Barkela, V. & Leuchter, M. (2024). Effectiveness of automated formative feedback in an online tutorial for promoting summarizing. *Journal of Educational Technology Development and Exchange*, *17*(1), 67-95. https://doi.org/10.18785/jetde.1701.04

Effectiveness of Automated Formative Feedback in an Online Tutorial for Promoting Summarizing

Veronika Barkela
RPTU Kaiserslautern-Landau
veronika.barkela@rptu.de

Miriam Leuchter
RPTU Kaiserslautern-Landau
miriam.leuchter@rptu.de

Abstract: *We conducted a study with the aim to investigate the effectiveness of automated formative feedback in improving students' ability to summarize. One-hundred and thirty-eight undergraduate students in an elementary education program were asked to summarize six scientific texts, with the experimental group (N=87) receiving automated formative feedback in a computer-based learning environment (FALB). FALB provides automated feedback about content coverage, copying words avoidance, redundancy avoidance, relevance, and length. Comparing the experimental group to a control group (N=51), results implied that summarizing skills could be fostered when interacting with FALB. In particular, the automated formative feedback promoted the adherence to the predefined length and the avoidance of copying words while maintaining a high content coverage, fostering cognitive processes essential for constructing a mental model of a text. In addition, students in the experimental group were able to maintain high quality summaries in their final session when not scaffolded. In conclusion, FALB supports the alignment of internal standards with external standards and provides an incentive to revise and engage with texts.*

Keywords: formative feedback, automated summary evaluator, summary writing, technology-enhanced learning environments, pre-service teacher students

1. Introduction

University students are anticipated to quickly extract and assimilate relevant information from scientific literature and articulate it with precision and conciseness (Kürschner & Schnotz, 2007; van Dijk & Kintsch, 1983). For this purpose, summarizing has been shown to be an adequate learning strategy (Mok & Chan, 2016; Stevens et al., 2019). Summarizing requires understanding a text, identifying relevant aspects, and reflecting on the topic (Perin et al., 2017; Westby et al., 2010). Furthermore, it entails writing a short, concise version of a text, maintaining the key aspects, and formulating them in one's own words, while avoiding redundant and irrelevant parts (Dunlosky et al., 2013; Kirkland & Saunders, 1991). The quality of a summary is highly associated with a person's mental model of the original text (Kim & McCarthy, 2021; Schnotz, 2006). Mental models are representations of a text and encompass both explicitly stated

information from the text and inferences drawn from the text by connecting related information with prior knowledge (van Dijk & Kintsch, 1983). People with more prior knowledge about the topic of a text are more likely to create a comprehensive mental model and write a good summary (Kim et al., 2019).

However, students' representations of how to learn with summaries are often incomplete and, they lack effective skills for creating a mental model of a text and summarizing it (Friend, 2001). They tend to simply copy phrases, forgo reflecting on the content of a text, and refrain from condensing the text to its key aspects, thereby depriving themselves of learning effectively from text (Ahn, 2022; Duke & Pearson, 2009). Such behavior implies that students' internal reference standards of a good summary and how to summarize differ from the external standard of a high quality summary and successful summarizing strategies. Thus, effective summarizing skills do not develop naturally, but must be learned (Ahn, 2022; Keck, 2006). Yet, summarizing strategies are often taught merely in elementary school and not emphasized in later grades, limiting students' ability to use summarizing as an effective learning strategy (McNamara et al., 2019). Therefore, an important endeavor at the university is to teach students effective summarizing strategies to help them succeed in their studies. However, providing learning opportunities that allow students to develop internal reference standards according to an external standard and thus improve their summarizing skills is hardly feasible for large classes with limited resources (Allen et al., 2016). Automated feedback systems for summarizing overcome this dilemma by providing the opportunity to assess many students immediately and as often as desired while meeting premises of effective feedback (Deeva et al., 2021; Strobl et al., 2019).

The field of technology-enhanced learning is undergoing rapid transformation with the integration of generative AI such as ChatGPT (OpenAI et al., 2023) to reshape teaching methods and assessment practices. Nevertheless, a constrained expert system utilizing established natural language processing techniques, like latent semantic analysis, may offer distinct advantages in fostering effective summarizing skills. The development of such a system is less resource-intensive compared to approaches relying on large language models, making it feasible even with limited resources. Furthermore, it enables the creation of a focused system that incorporates an external standard for evaluating students' work, allowing students to align their internal standards accordingly. Such systems have been implemented successfully for the English language (Kim & McCarthy, 2021; Li et al., 2018), French (Lemaire & Dessus, 2001), and Chinese (Sung et al., 2016), demonstrating their effectiveness.

Yet, to the best of the researchers' knowledge, such an automated feedback system for promoting university students' summarizing skills in German is still missing. This study seeks to fill this gap by expanding the evidence on the effectiveness of these systems to new languages, samples, and designs. Specifically, the aim is to assess the supportive potential of a German feedback system designed for undergraduate elementary education students. The researchers intend to scrutinize the key aspects of summarizing skills promoted by the tool and explore the potential association between the formative aspect of feedback and the quality of summaries. The approach involves renewing a German feedback system used in elementary and middle schools, initially focused on reading comprehension (Lenhard et al., 2013). Through this redesign, undergraduate elementary education students can engage in a computer-based learning

environment, fostering their summarizing skills and experiencing a tool that they may later employ as elementary school teachers to enhance reading comprehension in their students.

2. Theoretical Background

Summarizing requires multiple cognitive processes, including integrating new information into one's cognitive schema, determining the relevance of information, constructing a mental model of the text, translating the mental model into one's own words, and ultimately writing it down (Hidi & Anderson, 1986; Perin et al., 2017; Westby et al., 2010). Improved coordination of these processes contributes to individuals' proficiency in constructing comprehensive mental models and generating effective summaries (K. Kim et al., 2019).

Certain task designs support the acquisition of effective summarizing skills to better coordinate cognitive processes. For example, not seeing the text at the same time as writing the summary supports information retrieval and prevents word copying and redundancy (Hidi & Anderson, 1986). Moreover, limiting a summary's length encourages condensing content to key messages and deleting irrelevant information (Hill, 1991). Furthermore, various studies have emphasized the supportive role of formative feedback in encouraging iterative revisions, deep text processing, and adherence to task criteria according to an external standard (Graham, 2018; Kellogg & Raulerson, 2007). Several factors contribute to the effectiveness of feedback (Narciss, 2017; Nixon et al., 2016). For example, students derive greater benefits from immediate rather than delayed feedback (Shute, 2008), exhibit enhanced learning outcomes with elaborate as opposed to simple feedback (Hattie & Timperley, 2007), and better align their internal to an external reference standard when receiving individualized versus general feedback (Zhu et al., 2020). Moreover, effective feedback builds on pre-established assessment criteria and previous performance (Black & Wiliam, 2009). The same applies to automated formative feedback, which has been shown to be as effective and valid as human feedback (Seifried et al., 2012; Stevenson & Phakiti, 2014; van der Kleij et al., 2015).

Research in both offline and online learning environments has implied that the frequency of revising a draft has major impact on its text quality (J. A. Butler & Britt, 2010; Kirkland & Saunders, 1991; Roscoe et al., 2015; Sung et al., 2016). However, inexperienced writers tend to revise scarcely and superficially (Abba et al., 2018). Providing students with automated formative feedback might strengthen students' engagement in the learning process and encourage more revisions, thus supporting the alignment of internal and external reference standards (Link et al., 2020; Liu et al., 2017). Automated formative feedback can be implemented in a way that allows students to control the amount of feedback they receive by letting the algorithm evaluate their drafts as many times as they want. The number of feedback loops can be an indicator of the intensity of revision, thus positively affecting text quality.

The distinction between automated summary evaluators and automated writing evaluation has not always been clear and technological advancements have overlapped. However, for the purposes of this study, the researchers will mainly focus on the development of advancements in the realm of summarizing. Over the years, several automated summary evaluators have been developed. One notable system, *Summary Street* by Wade-Stein & Kintsch (2004) was among

the first to give feedback on summaries to elementary and middle school students, originally designed to enhance text comprehension. They followed a latent semantic analysis approach, which is a natural language processing technique to represent the content of texts. Their English-based computer-based learning environment included source texts about science topics, a text editor for summary composition, bar chart feedback on summary length and section coverage, and a redundancy and relevance check, that listed problematic sentences. The effectiveness of *Summary Street* was investigated using a within subject design with counterbalanced order of conditions. One condition provided feedback on length and spelling, while the other supplemented feedback on content coverage, redundancy, and relevance. The results indicated that automated feedback on content coverage significantly assisted students in enhancing the substance of their summaries. Based on this work, Lenhard et al. (2012) developed a similar system for German elementary school students. Their investigation into the effects of this automated summary evaluator revealed positive impacts on students' reading comprehension and fluency compared to control groups.

