Learning to write syntheses: the effect of process feedback and of observing models on performance and process behaviors

Nina Vandermeulen[1,2] · Elke Van Steendam[3] · Sven De Maeyer[4] ·
Marije Lesterhuis[5] · Gert Rijlaarsdam[6,7]

Abstract

Writing a synthesis text involves interacting reading and writing processes, serving the comprehension of source information, and its integration into a reader-friendly and accurate synthesis text. Mastering these processes requires insight into process' orchestrations. A way of achieving this is via process feedback in which students compare their process orchestration with examples. Access to such examples of enacted process orchestration models might have an additional learning effect. In the present study we replicated and extended the study of Vandermeulen et al. (*Written Communication*, *40*(1), 90–144, 2023) on the effect of keystroke logging data-based process feedback with feed-forward exemplars when compared to national baseline performances. In addition, we report the effect of a brief extension in which learners had the opportunity to observe an enacted model of their choice, showing one of three orchestrations of the initial stage of writing a synthesis task. A total of 173 10th—grade students were randomly assigned to a process feedback condition with or without added models. A baseline, consisting of a nationally representative sample of upper-secondary students' texts and processes, served as an alternative control group. Results showed that the process feedback, both with and without observation, had a significant effect on text quality. Regarding the process data, students in the feedback condition had a more prominent focus on the sources as they spent more time in them and switched more often between text and sources, compared to the baseline. The observation task magnified this effect.

Keywords Feedback · Writing process · Keystroke logging · Synthesis writing · Feedback instruction · Learning by exemplars · Compare and contrast · Observational learning

Extended author information available on the last page of the article

Introduction

Developing effective writing skills is indispensable for students' academic success (Graham et al., 2013). In both the research and educational work field, the importance of high-quality writing instruction is recognized (De Smedt et al., 2016). Writing instruction supports students in developing the highly complex skill that is writing. Over the last decades, the writing process has received more attention in several instructional approaches, such as strategy-focused instruction. This type of instruction, which has obtained positive results on students' writing performance (Graham & Perin, 2007; Graham et al., 2012), focuses on making students aware of their writing process by teaching them strategies for sub-processes such as idea planning, idea generation, goal-setting, or revising. The success of strategy-focused instruction paved the path for a process-oriented approach to writing instruction. A small but growing body of intervention studies offers feedback and instruction focusing on the writing process rather than on the writing product.

This process-oriented approach to writing instruction and feedback is in line with the recent conceptualization of feedback. Where feedback was traditionally seen as a product delivered to students (for example, comments provided by the teachers), the more recent feedback paradigm sees feedback rather as a process in which students actively engage with the feedback and make sense of it (Winston & Carless, 2020). This view entails that feedback should be designed in such a way that it provides students with opportunities to reflect on their learning and to monitor or self-regulate their learning (Nicol & MacFarlane-Dick, 2006).

The writing process can be defined as a functional dynamic system (Rijlaarsdam & Van den Bergh, 2006; van den Bergh et al., 2016) that consists of categories of cognitive and cognitive-linguistic activities, such as generating and organizing content, from memory and external sources, rhetorical goal-setting, formulation, rereading already written texts, evaluating and revising process progress and text-written-so-far. These basic activities can form all possible strings of activities in which an activity serves another activity in terms of goal-mean relations. The functional perspective refers to the relations between the activities that form strings of means-end relations. The dynamic perspective denotes that the functional relations between these strings may vary during the process. This writing process model shows the complexity of orchestrating writing process behaviors.

Hayes and Flower (1980) observed that "a great part of skill in writing is the ability to monitor and direct one's own composing processes" (p.39). Self-regulation plays an important role in this. Graham and Harris (2000) argue that writing competence depends on high levels of self-regulation. Students need support to learn to monitor, direct and regulate their actions (Hattie & Timperley, 2007). In other words, they must regulate their writing process. If the aim is to support students in developing their writing skills, they need guidance in developing effective writing process organizations (Vandermeulen et al., 2023). As several studies have confirmed the relationship between writing process activities (and

their orchestration) and text quality (Breetvelt et al., 1994; Vandermeulen et al., 2020c), also their texts would benefit/this would also enable them to improve their writing/one can also expect that their writing will improve as a result.

Keystroke logging tools have made it possible to gain insight into behavioral aspects of the writing process as by registering keystrokes, mouse movements and window switches, and time-stamp them (Leijten & Van Waes, 2013). Several studies have proposed ways of using keystroke logging data as a base to provide feedback and instruction on writing processes (Vandermeulen et al., 2020b).

In the present study, we tested the effect of two process feedback conditions on students' writing performance and writing behavior. For this study, we extended a previously carried out study (Vandermeulen et al., 2023), in which we tested the effect of data-based process feedback when compared to national baseline performances, by adding the observation of a writing process model. The previous study (Vandermeulen et al., 2023), on which our current study builds, provided students with a keystroke logging report (Vandermeulen et al., 2020b) with information on several writing process behaviors (such as source use, fluency, revision) that were found relevant in a national baseline study (Vandermeulen et al., 2020a). Subsequently, students compared their process data to the process data of higher-scoring students. This feedback intervention proved successful as participants displayed a more goal-oriented reading-writing strategy during their writing process, and they produced higher quality texts. The effect of the relatively short but intensive feedback intervention was comparable to the effect of one year of regular schooling. In the current study, we added an extra instructional component to the intervention, namely, the observation of a video model. We explored the added effect of a brief extension in which learners had the opportunity to observe an enacted model of their choice, showing one of three (empirically validated) orchestrations of the initial stage of writing a synthesis task. Participants in this study were randomly assigned to a process feedback condition with or without added models.

