.

The adaptive feedback prompt and integration were pretested and refined for clarity, accuracy and consistency. The prompt development process involved evaluating and iteratively refining several prompt versions, which were tested for the resulting feedback using fictitious student responses. These fictitious student responses ranged from a few-word response, to intermediate answers that were assumed to resemble typical responses, to a response that closely resembled the expert solution. The feedback was piloted with two educational sciences students, suggesting high face and content validity for the feedback messages. Embedded into the digital learning environment, the system provided adaptive feedback immediately after each task.

## Measurements

### Performance in statistical tasks

As an indicator for statistical skills, we assessed students' performance on the *t*-test tasks, based on their written responses. Responses were manually coded using a scheme developed by decomposing relevant entities from the expert solution. The scheme consisted of 10 items per task (see Table 2), indicating whether relevant statistical values were reported and interpreted correctly. Each task was scored from 0 to 10, with higher scores indicating better performance in statistical tasks (ie, greater statistical skills). To ensure consistency, two raters underwent training and interrater reliability was assessed by double coding 10% of the data, yielding good reliability ($\kappa = 0.86$), before one rater coded the remaining data. Reliability scores for each group and measurement point were satisfactory: McDonald's $\omega$ ranged from 0.77 to 0.78 for students' *performance in the pretest*, from 0.86 to 0.91 for students' *performance in the learning phase*, and from 0.77 to 0.85 for *performance in the posttest*. Mean scores with a possible range from 0 to 10 were calculated for each measurement point.

### Interest in statistics

The variables *pre-interest in statistics* and *post-interest in statistics* were assessed using an adapted version of the situational interest questionnaire by Rotgans and Schmidt (2011b) designed to measure subject-specific interest. Three of the scale's six items were selected

**TABLE 2** Example coding scheme for participants' responses to a *t*-test task, with task-general criteria and an example of task-specific details for one of the *t*-test tasks.

| Item | Criterion description | Example specification for one *t*-test task |
|---|---|---|
| 01 | The mean for Group 1 is correctly reported | $M=2.70$ |
| 02 | The standard deviation for Group 1 is correctly reported | $SD=0.88$ |
| 03 | The mean for Group 2 is correctly reported | $M=3.13$ |
| 04 | The standard deviation for Group 2 is correctly reported | $SD=1.12$ |
| 05 | The means of both groups are correctly compared | Group 1 < Group 2 |
| 06 | The *t*-value is correctly reported. | $t=-1.6$ |
| 07 | The degrees of freedom are correctly reported | $t(58)$ |
| 08 | The *p*-value is correctly reported | $p>.05$ or $p=.115$ |
| 09 | The *p*-value is correctly interpreted in words | For example: 'Not significant', 'not statistically supported', 'above the significance threshold' |
| 10 | The effect size is appropriately interpreted in words | For example: 'Weak', 'small effect', 'low effect size' |

because their wording was suitable for both pretest and posttest measures ('I want to know more about today's topic'; 'I enjoy working on today's topic'; and 'I think today's topic is interesting'). We translated the items into German and modified them to make their content more specific by replacing 'today's topic' with 'empirical research methods' (eg, 'I want to know more about empirical research methods') and '*t*-tests' (eg, 'I want to know more about *t*-tests'), resulting in a measurement of six items. Responses were rated on a 5-point Likert scale from 'Strongly disagree' to 'Strongly agree'.

The scale for *pre-interest in statistics* was administered before the performance pretest, showing satisfactory reliability (McDonald's $\omega=0.87$). *Post-interest in statistics* was measured after the performance posttest, with high reliability (McDonald's $\omega=0.93$). Mean scores with a possible range from 1 to 5 were computed for both interest measurements.

## Statistical analyses

The statistical analyses followed the recommendations of Braver and Braver (1988) for evaluating the Solomon four-group design. First, we tested for interaction effects between pretest participation and feedback type to assess whether pretest effects (eg, training or sensitisation) influenced feedback outcomes. If no interaction effects are found, it is recommended to interpret the treatment main effects (ie, of the feedback types).

To test H1, we used a MANOVA to analyse the effects of the feedback type (adaptive vs. static feedback) on performance in (a) the learning phase and (b) the posttest. As an exploratory follow-up analysis, we performed a repeated-measures ANOVA on the subsample with available pretest data to explore how students' statistical task performance changed across the three measurement points depending on feedback type. To test H2, an ANCOVA with the covariate pre-interest in statistics was used to control for the effects of prior interest in statistics on post-interest in statistics.

Assumptions of homogeneity of variances and normality were tested prior to the analyses. Levene's test confirmed homogeneity for all outcomes. Shapiro–Wilk tests indicated some violations of normality in a few experimental groups. However, given ANOVA-based methods' robustness to moderate normality violations (Glass et al., 1972), we proceeded with the planned parametric analyses. For the repeated-measures ANOVA, Mauchly's test indicated that the assumption of sphericity was violated, $W = 0.45$, $\chi^2(2) = 31.00$, $p < 0.001$, which is why Greenhouse–Geisser corrected results ($\varepsilon = 0.65$) were used for the within-subjects analysis. For the ANCOVA, homogeneity of regression slopes was tested using an interaction model. The interaction effects of feedback type and pre-interest in statistics ($F(1, 83) = 0.02$, $p = 0.891$), pretest participation and pre-interest in statistics ($F(1, 83) = 1.31$, $p = 0.256$), as well as feedback type, pretest participation and pre-interest in statistics ($F(1, 83) = 0.73$, $p = 0.396$) were all not significant, indicating that the assumption was met. All analyses were conducted in SPSS 29 with an alpha level of $\alpha = 0.05$.