Sung et al. (2016) conducted a study with Chinese elementary school students to compare the supportive potential of a summary evaluator providing semantic feedback based on text similarity in one condition (Foltz et al., 1999) and concept feedback based on concept maps in another condition (Schvaneveldt & Cohen, 2010). Results suggested positive effects of both semantic feedback and concept feedback on content coverage of the summaries. Furthermore, a decreasing submission count on the posttest indicated that students learned summarizing skills and did not rely on the support tool. Chew et al. (2019) developed an automated summary evaluator for undergraduate computer science students to learn and practice summarizing in the context of foreign language learning. They included concept maps, worked examples, and feedback on summarizing strategies, demonstrating positive effects on the improvement of the summaries' text quality (rated by teachers) from pretest to posttest. Despite these advancements, we have identified research gaps, specifically in our pursuit of promoting effective summarizing skills to German undergraduate elementary education students.

While existing systems have been predominantly used in school settings (Lenhard et al., 2012; Sung et al., 2016), or if in a university setting, for foreign language learning (Chew et al., 2019), an environment dedicated to explicit practice in processing scientific texts and communicating the information precisely and concisely remains undeveloped. Hence, this study introduces a computer-based learning environment designed to present short German scientific texts related to pedagogical content knowledge to elementary education students. The system offers automated feedback on content coverage and writing style aiming to provide learning opportunities for the development of effective summarizing strategies.

Various automated summary evaluators provide users with information on semantic similarity measures or concept maps of text content (Kim & McCarthy, 2021; Lenhard et al., 2012; Sung et al., 2016). In contrast, this study focuses on developing metrics aligned with the cognitive processes inherent in summarizing, including the identification of relevant information and the creation of a condensed and concise version of the text in one's own words. Therefore, in addition to details about content coverage and length, the automated feedback encompasses information on the avoidance of copied words, redundancy, and irrelevance.

Automated summary evaluators, designed to provide formative feedback, are intended to encourage students to consistently revise their drafts. Sung et al. (2016) utilized the quantity of feedback loops as a metric for tool utilization, indicative student engagement. Yet, to the best of our knowledge, the correlation between an increased number of feedback loops and the generation of higher-quality summaries remains unexplored. Therefore, this study seeks to elucidate the relationship between the number of feedback loops and the quality of summaries.

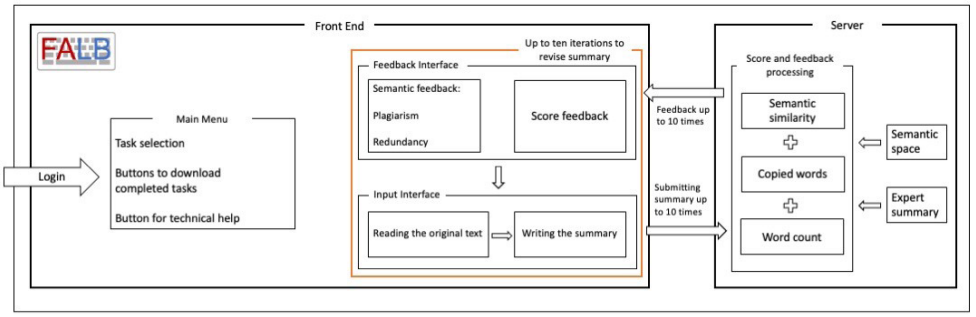
Methodologically, while many studies assess the effectiveness of automated formative feedback through posttest-to-pretest comparisons (Chew et al., 2019) or via case studies (Kim & McCarthy, 2021; Zhang, 2020), the researchers' approach involves evaluating change over multiple time points, employing learning trajectories. These trajectories depict students' probable cognitive developments as they progress in their task (Sztajn et al., 2012). This approach enables teachers to make diagnostic inferences and offer tailored feedback based on the data supplied, even with large sample sizes (Beese, 2019; Plass & Pawar, 2020). In computer-based contexts, understanding learning trajectories can help to anticipate learner behavior at different learning stages, designing customized learning environment, and implementing measures for additional support, such as individualized feedback (Lee & Tan, 2017; Schmid et al., 2022). In the following, we will describe the computer-based learning environment FALB and outline the pedagogical considerations.

3. FALB

The computer-based learning environment *FALB* was developed to provide learning opportunities for developing more sophisticated summarizing skills. It is based on principles of formative assessment (Black & Wiliam, 2009, 2018). *FALB* is composed of two main components (front end and server, Figure 1) which will be explained in the following.

Figure 1

Framework of the Computer-based Learning Environment FALB



3.1. Front End

The front end describes the platform which the user sees and interacts with. A two-page input interface presents the text to be summarized and a text box where the summary can be written

down. The original text is structured in two text sections and intentionally displayed separately from the text box to encourage the creation of a mental model and one's own phrasing (Hidi & Anderson, 1986). Rereading the original text is not limited; the text can be displayed by pressing a button. The feedback Interface displays the feedback and was taken from *conText* (Lenhard et al., 2013) which is based on 'summary street' (Wade-Stein & Kintsch, 2004). Automated formative feedback can be solicited up to ten times per text and session, and entails semantic feedback and score feedback

Semantic feedback provides information about copied words from the original text. Passages are marked as copied and underlined in red when three or more consecutive words are copied from the original text. Additionally, the system provides information about repeated expressions of the same idea (redundancy underlined in blue; see Figure 2) and specifies this information in more detail in a pop-up window, listing corresponding sentences with similar information in the same color. Furthermore, it provides information about irrelevant sentences which should help students stay with the text content (irrelevance underlined in grey). Moreover, unknown words are underlined because they are an indicator of spelling mistakes. Semantic feedback can be obtained by pressing a button labeled "submit text."

Score feedback provides information on how well the original text is covered for the three text sections separately. It also provides information on how well copied words or passages are avoided and how well repeating information is avoided. Moreover, it provides information about the length of the summary, which should not exceed 30% of the original text to obtain the maximum score. All scores are displayed in percentages as horizontal bars. They can be obtained by pressing a button and up to ten times if desired (see Figure 3). In the following, those feedback scores will be referred to as text quality scores.

Figure 2

Example of Semantic Feedback

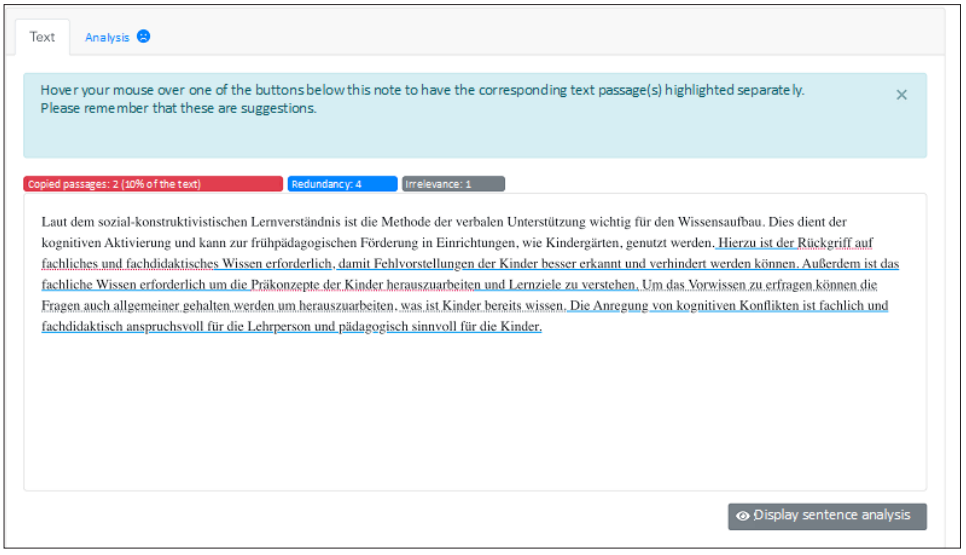
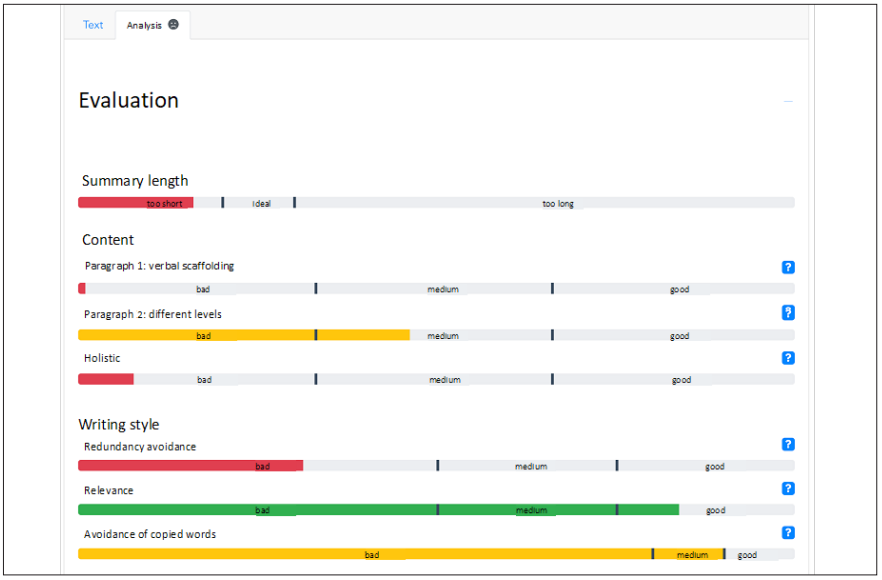


Figure 3

Example of Score Feedback



3.2. Server

The server's main task is to evaluate the summaries and to provide semantic feedback and text quality scores. For this purpose, the original texts, expert summaries, and semantic space were implemented on the server for calculation with Latent Semantic Analysis.