Literature review

We provide a brief overview of previous studies that provide feedback and instruction based on keystroke logging data (Sect. "Writing process-oriented interventions") and zoom in on two key elements of such interventions, namely, reflection and self-regulation (Sect. "Prompting reflection to stimulate self-regulation: three design principles"). We discuss the use of exemplars (Sect. "Compare and contrast: actual performance with exemplars") and observational learning with video models (Sect. "Compare and contrast: actual performance with models") as two methods to enhance comparing and contrasting activities that prompt reflection and self-regulation.

Writing process-oriented interventions

Previous studies offering process feedback or process instruction during interventions recurred to various keyloggers such as JEdit (Lindgren, 2004; Lindgren & Sullivan, 2003; Lindgren et al., 2009; Sullivan & Lindgren, 2002), Cywrite (Dux Speltz & Chukharev-Hudilainen, 2021; Ranalli et al., 2018), Inputlog (Vandermeulen et al., 2023; Bowen et al., 2022), and ProWrite (using concurrent keystroke logging and eye-tracking, Dux Speltz et al., 2022).

Information obtained from the keystroke logging programs was used in different ways to serve as input for the interventions. Some studies used a replay function (Lindgren, 2004; Lindgren & Sullivan, 2003; Lindgren et al., 2009; Sullivan & Lindgren, 2002), others worked with visualizations (Dux Speltz & Chukharev-Hudilainen, 2021), a report with numerical data and visualizations (Vandermeulen et al., 2023; Bowen et al., 2022), or screen replays followed by a remediation plan and real-time writing scaffolds (Dux Speltz et al., 2022).

A recurring key element in these interventions was the prompting of process reflection. The prompts for self-reflection varied from peer-discussions (Lindgren, 2004; Lindgren & Sullivan, 2003; Lindgren et al., 2009; Sullivan & Lindgren, 2002), or student–teacher discussions (Dux Speltz et al., 2022; Ranalli et al., 2018), to a set of guided questions and tasks (Bowen et al., 2022), or prompts to compare and contrast their process with example processes, varying from experienced writer examples (Ranalli et al., 2018), higher-scoring peers (Bowen et al., 2022), and higher- and equally-scoring peers (Vandermeulen et al., 2023).

In general, previous intervention studies reported positive effects on students' awareness of their own writing process, their writing motivation, self-regulation, and text quality.

Prompting reflection to stimulate self-regulation: three design principles

Reflection is a key component to self-regulated learning (Masui & De Corte, 2005) as it entails critical thinking about one's own task process and performance, leading to self-assessment (Quinton & Smallbone, 2010). Results from previous studies on effects of reflection on writing are not always consistent. Several studies on instructional programs aiming at performance improvement via self-regulated strategy instruction provide indirect evidence that reflective activities have positive effects since such programs improved writing performance (Graham et al., 2005; MacArthur et al., 2015) and writing self-efficacy (Zimmerman & Kitsantas, 1999). Interventions prompting writing process reflection based on keystroke logging data (as discussed in Sect. "Writing process-oriented interventions") reported a raise in students' consciousness of their own writing processes (Lindgren, 2004; Lindgren & Sullivan, 2003; Ranalli et al., 2018), higher writing motivation (Lindgren et al., 2009), more higher-order, textual revisions (Lindgren, 2004; Lindgren & Sullivan, 2003), changes in writing approach (Vandermeulen et al., 2023; Bowen et al., 2022; Dux Speltz et al., 2022), and better writing performance (Vandermeulen et al., 2023;

Bowen et al., 2022). However, Dux Speltz et al. (2022) did not find a significant effect of the intervention on students' text quality.

An instructional design to change writing processes via self-regulated learning is built on two pillars: the learning activity 'compare and contrast' and well-chosen objects to compare. One of these objects is one's own process. This leads to a first design principle: if learners have to reflect on their task processes, they should have optimal access to that process. The other object is a target object, namely a process or performance of another quality (i.e., a process or performance to acquire). Therefore, the second design principle is: if learners have to set goals, they should have access to process or performance models representing target behaviors. The third design principle is about prompting learners to relate the two objects (own process and a target process), which requires a specific form of analogic reasoning (Renkl, 2014). The instruction must stimulate students to find commonalities and differences between the two objects on a deeper level than the surface level. The third design principle is formulated as follows: if learners must learn from target models, they must compare and contrast the known and the target model to prepare goal-setting. Instruction based on target objects is known as example-based instruction (Van Gog & Rummel, 2010).

Renkl (2014) and Van Gog and Rummel (2010) distinguish two strands of research in example-based instruction. Instruction with worked examples is situated in cognitive theory and mostly studied in the area of problem-solving. The aim of learning by worked examples is constructing a strategy to solve certain problems, moving from the concrete (worked) examples to a more abstract schema via analogical reasoning (Renkl, 2014; Van Gog & Rummel, 2010). The second strand is grounded in Bandura's social-cognitive learning theory (Bandura, 1986). It refers to a natural human tendency to choose and observe model behavior to integrate in one's own processes and performances.