# RESULTS

## Effect on students' performance in statistical tasks

Descriptive statistics of students' performance in the learning phase and in the posttest indicated that students in all groups showed similar performance, with minor descriptive advantages in the AI-generated adaptive feedback groups (see Table 3). Overall, all students showed mediocre performance relative to the maximum of 10 achievable points per task. Following the recommendations by Braver and Braver (1988), we initially confirmed that there was no pretest sensitisation as there was no significant interaction effect of the factors feedback type and pretest participation on the performance in the learning phase, $F(1,86) = 0.80$, $p = 0.374$, $\eta_p^2 = 0.01$ (main effect of pretest participation on the learning phase performance was $F(1,86) = 0.97$, $p = 0.328$, $\eta_p^2 = 0.01$), or on the performance in the posttest, $F(1,86) = 0.43$, $p = 0.514$, $\eta_p^2 = 0.01$ (main effect of pretest participation on the posttest performance was $F(1,86) = 3.83$, $p = 0.053$, $\eta_p^2 = 0.04$).

The MANOVA used for investigating H1 found no significant difference between the two feedback types in their effect on students' performance in the learning phase, $F(1,86) = 2.83$, $p = 0.096$, $\eta_p^2 = 0.03$, or on the performance in the posttest, $F(1,86) = 0.90$, $p = 0.347$, $\eta_p^2 = 0.01$. These results suggest that the effects of the two feedback types on students' performance were not significantly different.

In an exploratory follow-up analysis, we conducted a repeated-measures ANOVA on the subsample of groups that participated in the pretest to examine the changes in students' statistical task performance across the three measurement points depending on the feedback type. The performance trajectories of this subsample exploration are shown in Figure 4. Although both types of feedback produced a very similar end-point performance in the posttest, adaptive feedback appeared to produce a more rapid initial increase in statistical task performance that subsided by the posttest, whereas static feedback produced smaller gains that also levelled off but appeared to follow a more consistent pattern.

There was no significant main effect of feedback type, $F(1,40) = 1.71$, $p = 0.198$, $\eta_p^2 = 0.04$, no significant main effect of measurement point, $F(1.29,51.68) = 2.11$, $p = 0.147$, $\eta_p^2 = 0.05$ (Greenhouse–Geisser corrected), and no significant interaction effect between feedback type and measurement point, $F(1.29,51.68) = 1.12$, $p = 0.312$, $\eta_p^2 = 0.03$ (Greenhouse–Geisser corrected), suggesting no consistent performance changes. To investigate a possible non-linear trajectory (performance increases followed by a plateau or decrease; see Figure 4), we decided to further examine polynomial trends. Consistent with the non-significant omnibus effect, the within-subjects contrasts indicated no significant linear trend, $F(1,40) = 0.35$, $p = 0.559$, $\eta_p^2 = 0.009$. However, a significant quadratic trend was found, $F(1,40) = 13.93$, $p < 0.001$, $\eta_p^2 = 0.258$, supporting the descriptive finding that performance increased from pretest to the learning phase and then levelled off at the posttest. The interaction between

TABLE 3 Descriptive statistics of the outcome variables.

| Feedback type | Pretest participation | Performance in the learning phase | Performance in the posttest | Post-interest in statistics |
|---|---|---|---|---|
| | | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) |
| Adaptive | Yes | 5.09 (3.10) | 3.70 (4.06) | 2.28 (1.00) |
| | No | 5.15 (3.48) | 5.65 (3.56) | 2.78 (1.18) |
| | Total | 5.12 (3.29) | 4.82 (3.86) | 2.56 (1.12) |
| Static | Yes | 3.38 (2.84) | 3.49 (3.11) | 3.01 (0.83) |
| | No | 4.63 (3.07) | 4.46 (3.40) | 3.06 (0.87) |
| | Total | 4.00 (2.99) | 3.97 (3.26) | 3.04 (0.84) |