Latent Semantic Analysis (LSA): Text quality scores were determined by LSA, which identifies and sorts words based on their context (Deerwester et al., 1990; Foltz et al., 1999; Landauer et al., 1998; Lenhard et al., 2007). LSA requires a large text corpus (semantic space). LSA projects sentences from the original text, the expert summary, and the student summary into the semantic space and computes a vector for each sentence. Based on the similarity of the vectors, LSA can determine the similarity between pairs of sentences. We used the LSA syntax of conText (Lenhard et al., 2013) and fed the corpus with 201,288 different meaningful German words about learning support in science education through teacher-student interaction.

Text Material: Six excerpts from German scientific texts of comparable length (445-649 words) were used and provided in a fixed order. All texts informed about teacher-student interaction in science education. The texts were selected to be at a similar level of readability, as determined by experts and a readability index (LIX: 66.0 – 80.2; Lenhard & Lenhard, 2014).

Expert Summaries: Two experts prepared summaries of the six texts used in this research. Both experts were specialists in teacher education and academic writing. After they each wrote a first draft, they discussed, revised, and combined the drafts according to the summary criteria (content, avoidance of copied words, avoidance of redundancy, relevance, and length).

Calculation of the Text Quality Scores: Content was calculated by comparing the relation of students' summary content coverage to the original text with the relation of the expert's summary content coverage to the original text. Avoidance of copied words was defined as the ratio of non-copied sentences or phrases to copied sentences or phrases. Redundancy avoidance was calculated based on the number of sentences containing repeated information. Relevance was defined as the ratio of irrelevant sentences to sentences containing relevant information. Length was measured as the ratio of the number of words in the summary to the number of words in the original text. All five text quality scores can theoretically range between 0 and 100%. While scores closer to 100% are desirable for content, avoidance of copied words, redundancy avoidance, and relevance, the optimal score for length is between 20 and 30%.

Summary Result: The text quality scores are interdependent. For example, a summary is more likely to capture the entire content of the original text if it is relatively long. However, the longer the summary, the more likely it is to contain irrelevant and redundant parts. Avoiding copied words or passages is easier if the content can be paraphrased, but at the expense of avoiding redundancy. If more redundancy is avoided, the content must be highly condensed, which may result in less content coverage. Therefore, to make the summaries comparable, the overall summary was calculated with a result that considers all these text quality scores in terms of their importance for the processing of a text. Garner (1982), Head et al. (1989), and Sung et al. (2016) proposed formulas to quantify text quality. Those formulas were modified to include text quality scores and assigned different weights depending on the importance we ascribed to them..

Content is weighted with the factor 0.5, avoidance of copied words is weighted with the factor 0.3, redundancy avoidance is weighted with the factor 0.15, and relevance is weighted with the factor 0.05. All these factors strongly contribute to the creation of mental models that indicate the transformation of the information in the text into individual knowledge (Schnotz, 2006; van Dijk & Kintsch, 1983). Following Lenhard et al. (2013), the optimal length of a summary was set between 20 and 30% of the original text; thus, the length in the formula was the ratio of the length of the students' summaries and the set length limit.

Content, avoidance of copied words, redundancy avoidance, and relevance add up in the formula's counter to display the text quality as a sum value. The high weighting of content derives from the importance of including all essential aspects of the original text in a summary that demonstrates a thorough understanding of the topic. Avoidance of copied words is also highly weighted and shows that the students can express thoughts in their own words. Redundancy avoidance contributes to a brief and concise presentation of the content, which is an essential aspect of a summary. This score is weighted less because it is relative to the length in the denominator. Relevance is weighted lightly because aspects of relevance are covered in the content factor. Additionally, most of the participants scored very high on the relevance score (85.1% > 90), indicating that students generally have no difficulty including relevant information. The longer the summary, the more likely it is that content and avoidance of copied words will score high, and redundancy avoidance and relevance will score low. Therefore, the sum of the content score, the copied word avoidance score, the redundancy avoidance score, and the relevance score is divided by the ratio of the student's summary length to the length limit (cf. Sung et al., 2016). The researchers' formula is as follows:

$$\text{Summary result} = \left(\frac{0.5*CT+0.3*CWA+0.15*RA+0.05*RV}{\frac{SLG}{LGL}} \right) * \frac{100}{150}\% \qquad \text{[Eq. 1]}$$

Students received automated feedback based on percentages derived from the formula as well as three levels (good, satisfactory, needs improvement). To validate the formula, 200 texts were randomly drawn from the sample. Two experts in scientific writing were blindly presented with these summaries and independently rated the summaries according to the summary criteria and three quality levels (good, satisfactory, needs improvement). The interrater reliability measured with Fleiss' kappa between the two experts was κ =.68, between rater 1 and the LSA-based summary scoring was κ =.73, between rater 2 and the LSA-based summary result was κ =.63, and between all three raters was κ =.68, which indicated substantial agreement (Landis & Koch, 1977; Seifried et al., 2012). Consequently, the formula's result satisfactorily represents the quality of the summaries as rated by humans.

4. The Present Research

The researchers examined the effects of automated formative feedback on students' summarizing skills during six sessions implemented in an online university tutorial for undergraduate elementary education students. Reading scientific texts that address core aspects of teaching (e.g., teacher-student interaction) helps students to value scientific literature as a basis for their continuing development in linking theory and practice (Kunina-Habenicht, 2020). However, as shown above, students need support to create a mental model of a text, identify relevant aspects, and summarize effectively (Ahn, 2022; Duke & Pearson, 2009; Friend, 2001). Hence, the researchers expected that providing automated formative feedback embedded in an online tutorial (*FALB*) would support the development of more sophisticated summarizing skills. Furthermore, they expected that the more frequent use of formative feedback would further positively impact summarizing skills. The researchers tested their assumptions by examining the effectiveness of *FALB* with a group that regularly interacted with *FALB* (experimental group) compared to a control group. For this purpose, the researchers of this study formulated three research questions:

(RQ 1) Does the experimental group achieve a higher summary result than a control group?

(RQ 2) Which aspects of summarizing skills (content, avoidance of copied words, redundancy avoidance, relevance, and length) are particularly promoted by the automated formative feedback?

(RQ 3) Do students (experimental group only) who completed more feedback loops write higher quality summaries?

5. Methods

5.1. Participants

A total of 138 cases were included in this study, of which 87 students studied B.Ed. elementary school education (experimental group) and 51 participants studied M.A. special education at the same university (control group). In accordance with the educational curriculum of the bachelor's elementary school education program, participation in the online tutorial was mandatory for all elementary school education students and was worth one credit point. Data for the experimental group were collected in the summer semester 2019. Special education students participated voluntarily in the online tutorial and received one credit point in return. Data for the control group was collected in the summer semester of 2022 which was after three years of intensive online learning due to the COVID-19 pandemic. Therefore, the control group's performance might be at a higher level than if the data had been collected before the pandemic (see limitations).

The participating students were between 19 and 36 years old (M = 23.70, sd = 2.77) and were 84.1% female. Seventy-eight point two percent of the experimental group and 89.6% of the control group had not yet taken a class on teacher-student interaction in science education which was the topic of the texts to be summarized.

5.2. Procedure

As part of an online tutorial, all participants completed a demographic questionnaire and were informed about the criteria of summarizing used in this study. Participants of the experimental group received information on how to decode the automated feedback. Students had to summarize six texts, with two weeks intervals, using the computer-based learning environment *FALB*. In the first and last session, participants of the experimental group submitted their summary but did not receive feedback. For the other four texts, they could write their summary, upload it, and receive automated formative feedback up to ten times. Participants of the control group also had to summarize the same six texts, with one week interval and did not receive any feedback or comments on their summaries (see Appendix A for the curriculum of the tutorials). The difference between the completion times of the two groups had no pedagogical reason but was due to the curricula of the tutorials. However, the researchers did not expect this difference to affect the results, as both groups were instructed to write the summary in 90 minutes without interruption. However, the experimental group could have suffered a slight disadvantage due to the two-week processing time.

The study was approved by the Institutional review board according to faculty regulations. The students provided informed consent for the use of their data. Confidentiality and personal data protection were guaranteed in accordance with relevant data privacy laws.