Both forms of example-based learning are especially effective in learning new skills. They are more effective than practicing these skills due to the lower cognitive load when reading worked-out examples or observing models (Van Gog & Rummel, 2010). Both forms can be included in instructional design prior to practice as well as following practice, as forms of feedback. In this paper we focus on their use in the feedback phase, to trigger reflection for goal-setting.

Compare and contrast: actual performance with exemplars

Exemplars can be defined as key examples of authentic student work that are typical of certain levels of quality or competence (Sadler, 1987). Stimulating students to learn from examples triggers reflection and stimulates self-regulation in student learning (Nicol, 2021; Scheiter, 2020). Research has shown that students benefit when they are encouraged to use exemplars as a point of reference when reflecting on their own work (Dixon et al., 2020). Several positive effects are attributed to exemplars. They can represent concrete task standards (Hendry et al., 2011), or show different strategies to complete tasks (Orsmond et al., 2002). Moreover, exemplars are considered encouraging by students as they are situated in the zone of proximal development and thus show a goal within reach (Hendry et al., 2011). The studies by

Bowen et al. (2022) and Vandermeulen et al. (2023) provided students with writing process data gathered with keystroke logging and visualizations of exemplar writing processes. The exemplars were labeled *feed-forward* as they consisted of exemplars of higher-scoring peers (known as the *feed-forward* notion- addressing the "Where to next?" question—from Hattie and Timperley's (2007) model). Both studies reported positive effects of the intervention on the participants' writing process approach and text quality.

Compare and contrast: actual performance with models

Observational learning involves the observation of (video) models in which a learner models a certain behavior. Several studies demonstrated the effectiveness of observational learning in the domain of writing (Couzijn, 1999; Groenendijk et al., 2013; Zimmerman & Kitsantas, 2002). For example, Braaksma et al. (2004) found that observational learning resulted in a higher amount of higher-level processes such as planning, with a more dynamic distribution of such processes across the writing process, and higher quality texts. Raedts et al. (2007) showed that students who learned by observing pairs of video-based peer models reported more extensive knowledge of effective strategies such as information gathering and planning. These students also wrote better texts compared to the control condition, and reported higher levels of self-efficacy.

Effects of observational learning are related to different factors. First, by observing writing behavior rather than performing it itself, there is a reduction in cognitive load (Braaksma et al., 2004). Students can focus on observing and reflecting instead of on executing a writing task. Secondly, observational learning provides students with insights into the variety of writing processes and into self-regulation during processes (Braaksma et al., 2004) as the model displays goal-setting via self-instructions, and via monitoring evaluation processes. Thirdly, when observing relatable peers, observational learning can also motivate learners, as it shows that the goals are within reach (Raedts et al., 2007).

Present study

The present study was guided by the two research questions:

1. Do the two interventions (feed-forward process feedback and feed-forward process feedback followed by observation of a video model) have an impact on writing performance and writing process behaviors?
2. Does this impact differ between the two intervention conditions?

We expect to replicate the effect of process feedback as demonstrated in Vandermeulen et al. (2023), with the process feedback based on comparison with exemplars positively affecting text quality and students' task process. Second, we expect that the addition of the opportunity to observe one of the three empirically validated

orchestrations in the form of an actual model-in-action will have an effect on the top of the feedback effect.

The present study builds on previous studies (addressed in Sect. "Literature review") and aims to fill some of the gaps. Studies on the effect of process feedback do not fully report how students are guided from data to goal-setting. Interpretation of the variability of effects is then difficult. Elements that are missing are the definition of a target process, and decision-making on which target process is most suitable for the participant. Therefore, in the current study we aim to provide students with target 'profiles' of how a synthesis task can be tackled by relatively proficient peers. Participants are encouraged to compare themselves to the target profiles. The focus of these comparisons is on the orchestration of the activities. Another element that is missing in previous studies is the connection between a numerical (i.e., as in the process report and the quiz) and verbal description of a profile on the one hand), and a profile model-in-action on the other. Therefore, we will study the effect of the addition of an observation task that provides students with an example-in-action.

Methodology

Participants

Participants were 185 Dutch 10-grade students from seven classes in three secondary schools. They were all enrolled in pre-university education (Dutch *VWO* stream). Successful completion of this program allows the candidates admission to university. Participants were between 15 and 18 years old (average age of 15.61) ($N_{female} = 107$, $N_{male} = 78$). Written consent was obtained from all students.

Participants within classes were randomly assigned to a condition in which participants received feed-forward process feedback ($N_{students}$ F = 87) and a condition in which participants received feed-forward process feedback followed by the observation of a video model ($N_{students}$ FO = 98). The imbalance in participants between the two conditions is due to random attrition (related to students dropping out because of illness, and technical difficulties with the keystroke logging tool). The distribution of male and female students across conditions, based on the participants that revealed their gender, did not differ to a statistically significant degree ($\chi^2(1, 171) = 0.718$, $p = 0.760$). Students received film tickets for their participation.