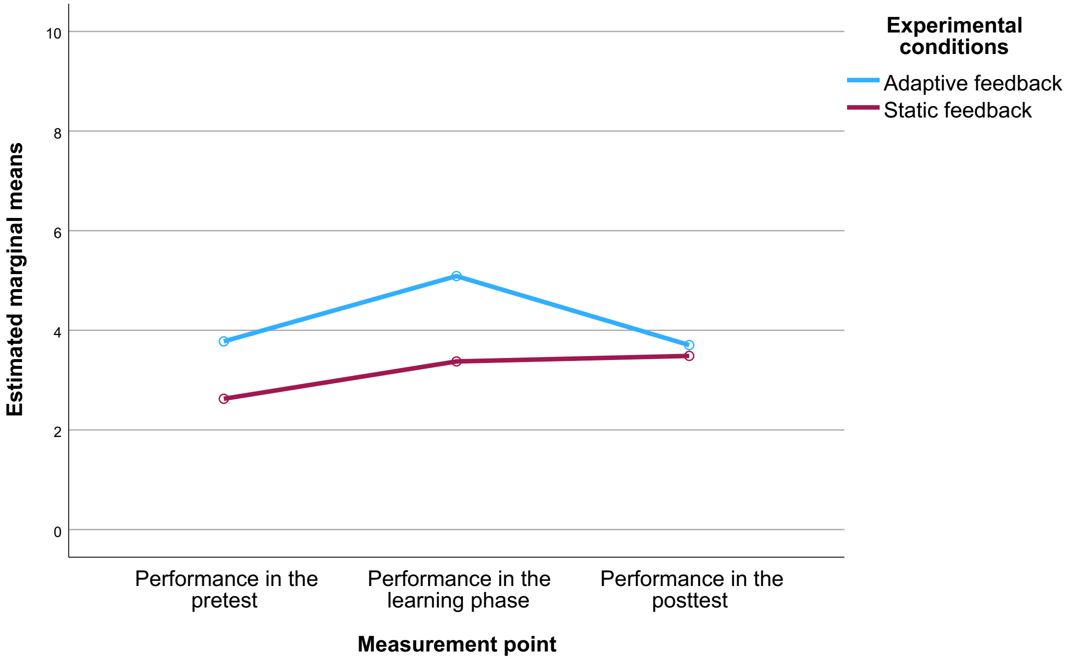


FIGURE 4 Exploratory subsample analysis of performance trends in the subsample assigned to the conditions with pretest participation.

feedback type and measurement point likewise indicated no linear trend, $F(1,40)=0.49$, $p=0.488$, $\eta_p^2=0.012$, but a significant quadratic interaction, $F(1,40)=5.30$, $p=0.027$, $\eta_p^2=0.117$, suggesting that the overall curvature of the trajectory differed by feedback type.

To further explore these effects at adjacent time points, we conducted within-subjects contrasts, which showed a significant performance improvement between the pretest and the learning phase, $F(1,40)=6.75$, $p=0.013$, $\eta_p^2=0.14$, while no significant performance difference between the learning phase and the posttest was observed, $F(1,40)=2.46$, $p=0.125$, $\eta_p^2=0.06$. There were no significant interactions between measurement point and feedback type for the more localised adjacent contrasts between the pretest and learning phase, $F(1,40)=0.50$, $p=0.484$, $\eta_p^2=0.01$, and between the learning phase and posttest, $F(1,40)=3.39.75$, $p=0.073$, $\eta_p^2=0.08$. Bonferroni-adjusted pairwise comparisons showed no significant performance differences between feedback conditions at any measurement point (pretest: $M_{diff}=1.15$, $SE=0.90$, $p=0.206$; learning phase: $M_{diff}=1.71$, $SE=0.92$, $p=0.070$; posttest: $M_{diff}=0.22$, $SE=1.11$, $p=0.845$), confirming the absence of initial group differences and suggesting that the feedback types did not lead to significant between-group differences at any time point. The significant quadratic trend confirms a non-linear trajectory in the feedback conditions, with performance improving from pretest to learning phase and then levelling off. Although the curvature of this trajectory differed by feedback type, no significant between-group differences emerged at any individual measurement point.

## Effect on students' interest in statistics

To test whether adaptive and static feedback differ in their effects on students' interest in statistics (H2), we conducted an ANCOVA with pre-interest in statistics as the covariate, $F(1,85)=68.61$, $p<0.001$, $\eta_p^2=0.45$. Again, we initially confirmed that there was no interaction effect of feedback type and pretest participation, $F(1,85)=0.36$, $p=0.551$, $\eta_p^2<0.01$ (main effect of pretest participation was $F(1,85)=0.96$, $p=0.329$, $\eta_p^2=0.01$).

In line with H2, there was a significant difference between the two feedback types in their effect on students' interest in statistics, with a medium-sized effect, $F(1,85)=5.69$, $p=0.019$, $\eta_p^2=0.06$. Students in the adaptive feedback condition showed significantly lower interest than students in the static condition (see Table 3). Therefore, the results support the hypothesis that adaptive and static feedback differ in their effects on students' interest in statistics, suggesting a favourable effect of static feedback compared with adaptive feedback.

# DISCUSSION

## The role of different feedback types for fostering students' statistical skills

Prior research has shown significant benefits of adaptive feedback over static feedback for reasoning tasks in simulation-based environments, with positive effects on justification quality but not on overall judgement accuracy (Bauer et al., 2025; Sailer et al., 2023). In the present study, however, we found no evidence that AI-generated adaptive feedback significantly outperformed static feedback in supporting students' statistical skills, as indicated by their performance in tasks on reporting statistical results. An exploratory subsample analysis indicated a non-linear trajectory: performance increased significantly from pretest to the learning phase and then levelled off by the posttest in both feedback conditions, without significant changes between learning phase and posttest. Trend analyses confirmed a significant quadratic trend and suggested a differential quadratic interaction by feedback type. Descriptive results indicated that the adaptive feedback seemed to produce a slightly higher initial gain in performance that subsided in the posttest, whereas static feedback yielded more gradual gains that then plateaued. However, these differences were not significant in the within-subjects contrasts at individual time increments, nor were there significant between-group differences at any discrete measurement point. Together, these results suggest that, despite slightly distinct learning dynamics, adaptive and static feedback produced comparable performance outcomes. While other studies found positive effects of adaptive feedback, they lacked control conditions or had no-feedback control groups (eg, Butterfuss et al., 2022; Escalante et al., 2023; Fidan & Gencel, 2022; Horbach et al., 2022; Lim et al., 2021; Meyer et al., 2024), making their results difficult to integrate with our findings, but not necessarily contradictory.