5.3. Data Analysis

Analysis was conducted using R (R Core Team, 2022), version 4.2.2, rStudio (Posit Team, 2022), the "psych" package (Revelle, 2022) for descriptive and correlational analyses, and the "lme4" (Bates et al., 2015) and "lmerTest" packages (Kuznetsova et al., 2017) to specify multilevel models of change. Tables were prepared with "apaTables" (Stanley, 2021) and models

were drawn with "sjPlot" (Ldecke et al., 2022). Cases with less than 10% content or more than 90% summary result at T0 were removed from the analysis because it either indicated the pretest was not summarized properly or students already possess skills to write high quality summaries (13 cases in the experimental group).

6. Results

Descriptive statistics for the summary result and the correlational analysis are shown in Table 1 and Table 2. The main interest of this study was to analyze students' improvement in summarizing across six time points when receiving automated formative feedback at four time points compared to a no treatment control group.

Table 1

Descriptive Statistics of the Summary Result over Time

			T0		T1		T2		T3		T4		T5	
Summary result	N	Min / Max	M	sd	M	sd	M	sd	M	sd	M	sd	M	sd
Control	51	10/100	50.65	15.40	53.77	16.57	49.57	17.13	47.98	15.29	46.92	14.97	53.06	15.06
Feedback	87		44.28	13.87	60.47	9.56	65.31	13.68	57.33	11.48	62.33	12.29	62.75	13.93

Table 2

Correlational Analysis of the Summary Result

Variable	1	2	3	4	5
1. Summary result_0					
2. Summary result_1	.19*				
3. Summary result_2	.00	.41**			
4. Summary result_3	-.10	.40**	.28**		
5. Summary result_4	.01	.34**	.49**	.39**	
6. Summary result_5	-.01	.27**	.35**	.32**	.22**

Note. * indicates $p < .05$. ** indicates $p < .01$.

First, the researchers checked for a multilevel structure in the data by calculating the intraclass correlation. This revealed that 20.3% of the variance in the summary result over time was explained by individual differences justifying a second level. Hence, a multilevel modeling of change was used with measurement points nested in students to account for interindividual as well as intra-individual change (Singer & Willett, 2003). To identify the optimal model, we tested several models with different functions of time as fixed and random effects using the deviance statistic (see Table 3). If time is included as a fixed effect, the change in the dependent variable is set equal for all individuals. This implies that differences are estimated for individuals' intercepts (e.g., the summary result at T0 in this study), but not for the rate of change. If time is included as

a random effect, both the intercepts and the change in the dependent variable can vary between individuals.

Table 3

Model Comparisons

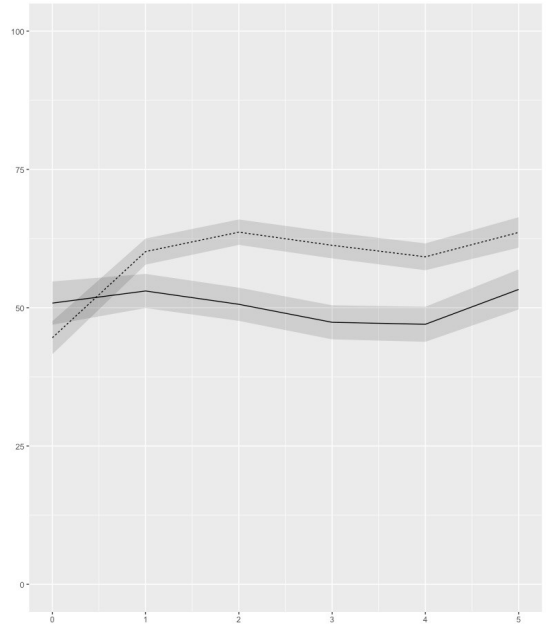
Model	Test of deviance
No time – fixed time	$\chi^2 = 30.38$, $df = 1$, $p < .001$
Fixed time – fixed time^2	$\chi^2 = 13.59$, $df = 1$, $p < .001$
Fixed time^2 – fixed time^3	$\chi^2 = 40.24$, $df = 1$, $p < .001$
Fixed time^3 – fixed time^4	$\chi^2 = 3.71$, $df = 1$, $p = .054$
Fixed time^3 – random time	$\chi^2 = 9.99$, $df = 1$, $p = .007$
Fixed time^3 – random time^2	$\chi^2 = 26.47$, $df = 1$, $p < .001$
Fixed time^3 – random time^3	–

Note. Time^2 = quadratic change in time. Time^3 = cubic change in time. Time^4 = quartic change in time.

Tests of deviances showed that the summary result followed a cubic change (*time*^{3}) and the effects of time differed between individuals. The model with a fixed quartic slope (*time*^{4}) did not explain the data better than the model with a cubic change as a fixed effect ($\chi^2 = 3.71$, $df = 1$, $p = .054$). Thus, the fixed *time*^{3} model was chosen the best and more parsimonious model. Next, we included *group* as a level-1 fixed effect to analyze different rates of change between the experimental group and the control group (see Figure 4).

Figure 4

Predicted Values of Summary Result (y-axis) depending on Time (x-axis) for Control Group (solid) and Experimental Group (dashed)



The model estimates are presented in Table 4. The summary result at T0 differed significantly between the two groups ($\gamma_{\Delta\ control\ -\ feedback}$ = -6.26, p = .014). The control group's summary result did not significantly change from T0 to T2, decreased slightly from T2 to T4, and then increased slightly once more from T4 to T5, thus following a cubic rate of change. On the contrary, the experimental group's summary result increased substantially from T0 to T2, decreased twice as much as the control group from T2 to T4, and then again increased slightly from T4 to T5 at the same rate of change as the control group. Thus, although the experimental group wrote significantly poorer summaries than the control group at T0, the experimental group benefited from the intervention and wrote significantly better summaries than the control group at T5 (RQ 1).

Table 4

Multilevel Model of Change in Summary Result

Fixed effects	***β***	***SE***	***p***
Time	0.64	3.28	.081
$Time^2$	-2.41	1.58	.009
$Time^3$	1.82	0.21	.003
Time ΔFB	1.99	4.13	.000
$Time^2$ ΔFB	-2.86	1.99	.013
$Time^3$ ΔFB	1.18	0.26	.121
Random effects	***Var***	***SD***	
Person	66.70	8.17	
Time	35.91	5.99	
$Time^2$	0.93	0.96	
Level-1 residual	141.54	11.90	
R^2_{total}	.12		

In a next step (RQ 2), we examined the single text quality scores (content, length, avoidance of copied words, redundancy avoidance; see Appendix B for descriptive statistics, model comparisons, and estimates) to better understand which aspects of a summary were particularly promoted by the automated feedback (see Figure 5). We omitted the relevance score from the analysis since more than 85.1% of all summaries had a relevance score over 90, indicating a ceiling effect.

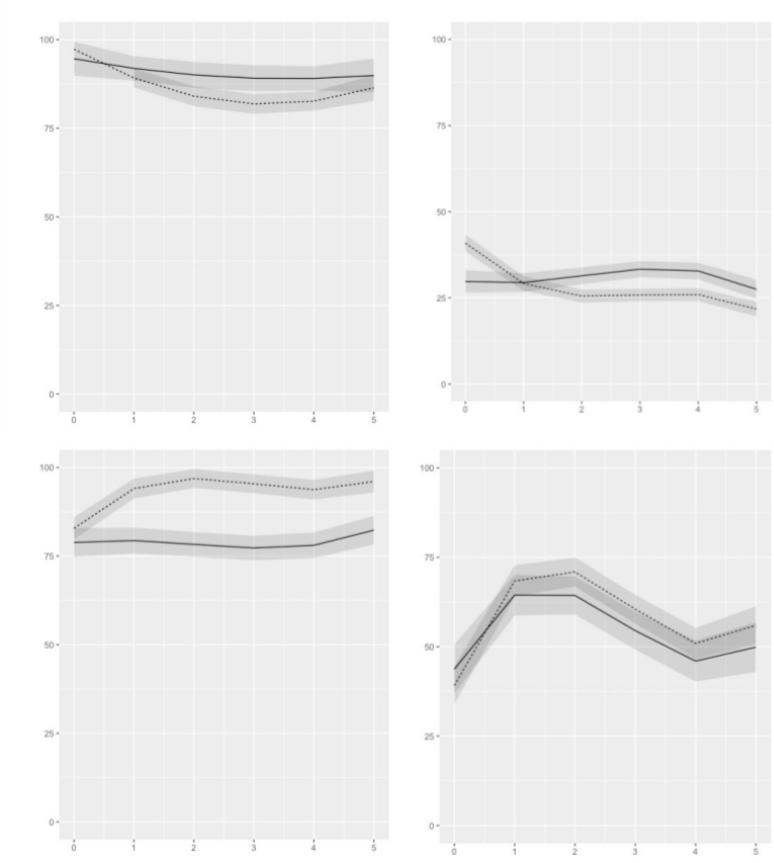
Content followed a quadratic change ($time^2$), but the effects of time did not differ between individuals. Content was over 95% at T0 for both groups indicating a good content coverage. For the control group, the effects of time were not significant and content coverage remained at a high level. For the experimental group, content coverage decreased significantly from T0 to T3 and then increased from T3 to T5 ending at a level around 85%. Length followed a cubic change ($time^3$) and the effects of time differed between individuals. Overall, the length values ranged from just over 40% to just under 25%, slightly outside the intended range of 20 to 30% length of the original text. The control group's length was around 30% at T0. It increased slightly until T3 and decreased from T3 to T5. At T5, the mean length was a little less than 30%. Conversely, the experimental group's length started significantly higher at T0 and decreased until T5, ending with an average length below 25%. *Avoidance of copied* followed a cubic change ($time^3$) and the

effects of time differed between individuals. For the control group, the effects of time were not significant, and avoidance of copied words remained at the same level over time. By contrast, the experimental group reduced the copying of words significantly from T0 to T2, increased slightly from T2 to T4 and decreased again from T4 to T5.