Research design

We implemented a repeated intervention design (Table 1) with three sessions (M1, M2, M3). Each session consisted of a before-writing phase, a writing phase, and an after-writing phase. The measurement design in Table 1 shows the activities carried out in each of the three sessions. Participants wrote an informative synthesis text during each session. Processes were logged with keystroke logging software Input-log. In sessions M2 and M3 students received feedback on their writing process of the texts written in the earlier session. Two feedback conditions were implemented

Table 1 Research design

| | Measurement Design | | | | | | | | |
| | Session M1: Text 1 | | | Session M2: Text 2 | | | Session M3: Text 3 | | |
	Pretest	Writing	After	Before	Writing	After	Before	Writing	After
Self-efficacy: long form*	×								×
Writing Style Inventory*	×								
Text Quality		×			×			×	
Synthesis Process Behaviors		×			×			×	
Perception Writing Task*			×			×			×
	Intervention Design								
	Session M1: Text 1			Session M2: Text 2			Session M3: Text 3		
	Pretest	Writing	After	Before	Writing	After	Before	Writing	After
Self-efficacy: short form				F	FO			F	FO
Feedback Evaluation: short				F	FO			F	FO
Exemplars				F	FO			F	FO
Observing Models					FO				FO
Feedback Evaluation: long*				F	FO			F	FO
Evaluating Observation*					FO				FO

The questionnaires indicated with * are not included in the present study. A description of all instruments and the respective questions of the questionnaires can be found in Vandermeulen et al. (2020a, 2023)

F: Feed-forward condition; FO: Feed-forward with Observing models condition

(Table 1, Intervention Design): the Feed-forward Feedback condition (F), in which the participants received feed-forward exemplars, and the Feed-forward Feedback condition plus Observation (FO), in which the participants additionally observed a model in action, on video. In the present paper, we focus on two types of variables: synthesis process variables and text quality.

Instructional design

Feed-forward process feedback

We designed the procedure according to three design principles, distilled from literature. Feedback was designed (1) to help students bridge the gap between their current and targeted performance, (2) to incite self-regulated learning in which understanding, reflection, evaluative judgment and monitoring were key (Panadero et al., 2019) and (3) to activate them (Nicol & MacFarlane-Dick, 2006). Feedback was based on the comparison of actual process data (Design principles 1 and 2). To support students to get an insight into their actual performance and process, students received: (1) feedback on their position on a text quality scale, and (2) quantitative

data about their writing process as outlined in Vandermeulen et al. (2020b) . To activate students to process the quantitative writing process data, they had to go through a series of learning activities, that is, compare and contrast and set themselves goals (Design Principle 3). To stimulate goal-setting, holistic exemplars of task processes were provided, based on an analysis of processes in a national representative sample. The exemplars consisted of exemplary writing processes of high-scoring writers, engaging the participants in the feed-forward aspect of feedback (i.e., Hattie and Timperley's (2007) "Where to next?" question).

In the following sections, we describe the guiding steps that were presented to participants to get insight into their process feedback data. In Sect. "Data sources", we provide details of the data that students had available. Sect. "Comparing, contrasting and goal-setting" presents all the learning activities students carried out to get grip on the data and set goals.

Data sources

Text quality Texts were rated by two independent raters. Scores were reported next to a text quality scale based on a national representative sample (Vandermeulen et al., 2020a) (Fig. 1).

Task process information Based on keystroke logging data, we generated individual reports on several writing behaviors, most of them in equal intervals of the process. This report followed the principles of Vandermeulen et al. (2020b).

Target task information To provide target processes, we constructed empirically validated exemplars and models of the initial stage of the writing process. We analyzed the initial stage (first interval) of the writing processes of the 50% best performances of upper-secondary students in a national assessment on (informative) synthesis writing (Vandermeulen et al., 2020a). A Latent Class analysis on these processes
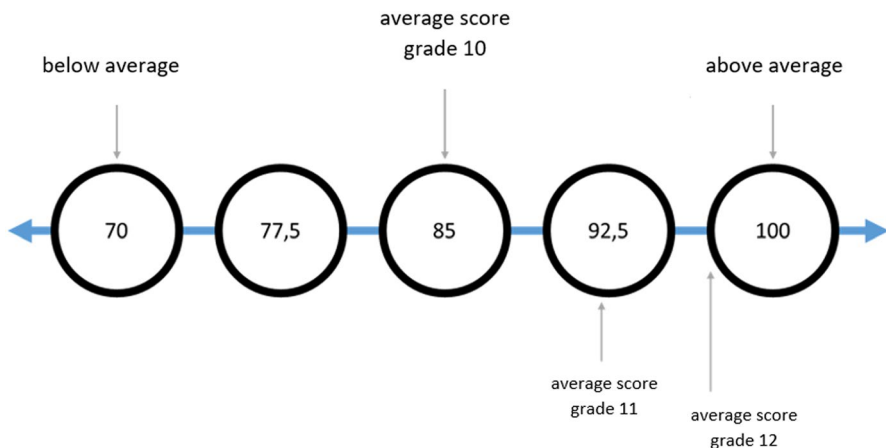

Fig. 1 Scale with average scores

resulted in three distinct task profiles, labeled as: the notetaker, the source reader and the thinker/focused reader. For the verbatim description of the profiles provided to students see Online Supplementary Materials A. The notetaker is characterized by heavy text production and considerable production speed at the same time as a more than average switching between the sources and one's own text process-initially, that is, compared to the other two profiles. The source reader displays a prominent focus on source use at the beginning of the writing process reflected in a more than average time in the sources and switches between these sources compared to the other two profiles. Finally, compared to the notetaker and the source reader, the thinker/focused reader spends a lot of time pausing at the beginning of the writing process which reflects reading and thinking. Sect. "Comparing, contrasting and goal-setting" describes how students in the different conditions were meant to process and make use of the components of the feedback, more specifically, of the task process information and the (descriptive) target task information.