Discrepancies in the findings might stem from differences in task complexity, that is, the amount and ambiguity of the information that needs to be processed (Stadler et al., 2019). Simulated reasoning tasks, as studied by Bauer et al. (2025) and Sailer et al. (2023), involve processing large, ambiguous information where personalised guidance is essential—particularly for achieving high-quality justifications, which require integrating multiple pieces of evidence into coherent arguments (Bauer et al., 2022). As such learning contexts can involve higher cognitive load, adaptive feedback might play a bigger role in preventing learners' cognitive resource depletion in these learning contexts (Chen et al., 2018; Fyfe et al., 2015). In contrast, the technical and structured nature of the statistical tasks in the present study may allow learners to self-assess effectively using the static expert feedback without necessarily needing adaptive support to relieve their cognitive resources. This aligns with findings from digital game-based learning, where feedback adjusted to task complexity—simple hints for easy tasks and detailed explanations for complex ones—improved learning and engagement over standardised feedback support (Mao et al., 2024). Thus, matching the feedback type to the complexity of the task may help to achieve optimal outcomes.

Although, as indicated by the descriptive findings, AI-generated adaptive feedback may have offered minor benefits through additional guidance (Deeva et al., 2021; Maier & Klotz, 2022; Plass & Pawar, 2020), these benefits might have been counterbalanced by enhanced cognitive engagement through self-assessment processes in the static feedback condition (Black & Wiliam, 2009; Nicol, 2021). Static feedback prompts learners to actively compare their responses with the expert solution, which might foster knowledge construction (Chi & Wylie, 2014). Feedback effectiveness may therefore depend on task characteristics—whether learners can independently understand areas for improvement or need adaptive support. Therefore, while adaptive feedback seems critical for complex, ambiguous tasks, static feedback may suffice or even promote beneficial engagement in highly structured tasks.

Findings from the exploratory subsample analysis support the interpretation that students' performance in both feedback conditions improved during the learning phase, but performance then levelled off—attenuating but not significantly declining—by the posttest where no feedback was provided, as indicated by a significant quadratic trend and within-subjects contrasts. Given that feedback can serve as a reward (Jansen et al., 2025), its absence during the posttest may have contributed to a motivational decline (Deci & Ryan, 2000).

## Static feedback for supporting students' interest in statistics

Our findings suggest that static feedback was favourable over AI-generated adaptive feedback for students' interest in statistics. Given the reciprocal relationship between learners' interest and task performance, the emerging differences in interest might have contributed to the descriptive trends observed in the exploratory subsample analysis, where the performance attenuation appeared descriptively—though not significantly—more pronounced in the adaptive feedback subgroup than in the static feedback subgroup. Since interest is considered a driver of intrinsic motivation, lower interest levels might have amplified motivational declines (Deci & Ryan, 2000). However, since the subgroup differences in learners' performance were not statistically significant, these trends should not be overinterpreted.

Interest development progresses from triggered to maintained situational interest under supportive conditions (Hidi & Renninger, 2006; Krapp, 2002), driven by fulfilment of basic psychological needs—autonomy, competence and relatedness—which promote positive emotional and cognitive engagement (Deci & Ryan, 2000; Krapp, 2005). The results indicate that, in structured statistical tasks, static feedback may better support these needs than adaptive feedback.

Static feedback might have fostered autonomy by providing a clear expert solution, enabling students to self-assess, compare their responses at their own pace and autonomously decide about the depth of their assessment (Black & Wiliam, 2009; Nicol, 2021). This flexibility allowed learners to control their effort in understanding mistakes, potentially facilitating a sense of independence. In contrast, adaptive feedback, though personalised, might have constrained autonomy by directing learners towards specific areas of improvement, limiting independent exploration, which may have been pronounced in the structured statistical tasks where students might not have been overwhelmed by complexity.

The potential effect of feedback on perceived competence appears more nuanced. Adaptive feedback can reinforce correct elements and can thereby enhance perceived competence and interest (Jansen et al., 2025). However, considering it also emphasises mistakes, it can as well undermine perceived competence, especially in our study context of learning statistics where students often face difficulties (Berndt et al., 2021). Research shows that feedback on errors can lower perceived competence and affective outcomes (Kuklick et al., 2023). In our study, participants' mediocre task performance suggests that adaptive feedback often had to correct mistakes and propose improvements, which might have negatively impacted students' perceived competence. Positive effects of AI-generated adaptive feedback can potentially be achieved through enhancing its affective design and exploring how to generate encouraging corrective feedback.

Although adaptive feedback offers higher interactivity through personalised responses (Plass & Pawar, 2020), in our study, this likely did not compensate for the known issues with fostering students' sense of relatedness in digital learning environments (Salikhova et al., 2020). However, prior research found no difference in intrinsic motivation between adaptive feedback and digital peer feedback (Fidan & Gencel, 2022), indicating that digital feedback—regardless of its source—may not fully replicate the relational benefits of face-to-face interactions with teachers or peers, which seem to be vital for fostering relatedness in learning environments.

The overall pattern of our findings suggests that while adaptive feedback offers personalised guidance, static feedback may better foster interest in structured yet challenging subject contexts like reporting statistical results, possibly by more effectively supporting autonomy and competence needs. However, as this study did not directly measure psychological needs, these interpretations require further investigation.

## Generative AI for providing adaptive feedback in higher education

This study suggests that AI-generated adaptive feedback in higher education is not superior to alternative feedback options, such as static expert feedback in the context of learning with structured statistical tasks, as we found no significant differences in task performance and a relative disadvantage for adaptive feedback on students' interest in statistics. While both feedback types seemed to have similar effects on students' task performance, initially increased performance levelled off in the posttest without feedback, suggesting no consistent pattern in the performance changes. While prior research has shown advantages of adaptive over static feedback in complex and ambiguous reasoning tasks within simulations (Bauer et al., 2025; Sailer et al., 2023), this might be explained by task characteristics, such as complexity and structure, moderating the effects of feedback adaptivity.