Despite those variations, the level of avoidance of copied words always remained above 90% from T2 to T5. Redundancy avoidance followed a cubic change (time3), but the effects of time did not differ between individuals. Redundancy avoidance increased significantly from T0 to T2, decreased from T2 to T4, and then increased slightly from T4 to T5, all at a lower level than content and avoidance of copied words. No group differences were observed.

Figure 5

Predicted Profiles of Content (top-left), Length (top-right), Avoidance of Copied Words (bottom left), and Redundancy Avoidance (bottom right; all depended constructs on y-axis) depending on Time (x-axis) for Control group (solid) and Experimental group (dashed)



Last, the researchers analyzed the experimental group individually regarding how frequently they used formative feedback (iteration) to improve the summary result (RQ 3; see Appendix C

for descriptive statistics and correlations). The optimal model implied a small significant effect for the summary result on iteration, indicating that students who completed more feedback loops tended to write better summaries (β = 0.14, p = .002; Table 5). Taken together, the intervention with automated formative feedback supported students in writing better summaries, and specifically promoted adherence to the length requirement and avoidance of word copying. Furthermore, more feedback loops resulted in higher quality summaries.

Table 5

Fixed Effects of Summary Result with Time and Iteration as Independent Variables, Experimental Group only

Fixed effects	***β***	***SE***	***p***
Time	8.62	14.33	0.000
$Time^2$	-20.36	6.33	0.000
$Time^3$	13.74	0.84	0.000
Iteration	0.14	0.32	0.002
R^2_{total}	.34		

7. Discussion

The present study was conducted to examine the effectiveness of an online tutorial with automated formative feedback on promoting undergraduate elementary education students' improvement in summarizing. Summarizing skills are important to extract relevant information from scientific texts and to succeed in studying and graduating successfully. Our study showed that summarizing skills can be fostered by automated formative feedback (Kim & McCarthy, 2021; Wade-Stein & Kintsch, 2004).

First, the summary result was analyzed. The summary result is an overall evaluation of the summaries, including the five text quality scores and their relationship to each other. This allows comparing students' overall performance and inferring their improvement in summarizing skills. In the experimental group, during the four-session intervention, feedback highlighted the weaknesses in students' drafts and formatively verified their conformity to the summary criteria. The increase in the summary result for the experimental group from T0 to T1 indicates that the alignment of the internal and external reference standards was particularly strong. However, the summary result continued to be on a high level in the following sessions. In contrast, the summary result of the control group students did not change from T0 to T2, decreased slightly from T2 to T4, and then increased marginally from T4 to T5.

Thus, the formative feedback might have induced the students in the experimental group to regularly evaluate their drafts against the external reference standard and align the external to the internal reference standard (Narciss, 2017). Accordingly, the summaries submitted not only reflect the experimental group students' own ideas of what constitutes a good summary but are also the result of their understanding and internalization of the summary criteria. Furthermore, research has shown that task engagement declines over the course of a seminar (Darby et al.,

2013). However, the feedback may have challenged experimental group students to consistently work at a higher level than control group students. This may suggest that not only the acquisition of effective summarizing skills was promoted, but also motivational regulation (Black & Wiliam, 2018; Clark, 2012). In the last session, students in the experimental group did not receive feedback while composing their summaries. However, the feedback group students' summary result remained at a significantly higher level than the control group students' summary result. The researchers infer that the experimental group students may have transferred feedback insights while composing the final summary and thus achieved better summary results than the control group.

The single text quality scores demonstrate the interdependence of the summary criteria and provide more detailed information about how students have met each summary criterion over time. The longer and more redundant the summary, the higher the possibility of covering full content and avoiding copying words. At T0, students in both the feedback and control groups wrote summaries with a high content coverage, yet they copied more than 20% words from the original text, included more than 50% redundant passages and wrote summaries that were 30% (control group) or longer (experimental group) of the original text. This could indicate that students may not have effectively encoded and integrated textual information into their cognitive schema. In the tutorial, students were expected to write summaries that cover close to 100% of the content, while also scoring high on avoiding copied words and redundant passages as well as adhering to the 20-30% length limit. With this, the researchers intended to stimulate deep processing of the original text and the development of a valid mental model. The profiles of the text quality scores illustrate the learning processes of the experimental group compared to the control group, which only had little rates of change and rather remained at baseline.

At T1, when students in the experimental group received formative feedback for the first time, they aligned their internal reference standard in terms of length and avoidance of copied words by submitting summaries within the optimal range of 20-30% of the original text and avoidance of copied words over 90%. However, the level of content coverage declined. From T2 to T3, while students maintained optimal levels of length and avoidance of copied words, content coverage continued to decrease slightly. Yet, in T4, content coverage increased again while the levels of length and avoidance of copied words remained in the optimal range. This suggested that students in the experimental group improved their summarizing skills over the four-session intervention by learning to coordinate summary criteria requirements and thus, wrote short summaries in their own words while maintaining a high content coverage. These aspects address cognitive processes needed to create a mental model of a text (cf. Dunlosky et al., 2013; Friend, 2001; van Dijk & Kintsch, 1983). Consequently, it might be inferred that they learned to create more valid mental models of the texts. This assumption was further supported by the fact that students not only maintained an optimal length and few copied words throughout the intervention when they had to meet with the criteria, but also maintained these criteria on the final summary when they were not formatively monitored for adherence to the criteria or given formative feedback on their summaries.

The automated formative feedback could not foster redundancy avoidance. At T0, both the control and experimental groups started at a much lower level of redundancy avoidance than content and avoidance of copied words. This was partly because the summaries exceeded the

length. limit, increasing the risk of redundancy. Moreover, it might indicate insufficient skills in writing concisely. Additionally, students may have had unelaborated prior knowledge since 90% of all participants had not yet attended lessons about teacher-student interaction in science education (the topic of the original texts). With little prior knowledge, it is difficult to make inferences, condense the gist of a text, and reorganize its ideas, which hinders the ability to write concise summaries (Kim et al., 2019; van Dijk & Kintsch, 1983). From T1 to T5, group differences were not significant, and redundancy avoidance varied widely across texts, but remained at a lower level than content and avoidance of copied words. In the experimental group, the automated formative feedback fostered the revision phase in the writing process but did not explicitly address the planning phase. As a result, students may not have learned strategies to thoroughly condense the core aspects of the text and sufficiently restructure the ideas of the text; activities that often occur in the planning phase (Chew et al., 2019). Thus, the students were unable to condense the content into a concise summary and avoid redundancy at a higher level. For future designs, it would be worth investigating whether additional prompts that specifically promote the planning phase and activation of prior knowledge could help students develop strategies to better avoid redundancy. In addition, students may have had difficulty following the formative feedback on redundancy avoidance and understanding why some passages were marked as redundant. Thus, they may not have been able to benefit from the feedback.

Automated formative feedback can immediately provide valid feedback to almost an unlimited number of students (Lenhard, 2008; Seifried et al., 2012). In this study, more feedback loops (iterations) positively impacted the summary result. This observation supports findings from previous research that the frequency of revisions highly influences the quality of a summary (Kirkland & Saunders, 1991; Link et al., 2020; Roscoe et al., 2015; Zhu et al., 2020). Students in the experimental group who completed multiple feedback loops might have engaged more deeply in summarizing than students who sought less feedback (Zhang, 2020). These students may also have had more sophisticated skills in seeking and processing feedback (Narciss, 2017). They could have judged external feedback as relevant, understood it, and accepted it (Brown et al., 2016). Thus, they might have been willing to change their internal standards rather than reinterpret the automated feedback according to their internal standards (Butler & Winne, 2016). It would be beneficial to conduct a subsequent study to examine more closely how students engage with automated formative feedback.

8. Limitation

First, the data collection period of the two groups was far apart. Data collection for the feedback group was in 2019 and for the control group in 2022. Between these years, the COVID-19 pandemic occurred, which greatly changed teaching at universities by offering many courses online. Therefore, these groups are comparable to a limited extent, as the control group is more accustomed to a fully online tutorial. Compared to the students in the feedback group, the control group may have been less distracted from reading and writing the texts and their summaries on the computer. They may also have had more practice in summarizing because they may have had to document their work progress more frequently for other courses. This is also reflected in the control group's higher text quality at T0.