Comparing, contrasting and goal-setting

We designed a series of learning activities to move from data to goal-setting (Table 2). In Sessions 2 and 3 students received and processed feedback on the former synthesis process and set goals by working through the learning activities in an Interpretation Guide. Students in the Observation condition additionally chose and observed a video model.

To guide students in the processes of comparing, contrasting and goal-setting, three steps were designed: getting a grip on the data provided, comparing their process to exemplars and setting process goals. The interpretation guide led students through the different steps.

Getting a grip The process started with self-assessment (self-efficacy), insight in the position on the text quality scale (Fig. 1), and reading the report with writing process data (Fig. 2). Students first filled in the short self-efficacy form with six questions probing for their self-efficacy beliefs in their selection and integration of source information, the ease with which they believed to be able to paraphrase source information, text structure, language use and conciseness. The short self-efficacy ques-

Table 2 Learning activities in conditions in both feedback sessions, before writing a new synthesis text

Activity	Conditions	
	Feed-Forward	Observation
Feedback Processing towards Exemplars	x	x
Choosing and Observing a Model		x
Goal-setting	x	x
Evaluating Feedback	x	x
Evaluating Video		x

Time use. When you write a text, you have to divide your time between several subtasks. A part of the time you spend writing (we call this: active writing time). The other part of the time you don't write. You read the sources, you reread your own text, or you are thinking.

How did you divide your time?

- Your total active writing time: 39.56 minutes

- Your total pausing time: 4.26 minutes

- Your total source time: 13.73 minutes

We have divided your process into three equal phases. In your case, each phase took 14.6 minutes. In the beginning, middle and end of your writing process, you use your time differently. In some phases, you might mainly engage in active writing, while in other phases you are more likely to read sources or to think. How did you divide your time between the different activities in the different phases?

	Phase in Your Writing Process		
Activity	Initial	Middle	Final
Active writing time	33,99	65,53	77,12
Pause-thinking time	2,81	9,87	16,48
Source time	63,20	24,60	6,40

Fig. 2 Example of Time Use in the Feedback Report with writing process data

tionnaire, based on a validated, more extended questionnaire (Vandermeulen et al., 2020a; six scales with reliabilities (α) ranging between 0.70 and 0.87), can be found in Table B1 in Online Supplementary Materials B. Then students were provided with information on their position in a text quality scale after which they were presented with a writing process data report.

Compare current process with exemplars Students completed a quiz that prompted them to go back to their data in the report, and which resulted in a position on or between the three exemplars (Fig. 3).

Goal-setting for the next synthesis task Two open questions were asked, starting with what students would like to keep, and then what they wanted to do differently (Fig. 4).

The design of the observational learning component

Students assigned to the Observation condition continued the learning phase with observational learning. They had a choice out of three models: the notetaker, the source reader, or the focussed reader (see Sect. "Data sources"). We based the videoclips on scripts with actions that reflected empirically grounded profiles. Following

How much time did you spend in the sources in the first stage of the writing process?
A   40% or less
B   more than 55%
O   between 40 and 55 %

After the nine questions, students had to count their letters.

Identify your type. How often have you circled the following letters?
A: .... B: .... C: .... O: ....

Then the key followed, with three descriptions of profiles. The text described the profile and asked questions to stimulate reflection, for instance:

"Most B's and O's: Your approach is probably most similar to that of the source reader. For instance, you spend much of your time in the early stages reading sources and switch between sources more often than an average student. This is a good approach, but also be aware of the pitfalls. For instance, have you succeeded in using all relevant information from the sources in your text? And do you not have to keep going back to the sources while writing?"

Fig. 3 Example of the quiz—comparing personal process data with exemplars

How will you write your new text?

You have now been introduced to different approaches. You also know what your process looks like. Now that you know this:

When writing your new text, what will you do the same? Why?

....

When writing your new text, what will you do differently? Why?

....

Fig. 4 Goal-setting questions

the recommendations by van Ockenburg et al. (2019) to provide students with freedom of choice to adapt strategies to their own preferences, participants were free to choose which model they observed. Videoclips are roughly 7 min in length (see website www.liftwritingresearch.wordpress.com). The empirically grounded construction and operationalization of the models can be found in Online Supplementary Materials A.

Procedure

The intervention took place in students' schools at three different measurement occasions in the course of seven days. Each session was supervised by a group of researchers and assistants acting as facilitators and making sure the protocol was

adhered to. Table 1 lists the different steps and measures at each measurement occasion.

At each session students had 50 min time to write an informative synthesis task. Students of both conditions were seated in separate classrooms. Processes were logged with keystroke logging software Inputlog (version 7.1.0.53). Students in the FO-condition received headphones to watch and listen to the video models.