While adaptive feedback can be technically well-implemented for structured tasks with clear solutions, this does not necessarily translate into higher learning outcomes. Structured tasks provide well-defined reference points (eg, expert solutions) that facilitate the alignment of AI-generated feedback with expected responses, although ensuring high-quality AI-generated feedback still requires good prompt engineering (Heston & Khun, 2023; Jacobsen & Weber, 2025; Stamper et al., 2024). However, in learning contexts comparable to our study, learners may be able to self-assess effectively using static feedback, reducing the added value of adaptive feedback. Thus, while AI can generate feedback for structured tasks, its added benefits for learning outcomes depend on whether adaptive support offers advantages beyond what well-designed static feedback already provides.

Although using a transformer-based LLM may have improved the alignment of adaptive feedback with student responses compared with earlier studies (Bauer et al., 2025; eg, Horbach et al., 2022; Sailer et al., 2023) that relied on less advanced algorithms (see Gombert et al., 2024), the present study highlights that analytical accuracy alone is not sufficient. Task context, feedback design and especially its affective elements might be relevant factors. Therefore, using generative AI for adaptive feedback requires thoughtful implementation, considering task complexity, learner needs and feedback design.

Practical implications of this study are that for highly structured statistical tasks, static expert feedback may be preferable to AI-generated adaptive feedback because static feedback achieves the same learning outcomes while better maintaining student interest, making it a scalable, resource-efficient choice for large courses. In addition, when exploring generative AI for feedback in digital learning environments, instructors may need to pay close attention to the affective design of the generated feedback messages, which might help maintain student interest and motivation in the learning tasks.

## Limitations and future research

This study offers insights into the effects of AI-generated adaptive feedback in higher education, though several limitations must be acknowledged. Students did not revise their responses after receiving feedback, which could have offered the opportunity for deeper engagement. Additionally, students' revisions could have provided valuable insight into the direct feedback uptake and its impact on students' outcomes, which could not be directly assessed with the data collected in this study. However, we assumed that students would be able to transfer insights from one task to the next given the high similarity between the *t*-test tasks. The study did not capture the AI-generated feedback messages given to students, as the field experiment relied on students using their own devices. This lack of direct evidence assessing feedback accuracy is a limitation. However, prior research highlights the strong performance of LLMs (Gombert et al., 2024) and careful prompt engineering with a clear expert solution as a reference point likely ensured accurate and relevant feedback. While newer LLM versions beyond the employed GPT-3.5 turbo model may exhibit even greater performance, the structured and clear nature of the tasks and reference information facilitated high feedback accuracy. Additionally, although students were not explicitly informed about the feedback variation before or during the experiment to prevent expectancy effects, some may have inferred the use of LLMs to generate the adaptive feedback. If so, this awareness could have influenced their response behaviour, either positively due to novelty effects associated with innovative technology or negatively due to algorithm aversion.

The relatively short intervention time, limited to one lecture session due to the field setting, may have restricted the ability to observe differential effects of adaptive versus static feedback. Possible benefits of adaptive feedback might emerge more clearly over extended intervention periods or with more opportunities for practice. A longer duration could also improve the stability and replicability of findings, as short-term measurements may be more vulnerable to situational influences and transient performance fluctuations, affecting reliability. Additionally, delayed posttest measures could help differentiate between participant fatigue and sustained performance declines.

The sample size, while sufficient for analysing the main effects of feedback type, was constrained by the available cohort size, potentially limiting the power to detect nuanced interaction effects, such as pretest sensitisation. Similarly, in the exploratory follow-up analysis of the subsample with pretest, the fewer observations per condition limited the power to detect more subtle within-subjects and interaction effects; consequently, non-significant effects in the follow-up analysis should be interpreted with caution, and future studies with larger samples will be needed to confirm the exploratory findings. Additionally, the generalizability of this study is limited by the fact that our sample only included education sciences students at a German university. It is therefore uncertain whether our results generalise to other social sciences study programmes or geo-cultural contexts.

This study also emphasised the role of psychological needs (competence, autonomy and relatedness) in explaining differences in students' interest in statistics. However, to minimise measurement time during the lecture session, these needs were not directly assessed. Future studies should address psychological needs more systematically to better understand how the influence of AI-generated adaptive feedback on student interest is explained by these mechanisms.

Despite its constraints, the field setting ensured high ecological validity by conducting the experiment under authentic learning conditions within a university lecture. This approach provides ecologically valid insights into the effects of AI-generated versus static expert feedback on educational sciences students' statistical skills and interest in a single-session intervention.

Future research should investigate how task complexity moderates the impact of feedback adaptivity on learning outcomes. While adaptive feedback may be less critical for structured tasks like statistical interpretation, it could play a more significant role in complex, ill-defined tasks. Additionally, refining the affective design of AI-generated feedback to ensure that corrective guidance remains encouraging could enhance both performance and interest outcomes. Finally, further studies should explore how these factors interact with learner characteristics and instructional contexts to optimise the cognitive and affective benefits of generative AI in diverse educational settings.

# CONCLUSION

This study investigated the effects of AI-generated adaptive feedback and static expert feedback on educational sciences students' statistical skills and interest in statistics when learning in a digital environment. While performance outcomes showed no significant differences between the two feedback types, static feedback demonstrated an advantage in fostering students' interest in statistics. Building on previous findings that highlight significant performance benefits of adaptive feedback in other contexts, our study suggests that the effects of the feedback types might depend on task characteristics, which influence students' need for personalised guidance. Additionally, in structured statistical tasks, static feedback may better support autonomy and competence needs.