Second, the sample of the present study consists of solely elementary and special education students of one German university. Therefore, the researchers do not know to what extent the results can be generalized to other university student populations and academic settings. For future studies, the sample could be expanded to other educational programs and student populations to evaluate the generalizability of these findings beyond the scope of elementary teacher education. Furthermore, a longitudinal study over several semesters could provide insight into the long-term effects of automated formative feedback on students' academic growth and skill retention.

Third, FALB is a recently developed computer-based learning environment. Therefore, it has not yet been evaluated in terms of learning and feedback experience, and usability. Research has shown that satisfaction (Doménech-Betoret et al., 2017), feedback acceptance (Seifried et al., 2016), and technology acceptance (Hanham et al., 2021) are highly associated with learning success in computer-based learning environments. Thus, future studies should consider these moderating variables and how they affect feedback engagement and learning outcomes.

Fourth, this study lacks control variables like prior knowledge, language capability, and time on task. Research has shown that prior knowledge influences the creation of mental models (Kim et al., 2019; van Dijk & Kintsch, 1983), limited language capability may increase mental load, reduce reading comprehension, and shift attention (Li, 2014; McCutchen, 2011), and time on task is a strong predictor of text quality (Butler & Britt, 2010). Therefore, in future studies, such variables should be controlled to further explain interindividual differences in working with *FAL*B.

9. Conclusion

Overall, this study provides valuable insight into the use of automated formative feedback through latent semantic analysis to teach effective summarizing skills to German university students. Furthermore, it highlights the necessity of imparting summarizing as an effective learning strategy and teaching effective summarizing strategies to students. The researchers aimed to extend the evidence on automated summary evaluators through a new language, sample, and design. The findings demonstrated the potential of *FALB* to support the development of more sophisticated summarizing skills in German undergraduate elementary education students. They indicated that *FALB* particularly encouraged short summaries without copied words, with the formative nature of the feedback contributing to improved text quality. Yet, redundancy avoidance was promoted only to a limited extent, potentially attributable to the insufficient emphasis on the planning phase within the computer-based learning environment. Future research could advance these insights by investigating the efficacy of additional prompts during the planning phase in fostering enhanced redundancy avoidance and the adoption of effective summarizing strategies. Notably, the research by van den Boom et al. (2004, 2007) insinuates a complementary effect of prompting and feedback that remains unexplored in computer-based learning environments with automated formative feedback on summarizing. Taken together, the insights gained from this study on the support mechanisms of *FALB* are valuable and contribute to enhancing the design of intelligent feedback systems for university applications..

Acknowledgement

We express our gratitude to Prof. Dr. Wolfgang Lenhard for conText, for his support in the development of FALB, and the enriching exchange. Additionally, we extend our thanks to Matthias Barde for the programming.

This project is part of the "Qualitaetsoffensive Lehrerbildung," a joint initiative of the Federal Government and the Laender that aims to improve the quality of teacher training [grant number 01JA2016]. The program is funded by the Federal Ministry of Education and Research, Germany. The authors are responsible for the content of this publication.

Author Information

Veronika Barkela
Department of Child and Adolescent Education
RPTU Kaiserslautern-Landau
August-Croissant-Str. 5, 76829 Landau, Germany
Telephone: 0049-6341-280 34 152
Telefax: 0049-6341-280 34 131
https://orcid.org/0000-0002-9704-1153
E-mail: veronika.barkela@rptu.de

Miriam Leuchter
https://orcid.org/0000-0002-7962-6561
miriam.leuchter@rptu.de

References

Abba, K. A., Zhang, S. S., & Joshi, R. M. (2018). Community college writers' metaknowledge of effective writing. *Journal of Writing Research, 10*(1), 85–105. https://doi.org/10.17239/jowr-2018.10.01.04

Ahn, S. (2022). Developing summary writing abilities of Korean EFL university students through teaching summarizing skills. *English Teaching, 77*(2), 25–43. https://doi.org/10.15858/engtea.77.2.202206.25

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), H*andbook of Writing Research* (second edition, pp. 316–329). Guildford Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Beese, E. B. (2019). A process perspective on research and design issues in educational

personalization. *Theory and Research in Education, 17*(3), 253–279. https://doi.org/10.1177/1477878519893963

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

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. A*ssessment in Education: Principles, Policy & Practice, 25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Brown, G. T. L., Peterson, E. R., & Yao, E. S. (2016). Student conceptions of feedback: Impact on self-regulation, self-efficacy, and academic achievement. *The British Journal of Educational Psychology, 86*(4), 606–629. https://doi.org/10.1111/bjep.12126

Butler, D. L., & Winne, P. H. (2016). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Butler, J. A., & Britt, M. A. (2010). Investigating instruction for improving revision of argumentative essays. *Written Communication, 28*(1), 70–96. https://doi.org/10.1177/0741088310387891

Chew, C. S., Idris, N., Loh, E. F., Wu, W. V., Chua, Y. P., & Bimba, A. T. (2019). The effects of a theory-based summary writing tool on students' summary writing. *Journal of Computer Assisted Learning, 35*(3), 435–449. https://doi.org/10.1111/jcal.12349

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review, 24*(2), 205–249. https://doi.org/10.1007/s10648-011-9191-6

Darby, A., Longmire-Avital, B., Chenault, J., & Haglund, M. (2013). Students' motivation in academic service-learning over the course of the semester. *College Student Journal, 47*(1), 185–191.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

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

Doménech-Betoret, F., Abellán-Roselló, L., & Gómez-Artiga, A. (2017). Self-efficacy, satisfaction, and academic achievement: The mediator role of students' expectancy-value beliefs. *Frontiers in Psychology, 8*, 1193. https://doi.org/10.3389/fpsyg.2017.01193

Duke, N. K., & Pearson, P. D. (2009). Effective practices for developing reading comprehension. *Journal of Education, 189*(1–2), 107–122. https://doi.org/10.1177/0022057409189001-208

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

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 939–944.

Friend, R. (2001). Effects of strategy instruction on summary writing of college students.

Contemporary Educational Psychology, 26(1), 3–24. https://doi.org/10.1006/ceps.1999.1022

Garner, R. (1982). Efficient text summarization costs and benefits. *The Journal of Educational Research, 75*(5), 275–279. https://doi.org/10.1080/00220671.1982.10885394

Graham, S. (2018). Instructional feedback in writing. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge Handbook of Instructional Feedback* (1st ed., pp. 145–168). Cambridge University Press. https://doi.org/10.1017/9781316832134.009

Hanham, J., Lee, C. B., & Teo, T. (2021). The influence of technology acceptance, academic self-efficacy, and gender on academic achievement through online tutoring. *Computers & Education, 172,* 104252. https://doi.org/10.1016/j.compedu.2021.104252

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

Head, M. H., Readence, J. E., & Buss, R. R. (1989). An examination of summary writing as a measure of reading comprehension. *Reading Research and Instruction, 28*(4), 1–11. https://doi.org/10.1080/19388078909557982

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research, 56*(4), 473–493. https://doi.org/10.3102/00346543056004473

Hill, M. (1991). Writing summaries promotes thinking and learning across the curriculum—But why are they so difficult to write? *Journal of Reading, 34*(7), 536–539. http://www.jstor.org/stable/40014578

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing, 15*(4), 261–278. https://doi.org/10.1016/j.jslw.2006.09.006

Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review, 14*(2), 237–242. https://doi.org/10.3758/BF03194058

Kim, K., Clarianay, R. B., & Kim, Y. (2019). Automatic representation of knowledge structure: Enhancing learning through knowledge structure reflection in an online course. *Educational Technology Research and Development, 67*(1), 105–122. https://doi.org/10.1007/s11423-018-9626-6

Kim, M. K., & McCarthy, K. S. (2021). Improving summary writing through formative feedback in a technology-enhanced learning environment. *Journal of Computer Assisted Learning, 37*(3), 684–704. https://doi.org/10.1111/jcal.12516

Kirkland, M. R., & Saunders, M. A. P. (1991). Maximizing student performance in summary writing: Managing cognitive load. *TESOL Quarterly, 25*(1), 105–121. https://doi.org/10.2307/3587030

Kunina-Habenicht, O. (2020). Wissen ist Macht: Ein Plädoyer für ein wissenschaftliches Lehramtsstudium. In C. Scheid & T. Wenzl (Eds.), *Wieviel Wissenschaft braucht die Lehrerbildung?* (pp. 109–126). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-23244-3_6

Kürschner, C., & Schnotz, W. (2007). Konstruktion mentaler Repräsentationen bei der Verarbeitung von Text und Bildern. *Unterrichtswissenschaft, 35*(1), 48–67. https://doi.org/10.25656/01:5486

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. J*ournal of Statistical Software, 82*(13). https://doi.org/10.18637/jss.v082.i13

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*(2 & 3), 259–284. https://doi.org/10.1080/01638539809545028

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lee, A. V. Y., & Tan, S. C. (2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics, 4*(3). https://doi.org/10.18608/jla.2017.43.5