At M2 and M3 prior to writing, students received feedback on their writing process. They had 25 min to process the feedback individually (see Sect. "Research design" for feedback). After having processed the feedback and prior to writing at M2 and M3, students also filled out a short feedback evaluation questionnaire. Students had to indicate the degree to which they agreed with the following statements: (1) I considered the information about my own writing process useful, (2) I considered the questionnaire about my own writing process useful, (3) In my opinion the description of the different approaches to writing (Note-taker, Source Reader, Focused Reader) is useful, (4) The videoclip is useful.

Measures

Questionnaires

Table 1 illustrates that students in both conditions filled in a short self-efficacy questionnaire prior to receiving the feedback at M2 and M3 to trigger awareness and encourage reflection. At the end of Sessions 2 and 3 after having processed the feedback and prior to writing a new text, a short feedback evaluation questionnaire was administered.

Self-efficacy short form The questionnaire is a shortened version of a previously validated longer self-efficacy questionnaire containing six questions (cf. Table B1 in Online Supplementary Materials B) corresponding to the selection and integration of source information (Dealing with sources and Integration of the sources), the ease with which students believed to be able to paraphrase source information (Elaboration of the sources), Text structure, Language use and Conciseness.

Feedback evaluation questionnaire The final step in the feedback flow prior to writing consisted of a short questionnaire to evaluate the feedback. Students were asked to indicate on a five-point-Likert scale the extent to which they agreed with five statements about the usefulness of the writing process information and questions (Interpretation Guide), the exemplar processes as described in the different profiles in the Interpretation Guide and in the FO-condition about the videoclips watched (cf. Table B2 in Online Supplementary Materials B).

Writing product

Tasks Writing assignments were selected from a national baseline study in the Netherlands with 658 students in upper-secondary education (grades 10–12). Task con-

struction is described in detail in Vandermeulen et al. (2020a). Sources were short informative texts. Students received additional explanations on the characteristics of an informative synthesis text, target-audience, text length and on using the sources. Additionally, they were briefly instructed on how to use Inputlog.

Rating Text quality was rated with Comproved, software based on the pairwise comparison method (Lesterhuis et al., 2016). Within pairwise comparison, raters are presented each time with two synthesis texts from which they have to select the better one based on a holistic evaluation of the text quality. Each text is compared multiple times. Based on the comparative judgements of all raters, using the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959), a rank order is calculated from worst to best text. For each text a logit score is calculated. Next, a reliability for the whole rank order is calculated, the Scale Separation Reliability, ranging from 0 to 1, with 1 representing a perfect consensus and full security on the position of the text on the rank order. In addition, the method of comparative judgements enables one to rescale the logit scores to, in this case, the baseline study (Vandermeulen et al., 2020a), by adding a text with a low and a text with a high-quality score to the comparisons.

The 23 raters all made 100 comparisons, so each text was compared about 10 times with another text. The raters were PhD students in an educational or language related program, and master students in an educational program, and were familiar with rating text quality and the Comproved software. They were explicitly asked to evaluate the texts holistically taking into account all of the following four criteria (Vandermeulen et al., 2020a): (1) relevance and correctness of the information, (2) integration of the sources into a new text with its own structure and overarching theme, (3) coherence and cohesion, and (4) language use. Raters received a financial compensation. The SSR index was 0.62, acceptable considering that we worked with untrained raters, which adds to the study's validity as they all brought their own, valid, perspective to text quality to the assessment.

Writing process/behavior

Filtering the data Prior to analyzing the keystroke logging data, we filtered them by using both the time filter and the source recoding functions in Inputlog. We used the last key students typed as an indication for the end of the writing process. This means that we removed the time from typing the last key to searching for the Inputlog stop button at the bottom of the screen. In line with Vandermeulen et al. (2020c), we used the source recoding function to group the sources identified by Inputlog into one of three source categories: given source texts, the synthesis text, and off-task sources.

Writing process/behavioral measures Writing processes were analyzed using Inputlog. In line with Vandermeulen et al. (2020b, 2023), and Van Steendam et al. (2022), we distinguish easily interpretable and meaningful relative process measures for two writing behavioral aspects: source use and production (Table 3). The studies illustrated that the variables were amenable to change. The variables selected also cor-

Table 3 Behavioral measures used for the analyses

Aspect	Behavioral measure interval 1	Short variable name
Source use	Proportion of time in sources	Source Time
	Number of switches per minute between sources	Switches Sources
	Number of switches per minute between synthesis text and sources	Switches Synthesis—Sources
Production and fluency	Proportion of active writing time	Active Writing Time
	Number of keystrokes per minute	Speed

responded to the variables selected for the three writing process profiles. Source use is represented by three variables: (1) (relative) amount of time spent in the sources, (2) number of switches between the sources (per minute), (3) number of switches between the synthesis text and the sources (per minute). Finally, the production measures were (1) active writing time, and (2) speed operationalized by the number of keystrokes per minute. The process measures were selected for the first (of three) intervals of the process. Writing processes were divided into three equal intervals, and we selected the first interval representing the beginning of the writing process where we expected source use activities to be most prominent.