Our study highlights that factors beyond analytical accuracy—such as task complexity, learner needs and the feedback design, including affective components—might be crucial for achieving added benefits for learning outcomes through AI support. Future research should explore these interacting factors to optimise the use of AI-generated feedback for both cognitive and affective outcomes in diverse educational contexts.

## ACKNOWLEDGEMENTS

This work was supported by the 'Stiftung Innovation in der Hochschullehre' (project 'Facilitating Competence Development through Authentic, Digital, and Feedback-Based Teaching-Learning Scenarios') under Grant FBM2020-EA-2620-01350. During the writing and revision of this paper, the authors used ChatGPT 4o to revise the draft for readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. Open Access funding enabled and organized by Projekt DEAL.

## CONFLICT OF INTEREST STATEMENT

We have no known conflict of interest to disclose.

## DATA AVAILABILITY STATEMENT

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

## ETHICS STATEMENT

Participants were informed about the purpose and nature of the research, and their participation was voluntary and anonymous. No sensitive personal data were collected, and all data were handled in accordance with data protection regulations.

## ORCID

*Elisabeth Bauer* https://orcid.org/0000-0003-4078-0999
*Amadeus J. Pickal* https://orcid.org/0000-0001-5897-3153

## REFERENCES

Ahmed, U. Z., Srivastava, N., Sindhgatta, R., & Karkare, A. (2020). Characterizing the pedagogical benefits of adaptive feedback for compilation errors by novice programmers. In G. Rothermel & D.-H. Bae (Eds.), *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Software Engineering Education and Training* (pp. 139–150). ACM. https://doi.org/10.1145/3377814.3381703

Bauer, E., Sailer, M. [. M.]., Niklas, F., Greiff, S., Sarbu-Rothsching, S., Zottmann, J. M., Kiesewetter, J., Stadler, M., Fischer, M. R., Seidel, T., Urhahne, D., Sailer, M. [. M.]., & Fischer, F. (2025). AI -based adaptive feedback in simulations for teacher education: An experimental replication in the field. *Journal of Computer Assisted Learning*, *41*(1), Article e13123. https://doi.org/10.1111/jcal.13123

Bauer, E., Sailer, M., Kiesewetter, J., Fischer, M., & Fischer, F. (2022). Diagnostic argumentation in teacher education: Making the case for justification, disconfirmation, and transparency. *Frontiers in Education*, *7*, 977631. https://doi.org/10.3389/feduc.2022.977631

Benlahcene, A., Kaur, A., & Awang-Hashim, R. (2021). Basic psychological needs satisfaction and student engagement: The importance of novelty satisfaction. *Journal of Applied Research in Higher Education*, *13*(5), 1290–1304. https://doi.org/10.1108/JARHE-06-2020-0157

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi (Ed.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3–15). Kluwer Academic Publishers. https://doi.org/10.1007/1-4020-2278-6_1

Berndt, M., Schmidt, F. M., Sailer, M., Maximilian, Fischer, F., Fischer, M. R., & Zottmann, J. M. (2021). Investigating statistical literacy and scientific reasoning & argumentation in medical-, social sciences-, and economics students. *Learning and Individual Differences*, *86*, 101963. https://doi.org/10.1016/j.lindif.2020.101963

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Braver, M. W., & Braver, S. L. (1988). Statistical treatment of the Solomon four-group design: A meta-analytic approach. *Psychological Bulletin*, *104*(1), 150–154. https://doi.org/10.1037/0033-2909.104.1.150

Bromage, A., Pierce, S., Reader, T., & Compton, L. (2022). Teaching statistics to non-specialists: Challenges and strategies for success. *Journal of Further and Higher Education*, *46*(1), 46–61. https://doi.org/10.1080/0309877X.2021.1879744

Butterfuss, R., Roscoe, R. D., Allen, L. K., McCarthy, K. S., & McNamara, D. S. (2022). Strategy uptake in writing pal: Adaptive feedback and instruction. *Journal of Educational Computing Research*, *60*(3), 696–721. https://doi.org/10.1177/07356331211045304

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Extending cognitive load theory to incorporate working memory resource depletion: Evidence from the spacing effect. *Educational Psychology Review*, *30*(2), 483–501. https://doi.org/10.1007/s10648-017-9426-2

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Chiu, T. K. F. (2023). Student engagement in K-12 online learning amid COVID-19: A qualitative approach from a self-determination theory perspective. *Interactive Learning Environments*, *31*(6), 3326–3339. https://doi.org/10.1080/10494820.2021.1926289

Dai, W., Tsai, Y.-S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, *7*, Article 100299, 1–10. https://doi.org/10.1016/j.caeai.2024.100299

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & de Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, *162*, 104094. https://doi.org/10.1016/j.compedu.2020.104094

Durik, A. M., Shechter, O. G., Noh, M., Rozek, C. S., & Harackiewicz, J. M. (2015). What if I can't? Success expectancies moderate the effects of utility value information on situational interest and performance. *Motivation and Emotion*, *39*(1), 104–118. https://doi.org/10.1007/s11031-014-9419-0

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), 1–20. https://doi.org/10.1186/s41239-023-00425-2