Lemaire, B., & Dessus, P. (2001). A system to assess the semantic content of student essays. *Journal of Educational Computing Research, 24*(3), 305–320. https://doi.org/10.2190/G649-0R9C-C021-P6X3

Lenhard, W. (2008). Bridging the gap to natural language: A review on intelligent tutoring systems based on Latent Semantic Analysis. https://opus.bibliothek.uniwuerzburg.de/files/2397/Lenhard_Bridging_the_Gap.pdf

Lenhard, W., & Lenhard, A. (2014). *Berechnung des Lesbarkeitsindex LIX nach Björnson.* Unpublished. https://doi.org/10.13140/RG.2.1.1512.3447

Lenhard, W., Baier, H., Endlich, D., Lenhard, A., Schneider, W., & Hoffmann, J. (2012). Computerunterstützte Leseverständnisförderung: Die Effekte automatisch generierter Rückmeldungen. *Zeitschrift Für Pädagogische Psychologie, 26*(2), 135–148. https://doi.org/10.1024/1010-0652/a000066

Lenhard, W., Baier, H., Hoffmann, J., & Schneider, W. (2007). Automatische Bewertung offener Antworten mittels Latenter Semantischer Analyse. *Diagnostica, 53*(3), 155–165. https://doi.org/10.1026/0012-1924.53.3.155

Lenhard, W., Baier, H., Lenhard, A., Hoffmann, J., & Schneider, W. (2013). *ConText: Förderung des Leseverständnisses durch das Arbeiten mit Texten: Manual*. Hogrefe.

Li, H., Cai, Z., & Graesser, A. C. (2018). Computerized summary scoring: Crowdsourcing-based latent semantic analysis. *Behavior Research Methods, 50*(5), 2144–2161. https://doi.org/10.3758/s13428-017-0982-7

Li, J. (2014). The role of reading and writing in summarization as an integrated task. *Language Testing in Asia, 4*(1), 3. https://doi.org/10.1186/2229-0443-4-3

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning, 35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323

Liu, M., Li, Y., Xu, W., & Liu, L. (2017). Automated essay feedback generation and its impact on revision. *IEEE Transactions on Learning Technologies, 10*(4), 502–513. https://doi.org/10.1109/tlt.2016.2612659

Lüdecke, D. (2022). sjPlot: Data visualization for statistics in social science. R package version 2.8.12, https://CRAN.R-project.org/package=sjPlot

McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Writing Research, 3*(1), 51–68.

McNamara, D. S., Roscoe, R., Allen, L., Balyan, R., & McCarthy, K. S. (2019). Literacy: From the perspective of text and discourse theory. *Journal of Language and Education, 5*(3), 56–69. https://doi.org/10.17323/jle.2019.10196

Mok, W. S. Y., & Chan, W. W. L. (2016). How do tests and summary writing tasks enhance long-term retention of students with different levels of test anxiety? I*nstructional Science, 44*(6),

567–581. https://doi.org/10.1007/s11251-016-9393-x

Narciss, S. (2017). Conditions and effects of feedback viewed through the lens of the interactive tutoring feedback model. In D. Carless, S. M. Bridges, C. K. Y. Chan, & R. Glofcheski (Eds.), *Scaling up Assessment for Learning in Higher Education* (Vol. 5, pp. 173–189). Springer Singapore.

Nixon, S., Brooman, S., Murphy, B., & Fearon, D. (2016). Clarity, consistency, and communication: Using enhanced dialogue to create a course-based feedback strategy. *Assessment & Evaluation in Higher Education, 42*(5), 812–822. https://doi.org/10.1080/02602938.2016.1195333

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). GPT-4 Technical Report. https://doi.org/10.48550/ARXIV.2303.08774

Perin, D., Lauterbach, M., Raufman, J., & Kalamkarian, H. S. (2017). Text-based writing of low-skilled postsecondary students: Relation to comprehension, self-efficacy, and teacher judgments. *Reading and Writing, 30*(4), 887–915. https://doi.org/10.1007/s11145-016-9706-0

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

Posit Team (2022). *RStudio: Integrated development environment for R*. Posit Software, PBC, Boston, MA. http://www.posit.co/

R Core Team (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Revelle, W. (2022). Package "psych." *The Comprehensive R Archive Network, 337*, 1–465.

Roscoe, R. D., Snow, E. L., Allen, L. K., & McNamara, D. S. (2015). Automated detection of essay revising patterns: Applications for intelligent feedback in a writing tutor. *Technology, Instruction, Cognition and Learning, 10*, 59–79.

Schmid, R., Pauli, C., Stebler, R., Reusser, K., & Petko, D. (2022). Implementation of technology-supported personalized learning—Its impact on instructional quality. *The Journal of Educational Research,* 1–12. https://doi.org/10.1080/00220671.2022.2089086

Schnotz, W. (2006). Was geschieht im Kopf des Lesers? Mentale Konstruktionsprozesse beim Textverstehen aus der Sicht der Psychologie und der kognitiven Linguistik. *Text-Verstehen. Grammatik und darüber hinaus*, 222–238.

Schvaneveldt, R. W., & Cohen, T. A. (2010). Abductive reasoning and similarity: Some computational tools. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based Diagnostics and Systematic Analysis of Knowledge* (pp. 189–211). Boston, MA: Springer US.

Seifried, E., Lenhard, W., & Spinath, B. (2016). Automatic essay assessment: Effects on students' acceptance and on learning-related characteristics. *Psihologija, 49*(4), 469–482. https://doi.org/10.2298/PSI1604469S

Seifried, E., Lenhard, W., Baier, H., & Spinath, B. (2012). On the reliability and validity of human and LSA-based evaluations of complex student-authored texts. *Journal of Educational Computing Research, 47*(1), 67–92. https://doi.org/10.2190/EC.47.1.d

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

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Stanley, D. (2021). *apaTables: Create American psychological association (APA) style tables.* R package version 2.0.8, https://CRAN.R-project.org/package=apaTables

Stevens, E. A., Park, S., & Vaughn, S. (2019). A review of summarizing and main idea interventions for struggling readers in grades 3 through 12: 1978–2016. *Remedial and Special Education, 40*(3), 131–149. https://doi.org/10.1177/0741932517749940

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19,* 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education, 131*, 33–48. https://doi.org/10.1016/j.compedu.2018.12.005

Sung, Y.-T., Liao, C.-N., Chang, T.-H., Chen, C.-L., & Chang, K.-E. (2016). The effect of online summary assessment and feedback system on the summary writing on 6th graders: *The LSA-based technique. Computers & Education, 95*, 1–18. https://doi.org/10.1016/j.compedu.2015.12.003

Sztajn, P., Confrey, J., Wilson, P. H., & Edgington, C. (2012). Learning trajectory based instruction: Toward a theory of teaching. *Educational Researcher, 41*(5), 147–156. https://doi.org/10.3102/0013189X12442801

van den Boom, G., Paas, F. G. W. C., & van Merriënboer, J. J. G. (2007). Effects of elicited reflections combined with tutor or peer feedback on self-regulated learning and learning outcomes. *Learning and Instruction, 17*(5), 532–548. https://doi.org/10.1016/j.learninstruc.2007.09.003

van den Boom, G., Paas, F., van Merriënboer, J. J. G., & van Gog, T. (2004). Reflection prompts and tutor feedback in a web-based learning environment: Effects on students' self-regulated learning competence. *Computers in Human Behavior, 20*(4), 551–567. https://doi.org/10.1016/j.chb.2003.10.001

van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475–511. https://doi.org/10.3102/0034654314564881

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* Academic Press.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*(3), 333–362. https://doi.org/10.1207/s1532690xci2203_3

Westby, C., Culatta, B., Lawrence, B., & Hall-Kenyon, K. (2010). Summarizing expository texts. *Topics in Language Disorders, 30*(4), 275–287. https://doi.org/10.1097/TLD.0b013e3181ff5a88

Zhang, Z. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing, 43*, 100439. https://doi.org/10.1016/j.asw.2019.100439

Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education, 143*, 103668. https://doi.org/10.1016/j.compedu.2019.103668