Treatment fidelity

Research assistants supervising the data-collection closely adhered to a script with the various steps in the procedure and were also instructed prior to the data-collection sessions.

To check the random assignment of the students to both conditions a multivariate model was fitted to estimate the differences between the FO-condition and the F-condition for all six dependent variables (text quality and the 5 behavioral process measures) at Session 1 (see Appendix A). The resulting highest density intervals are for all these variables situated around the value zero, indicating that we have no evidence that there were significant differences between both conditions at Session 1 for any of the dependent variables.

A check of the questions of the feedback evaluation questionnaire (prior to writing at Session 2) showed that these had been filled out by all students in both conditions. The feedback was positively evaluated in both conditions (means for the three feedback elements ranging from 3.79 to 3.93 out of 5). In the FO-condition, the model-videos were positively evaluated (3.5 on a scale from 1 to 5). No statistically significant differences between conditions were found for evaluation of feedback quality, for none of the feedback elements: not for the feedback report, nor for the quiz or the profile description: $F (3,145) = 0.228$, $p = 0.877$; Wilk's $\Lambda = 0.995$, partial $\eta^2 = 0.005$).

Analyses

Correlational analyses

We estimated the correlations between text quality and the behavioral measures at each of the three sessions. We used a Bayesian framework to estimate the posterior probability distribution for each of the correlation coefficients so that we can make use of its property that all inferences are probabilistic, giving us a framework to express inferential and predictive uncertainty (Gelman et al., 2021). Analyses are implemented in the probabilistic programming language *Stan* (Carpenter et al., 2017) making use of *R* (R Core Team, 2020) and the package *brms* (Bürkner, 2017). To estimate the models we made use of the default weakly informative priors defined in brms. Given that we have a sufficient amount of data, weakly informative priors are capable to keep the inferences in a reasonable range without making the inferences sensible to the choice of priors (Gelman et al., 2021). Estimation of the models was based on a Markov Chain Monte Carlo (MCMC) sampling algorithm, making use of 6 chains of 6,000 iterations each, with 1,500 burn-in iterations to guarantee convergence (see Lambert, 2018 for an in-depth explanation on the MCMC sampling algorithm).

Estimating the differences between both conditions

To analyze the potential different impact of both conditions (F vs. FO), we rely on three alternative models, that are an implementation of a multivariate model, assuming that the scores for each session are three separate dependent variables that will be modeled simultaneously. This approach allows us to model the variances and covariances between students for each session, not assuming a similar variance between students at each session. The models were run for all six dependent variables: text quality and the five behavioral variables measuring the writing process.

A first model (Model 1) is an unconditional model not taking into account the possible effects of the intervention. This model only assumes three average scores (one for each session), three variances between students (one for each session) and covariances between the scores at each session for students. The following online supplementary material presents an elaborate formal definition of the model (https://majestic-pixie-28398c.netlify.app/).

To estimate the difference between both conditions, we introduce the effect of condition as a dummy variable (indicating whether a student was assigned to the F-condition (=0) or to the FO- condition (=1)) in a Model 2. The dummy variable is added to the model as a predictor for scores on Sessions 2 and 3. Specifically, in Model 2 we assume that the effect of the condition is identical at both sessions. In other words, there is a single parameter expressing the effect of condition, which is a parsimonious model.

In Model 3 we loosen this assumption of an equal difference between both conditions at both Sessions 2 and 3, allowing for a session-specific effect of condition. If Model 3 fits the data better than Model 2, the condition effect varies at sessions.

Bayesian models are used as they are ideal for reliable parameter estimation given that we can derive a posterior probability distribution of the parameter values of interest (McElreath, 2020). The posterior probability distribution of these parameters can be used to propagate the uncertainties around the parameter estimates when calculating contrasts. Moreover, as our design lacks a control group, we will follow van den Bergh et al.'s (2023) suggestion and use a national baseline study as a point of reference to evaluate the effect of the intervention which necessitates a Bayesian approach (see Sect. "Comparison to a national baseline").

The models are estimated for each of the dependent variables separately. By comparing the three models on their model fit, making use of *leave-one-out cross validation* (Gelman et al., 2021), we select the best fitting model for each of the dependent variables. Posterior probability distributions of the parameter estimates based on these best fitting models are explored by describing 89% credible intervals (McElreath, 2020) and the probability of direction (Makowski, et al., 2019).

We implemented the model estimation in the probabilistic programming language *Stan* (Carpenter et al., 2017) making use of *R* (R Core Team, 2020) and the package *brms* (Bürkner, 2017), making use of the default weakly informative priors defined in *brms*, using 6 chains of 6,000 iterations each, with 1,500 burn-in iterations.

Comparison to a national baseline

Besides estimating the differences between the two conditions, we also want to learn more about the impact of the intervention on text quality and the five behavioral measures. Therefore, we compare the change in scores from Session 1 to Session 2 and Session 3. This comparison is hindered by task effects: at each session, another writing task was implemented (Vandermeulen et al., 2020a). Following the suggestion by van den Bergh et al. (2023), we made use of the data available from a broader national baseline study (Vandermeulen et al., 2020a) with the same tasks, measurements of text quality, and behavioral measures as used in the current intervention study. The baseline consists of a nationally representative sample of students ($N = 658$) from three grades (10–12) who each wrote multiple synthesis texts ($N = 2310$) without receiving feedback (Vandermeulen et al., 2020a). Comparison with the baseline allowed us to compare the feedback effect with the grade effect, namely, the effect of the progress students make over time or over grades. Since national baseline studies offer a large number of observations on a large variety of tasks, they provide a much richer account of the general level of achievement than a traditional control group. For a more detailed description of the baseline study and its methodological considerations, we refer to Vandermeulen et al. (2020a).