Fidan, M., & Gencel, N. (2022). Supporting the instructional videos with chatbot and peer feedback mechanisms in online learning: The effects on learning performance and intrinsic motivation. *Journal of Educational Computing Research*, *60*(7), 1716–1741. https://doi.org/10.1177/07356331221077901

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Schaeffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/gaise/gaiseprek-12_full.pdf

Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2015). When feedback is cognitively-demanding: The importance of working memory capacity. *Instructional Science*, *43*(1), 73–91. https://doi.org/10.1007/s11251-014-9323-8

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1–25. https://doi.org/10.1111/j.1751-5823.2002.tb00336.x

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. https://doi.org/10.3102/00346543042003237

Gombert, S., Fink, A., Giorgashvili, T., Jivet, I., Di Mitri, D., Yau, J., Frey, A., & Drachsler, H. (2024). From the automated assessment of student essay content to highly informative feedback: A case study. *International Journal of Artificial Intelligence in Education*, *34*, 1378–1416. https://doi.org/10.1007/s40593-023-00387-6

Griendling, L. M., VanUitert, V. J., & McDonald, S. D. (2022). Are students' basic psychological needs fulfilled in remote learning environments? A mixed methods study. *Middle Grades Review*, *8*(2), 1–15. https://eric.ed.gov/?id=ej1364925

Guggemos, J., Moser, L., & Seufert, S. (2022). Learners don't know best: Shedding light on the phenomenon of the K-12 MOOC in the context of information literacy. *Computers & Education*, *188*, 104552. https://doi.org/10.1016/j.compedu.2022.104552

Haller, H., & Kraus, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, *7*(1), 1–20. https://doi.org/10.5283/epub.34338

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

Heston, T., & Khun, C. (2023). Prompt engineering in medical education. *International Medical Education*, *2*(3), 198–205. https://doi.org/10.3390/ime2030019

Hetmanek, A., Engelmann, K., Opitz, A., & Fischer, F. (2018). Beyond intelligence and domain knowledge: Scientific reasoning and argumentation as a set of cross-domain skills. In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation* (pp. 203–226). Routledge.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, *41*(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4

Hidi, S., Renninger, K. A., & Krapp, A. (2004). Interest, a motivational variable that combines affective and cognitive functioning. In D. Y. Dai & R. J. Sternbergm (Eds.), *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development* (pp. 89–115). Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410610515-11

Horbach, A., Laarmann-Quante, R., Liebenow, L., Jansen, T., Keller, S., Meyer, J., Zesch, T., & Fleckenstein, J. (2022). Bringing automatic scoring into the classroom—Measuring the impact of automated analytic feedback on student writing performance. In *Linköping Electronic Conference Proceedings, Proceedings of the 11th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2022)* (pp. 72–83). Linköping University Electronic Press. https://doi.org/10.3384/ecp190008

Hui, Y. K., Li, C., Qian, S., & Kwok, L. F. (2019). Learning engagement via promoting situational interest in a blended learning environment. *Journal of Computing in Higher Education*, *31*(2), 408–425. https://doi.org/10.1007/s12528-019-09216-z

Jacobsen, L. J., & Weber, K. E. (2025). The promises and pitfalls of large language models as feedback providers: A study of prompt engineering and the quality of AI-driven feedback. *AI*, *6*(2), Article 35, 1–17. https://doi.org/10.3390/ai6020035

Jansen, T., Höft, L., Bahr, J. L., Kuklick, L., & Meyer, J. (2025). Constructive feedback can function as a reward: Students' emotional profiles in reaction to feedback perception mediate associations with task interest. *Learning and Instruction*, *95*, 102030. https://doi.org/10.1016/j.learninstruc.2024.102030

Jansen, T., Höft, L., Bahr, L., Fleckenstein, J., Möller, J., Köller, O., & Meyer, J. (2024). Comparing generative AI and expert feedback to students' writing: Insights from student teachers. *Psychologie in Erziehung und Unterricht*, *71*(2), 80–92. https://doi.org/10.2378/peu2024.art08d

Jeon, J. (2022). Exploring a self-directed interactive app for informal EFL learning: A self-determination theory perspective. *Education and Information Technologies*, *27*(4), 5767–5787. https://doi.org/10.1007/s10639-021-10839-y

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Knogler, M., Harackiewicz, J. M., Gegenfurtner, A., & Lewalter, D. (2015). How situational is situational interest? Investigating the longitudinal structure of situational interest. *Contemporary Educational Psychology*, *43*, 39–50. https://doi.org/10.1016/j.cedpsych.2015.08.004

Krapp, A. (2002). An educational-psychological theory of interest and its relation to self-determination theory. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 405–427). University Rochester Press. https://psycnet.apa.org/record/2002-01702-018

Krapp, A. (2005). Basic needs and the development of interest and intrinsic motivational orientations. *Learning and Instruction*, *15*(5), 381–395. https://doi.org/10.1016/j.learninstruc.2005.07.007

Krapp, A., & Prenzel, M. (2011). Research on interest in science: Theories, methods, and findings. *International Journal of Science Education*, *33*(1), 27–50. https://doi.org/10.1080/09500693.2010.518645

Kuklick, L., Greiff, S., & Lindner, M. A. (2023). Computer-based performance feedback: Effects of error message complexity on cognitive, metacognitive, and motivational outcomes. *Computers & Education*, *200*, 104785. https://doi.org/10.1016/j.compedu.2023.104785

Kulacki, A. R., & Aikens, M. L. (2024). Examining motivational attitudes toward statistics and their relationship to performance in life science students. *Journal of Statistics and Data Science Education*, 1–12. https://doi.org/10.1080/26939169.2024.2365892