Appendices

Appendix A: Additional Information on FALB

Table A1

List of Text Material

Time	Text material	Word count	LIX
T0	Kleickmann, T.; Hardy, I.; Möller, K.; Pollmeier, J.; Tröbst, S. & Beinbrech, C. (2010). Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter: Theoretische Konzeption und Testkonstruktion. Zeitschrift für Didaktik der Naturwissenschaften, 16; 268 – 269.	637	73,9
T1	Giest, H. (2015). Methodisches Erschließen. In: Kahlert, J.; Fölling-Albers, M.; Götz, M.; Hartinger, A.; Miller, S.; Wittkowske, S. (Hrsg.). Handbuch Didaktik des Sachunterrichts. Bad Heilbrunn: Verlag Julius Klinkhardt.	445	80,2
T2	Sodian B., Mayer D. (2013) Entwicklung des wissenschaftlichen Denkens im Vor- und Grundschulalter. In: Stamm M., Edelmann D. (eds) Handbuch frühkindliche Bildungsforschung. Springer VS, Wiesbaden.	531	66,0
T3	Leuchter, M. & Saalbach, H. (2014). Verbale Unterstützungsmaßnahmen im Rahmen eines naturwissenschaftlichen Lernangebots in Kindergarten und Grundschule. Unterrichtswissenschaft, 42(2), 117-131.	649	67,2
T4	Klieme, E.; Bürgermeister, A; Harks, B.; Blum, W.; Leiß. D & Rakoczy, K. (2010). Leistungsbeurteilung und Kompetenzmodellierung im Mathematikunterricht. Projekt Co2CA. Zeitschrift für Pädagogik, Beiheft; 56. Weinheim; Basel: Beltz.	618	74,6
T5	Möller, K., Steffensky, M. (2010). Naturwissenschaftliches Lernen im Unterricht mit 4- bis 8-jährigen Kindern. Kompetenzbereiche frühen naturwissenschaftlichen Lernens. In M. Leuchter (Ed.), Didaktik für die ersten Bildungsjahre. Unterricht mit 4- bis 8-jährigen Kindern. Seelze: Friedrich Verlag.	638	71,7

Table A2

Curriculum of the Experimental Group's Online Tutorial

Week	Assignment
1 + 2	Summarize Text 1 in FALB
3 + 4	Summarize Text 2 and interact with FALB
5 + 6	Summarize Text 3 and interact with FALB
7 + 8	Summarize Text 4 and interact with FALB
9 + 10	Summarize Text 5 and interact with FALB
11 + 12	Summarize Text 6 in FALB

Table A3

Curriculum of the Control Group's Online Tutorial

Week	Assignment
1	Summarize Text 1
2	Summarize Text 2
3	Summarize Text 3
4	Summarize Text 4
5	Summarize Text 5
6	Summarize Text 6
7	Recognizing persuasive arguments
8	Identifying persuasive argument structures
9	Write an argumentative essay in the field of pedagogy
10	Write an argumentative essay in the field of sustainability
11	Argumentation based on the Toulmin model
12	Recognizing and formulating persuasive arguments

Appendix B: Additional Information on RQ 2

Table B1

Descriptive Statistics of Content, Length, Avoidance of Copied Words, Redundancy Avoidance

Variable	*min./ max.*	T0		T1		T2		T3		T4		T5	
	0/100	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1. Content	Control	93.50	18.65	91.42	20.32	98.06	7.67	78.96	29.25	92.71	19.57	89.93	18.80
	Feedback	98.25	6.49	84.19	15.56	94.63	8.23	71.22	21.6	87.97	20.54	85.34	17.94
2. Length	Control	30.07	13.84	27.48	12.2	36.25	14.11	27.54	14.24	36.16	15.05	26.74	12.01
	Feedback	41.13	14.64	28.20	5.93	27.05	7.70	24.87	4.82	26.24	5.49	21.77	5.31
3. Avoid. cop. words	Control	78.47	19.31	81.07	18.33	75.04	20.66	80.42	21.97	76.54	18.96	82.59	20.80
	Feedback	82.51	16.15	95.41	6.90	94.70	8.51	96.97	5.28	93.25	10.08	96.07	7.49
4. Red. avoid	Control	42.16	25.67	68.95	29.56	61.74	24.04	50.60	27.75	51.14	26.77	47.96	29.39
	Feedback	38.05	19.80	70.56	24.17	73.16	23.66	51.61	24.77	58.65	24.19	53.77	24.84

Table B2

Model comparisons for Content

Model	Test of deviance
No time – fixed time	$\chi^2 = 24.29$, df = 1, p = .000
Fixed time – fixed time^2	$\chi^2 = 21.67$, df = 1, p = .000
Fixed time^2 – fixed time^3	$\chi^2 = 0.97$, df = 1, p = .324
Fixed time^2 – random time	-

Table B3

Model Comparisons for Length

Model	Test of deviance
No time – fixed time	$\chi^2 = 93.66$, df = 1, p = .000
Fixed time – fixed time^2	$\chi^2 = 7.88$, df = 1, p = .005
Fixed Time^2 – fixed time^3	$\chi^2 = 36.68$, df = 1, p = .000
Fixed time^3 – random time	$\chi^2 = 35.68$, df = 1, p = .000
Fixed time^3 – random time^2	-

Table B4

Model Comparisons for Avoidance of Copied Words

Model	Test of deviance
No time – fixed time	$\chi^2 = 37.73$, df = 1, p = .000
Fixed time – fixed time^2	$\chi^2 = 14.31$, df = 1, p = .000
Fixed Time^2 – fixed time^3	$\chi^2 = 26.42$, df = 1, p = .000
Fixed time^3 – random time	$\chi^2 = 19.18$, df = 1, p = .000
Fixed time^3 – random time^2	-

Table B5

Model Comparisons for Redundancy Avoidance

Model	Test of deviance
No time – fixed time	$\chi^2 = 0.00$, df = 1, p = .960
No time – fixed time^2	$\chi^2 = 61.97$, df = 1, p = .000
Fixed Time^2 – fixed time^3	$\chi^2 = 75.00$, df = 1, p = .000
Fixed Time^3 – random time	-

Table B6

Multilevel Model of Change in Content

Fixed effects	***Estimates***	***SE***	***p***
Intercept (control)	94.61	2.43	.000
Time (control)	-3.17	1.99	.112
Time^2 (control)	0.45	0.38	.245
Intercept ΔFB	2.68	3.06	.382
Time ΔFB	-6.42	2.51	.011
Time^2 ΔFB	1.04	0.48	.032
Random Effects	***Var***	***SD***	
Person	72.95	8.54	
Level-1 residual	278.72	16.70	

Table B7

Multilevel Model of Change in Length

Fixed effects	***Estimates***	***SE***	***p***
Intercept (control)	29.78	1.64	.000
Time (control)	-2.28	2.08	.272
$Time^2$ (control)	2.34	1.03	.023
$Time^3$ (control)	-0.39	0.13	.004
Intercept ΔFB	11.12	2.07	.000
Time ΔFB	-14.87	2.62	.000
$Time^2$ ΔFB	3.78	1.30	.004
$Time^3$ ΔFB	-0.30	0.17	.083
Random Effects	***Var***	***SD***	
Person	79.77	8.93	
Time	2.37	1.54	
Level-1 residual	60.20	7.76	

Table B8

Multilevel Model of Change in Avoidance of Copied Words

Fixed effects	***Estimates***	***SE***	***p***
Intercept (control)	78.82	2.09	.000
Time (control)	1.95	2.37	.410
$Time^2$ (control)	-1.68	1.17	.152
$Time^3$ (control)	0.29	0.15	.064
Intercept ΔFB	4.01	2.63	.127
Time ΔFB	14.85	2.99	.000
$Time^2$ ΔFB	-4.58	1.48	.002
$Time^3$ ΔFB	0.40	0.19	.039
Random Effects	***Var***	***SD***	
Person	147.01	12.13	
Time	3.34	1.83	
Level-1 residual	78.15	8.84	

Table B9

Multilevel Model of Change in Redundancy Avoidance

Fixed Effects	***Estimates***	***SE***	***p***
Intercept (control)	43.71	3.45	.000
Time (control)	34.82	6.04	.000
$Time^2$ (control)	-15.95	3.00	.000

Time^3 (control)	1.85	0.39	.000
Intercept ΔFB	-4.56	4.41	.302
Time ΔFB	12.37	7.60	.104
Time^2 ΔFB	-4.30	3.78	.255
Time^3 ΔFB	0.45	0.50	.364
Random Effects	***Var***	***SD***	
Person	132.37	11.51	
Level-1 residual	512.60	22.64	

Appendix C: Additional Information on RQ 3

Table C10

Means, Standard Deviations, and Correlations of Summary Result and Iteration

Variable	***M***	***SD***	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. tq_1	60.47	9.56							
2. tq_2	65.31	13.68	.46**						
3. tq_3	57.33	11.48	.24*	.10					
4. tq_4	62.33	12.29	.23*	.14	.26*				
5. iteration_1	2.39	1.77	.25*	.16	.23*	.28**			
6. iteration_2	2.53	2.08	.19	.18	.21	.05	.55**		
7. iteration_3	2.71	2.31	.23*	.11	.26*	.24*	.40**	.43**	
8. iteration_4	2.47	2.10	.24*	.33**	.20	.21	.48**	.43**	.31**

Note. * indicates $p < .05$. ** indicates $p < .01$.

Table C11

Multilevel Model of Change in Summary Result with Feedback Group only and Iteration as Independent Variable

Fixed effects	***Estimates***	***SE***	***p***
Time	72.08	14.33	.000
Time^2	-32.67	6.33	.000
Time^3	4.37	0.84	.000
Iteration	0.98	0.32	.000
Random Effects	***Var***	***SD***	
Person	23.62	4.86	
Level-1 residual	110.61	10.52	