Lim, L.-A., Gentili, S., Pardo, A., Kovanović, V., Whitelock-Wainwright, A., Gašević, D., & Dawson, S. (2021). What changes, and for whom? A study of the impact of learning analytics-based process feedback in a large course. *Learning and Instruction*, *72*, 101202. https://doi.org/10.1016/j.learninstruc.2019.04.003

Linnenbrink-Garcia, L., Patall, E. A., & Messersmith, E. E. (2013). Antecedents and consequences of situational interest. *The British Journal of Educational Psychology*, *83*(Pt 4), 591–614. https://doi.org/10.1111/j.2044-8279.2012.02080.x

Lloyd, S. A., & Robertson, C. L. (2012). Screencast tutorials enhance student learning of statistics. *Teaching of Psychology*, *39*(1), 67–71. https://doi.org/10.1177/0098628311430640

Ma, B., Chen, L., & Konomi, S. (2024). Enhancing programming education with ChatGPT: A case study on student perceptions and interactions in a python course. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Communications in Computer and Information Science: Vol. 2150. Artificial Intelligence in Education. Posters and Late Breaking Results, workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (Vol. *2150*, pp. 113–126). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64315-6_9

Maier, U., & Klotz, C. (2022). Personalized feedback in digital learning environments: Classification framework and literature review. *Computers and Education: Artificial Intelligence*, *3*, 100080. https://doi.org/10.1016/j.caeai.2022.100080

Mao, P., Cai, Z., Wang, Z., Hao, X., Fan, X., & Sun, X. (2024). The effects of dynamic and static feedback under tasks with different difficulty levels in digital game-based learning. *The Internet and Higher Education*, *60*, 100923. https://doi.org/10.1016/j.iheduc.2023.100923

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, *6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199

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*(2), 56–76. https://doi.org/10.1016/j.compedu.2013.09.011

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education*, *46*(5), 756–778. https://doi.org/10.1080/02602938.2020.1823314

Nikiforidou, Z., Lekka, A., & Pange, J. (2010). Statistical literacy at university level: The current trends. *Procedia - Social and Behavioral Sciences*, *9*, 795–799. https://doi.org/10.1016/j.sbspro.2010.12.236

OpenAI. (2022). *ChatGPT (Version 3.5 Turbo) [Large language model]*. https://chat.openai.com/chat

Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, *52*(3), 275–300. https://doi.org/10.1080/15391523.2020.1719943

Prenzel, M. (1992). The selective persistence of interest. In K. A. Renninger, S. Hidi, A. Krapp, & A. Renninger (Eds.), *The role of interest in learning and development* (pp. 71–98). Lawrence Erlbaum Associates.

Rotgans, J. I., & Schmidt, H. G. (2011a). Situational interest and academic achievement in the active-learning classroom. *Learning and Instruction*, *21*(1), 58–67. https://doi.org/10.1016/j.learninstruc.2009.11.001

Rotgans, J. I., & Schmidt, H. G. (2011b). The role of teachers in facilitating situational interest in an active-learning classroom. *Teaching and Teacher Education*, *27*(1), 37–42. https://doi.org/10.1016/j.tate.2010.06.025

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3), 1–12. https://doi.org/10.1080/10691898.2002.11910678

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction*, *83*, 101620. https://doi.org/10.1016/j.learninstruc.2022.101620

Salikhova, N. R., Lynch, M. F., & Salikhova, A. B. (2020). Psychological aspects of digital learning: A self-determination theory perspective. *Contemporary Educational Technology*, *12*(2), ep280. https://doi.org/10.30935/cedtech/8584

Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist*, *26*(3–4), 299–323. https://doi.org/10.1080/00461520.1991.9653136

Sharma, S. (2017). Definitions and models of statistical literacy: A literature review. *Open Review of Educational Research*, *4*(1), 118–133. https://doi.org/10.1080/23265507.2017.1354313

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

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin*, *46*(2), 137–150. https://doi.org/10.1037/h0062958

Stadler, M., Niepel, C., & Greiff, S. (2019). Differentiating between static and complex problems: A theoretical framework and its empirical validation. *Intelligence*, *72*(185), 1–12. https://doi.org/10.1016/j.intell.2018.11.003

Stamper, J., Xiao, R., & Hou, X. (2024). Enhancing LLM-based feedback: Insights from intelligent tutoring systems and the learning sciences. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Communications in Computer and Information Science: Vol. 2150. Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (Vol. *2150*, pp. 32–43). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64315-6_3

Stark, R., Puhl, T., & Krause, U.-M. (2009). Improving scientific argumentation skills by a problem-based learning environment: Effects of an elaboration tool and relevance of student characteristics. *Evaluation & Research in Education*, *22*(1), 51–68. https://doi.org/10.1080/09500790903082362

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3–46. https://doi.org/10.52041/serj.v2i2.553

Williams, M., Payne, G., Hodgkinson, L., & Poade, D. (2008). Does British sociology count? *Sociology*, *42*(5), 1003–1021. https://doi.org/10.1177/0038038508094576

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

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*, *8*(10), 1839–1850. https://doi.org/10.1038/s41562-024-02004-5

**How to cite this article:** Bauer, E., Richters, C., Pickal, A. J., Klippert, M., Sailer, M., & Stadler, M. (2025). Effects of AI-generated adaptive feedback on statistical skills and interest in statistics: A field experiment in higher education. *British Journal of Educational Technology*, *56*, 1735–1757. https://doi.org/10.1111/bjet.13609