Using data from 10th graders in that baseline study ($N_{students} = 242$, $N_{texts,\ processes} = 435$), we estimated a mixed effects model for each of the dependent variables containing random effects for schools and tasks using a Bayesian modeling

approach. Based on these models, we extracted the average score of 10th graders for the specific tasks used in our intervention study and the uncertainties around those average scores based on the posterior probability distribution. These posterior distributions coming from the national baseline data formed a point of reference to evaluate the posterior distributions for the average scores derived from the models applied on the intervention data.

Open data and code statement

All the data for the Bayesian analyses, Stan code, R scripts and output are available on the OSF (https://osf.io/3sb7m/?view_only=b89adb32abaa46c59676cc7c6a47a62f). In an online walkthrough we further documented the analyses and code. This online document is accessible via: https://majestic-pixie-28398c.netlify.app/.
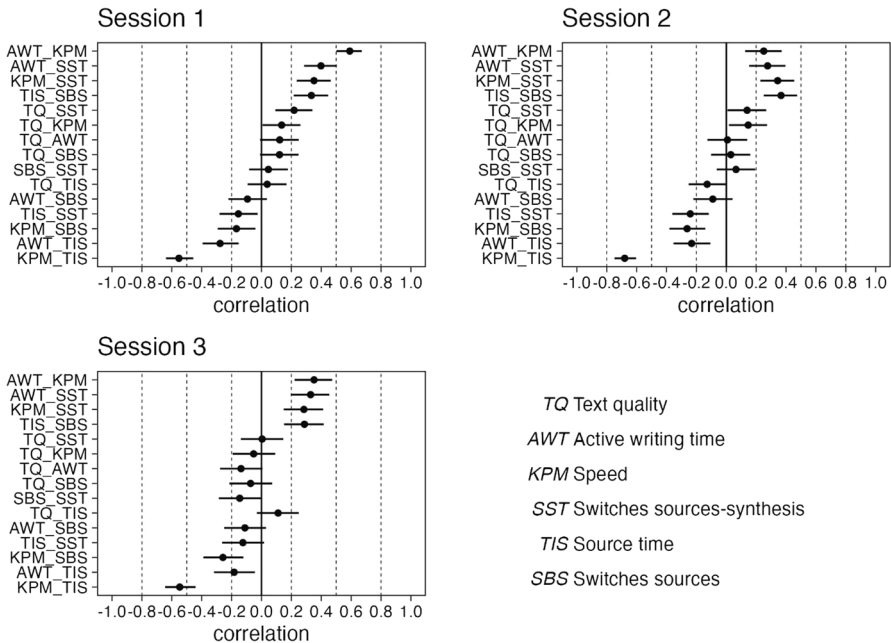
Fig. 5 Posterior probability distributions for the correlations between text quality and the process variables at each session and the accompanying 89% credible intervals based on Bayesian estimation. Correlations are arranged from largest positive to largest negative in session 1

Results

Correlations

A first step in the analyses is the estimation of the correlations between the dependent variables (text quality and the five behavioral process measures). Figure 5 visualizes these correlations for each of the three sessions.

The correlations between text quality and the different process measures are all small or negligible, at each session.

The only constant moderate correlation in each session is the negative correlation between keystrokes per minute and time in sources. At Session 1 there is also a moderate positive correlation between active writing time and keystrokes per minute. This correlation is lower in Sessions 2 and 3 as all correlations become weaker.

Model comparisons

After all three models were estimated for each of the dependent variables, we compared the model fit of these models, based on a leave-one-out cross-validation (Online Supplementary Materials C). We rely on the *looic* to select the best fitting models.

From these model comparisons we learn that only for text quality Model 1 outperforms the other models (lowest looic), meaning that for text quality we have no convincing evidence that students in the F-condition wrote better or worse texts than students in the FO-condition (RQ2).

For the process variables we see that Model 2 shows the best fit for three behaviors: active writing time, time in sources and switches between sources and text. This indicates that students in the two conditions score differently for these variables both at Sessions 2 and 3, with this difference being equal for both sessions. Finally, for keystrokes per minute and switches between sources Model 3 shows the best model fit. Students in both conditions score differently for these variables at Sessions 2 and 3 and the difference is not equally large for Session 2 as it is for Session 3 (RQ2).

In what follows we describe these best fitting models for each of the dependent variables.

Text quality

The best fitting model for Text Quality was the unconditional model (Online Supplementary Materials C, Table C1). From the posterior probability distributions for the contrasts we learn that there is an increase in text quality at each session (RQ1).

Next, we want to test whether the effects of the interventions (as shown by an increase in text quality at Sessions 2 and 3) can be explained by the fact that different tasks are used at each session. Therefore, we make a comparison with information from the baseline study (RQ1). In Fig. 6, we plot the posterior probability
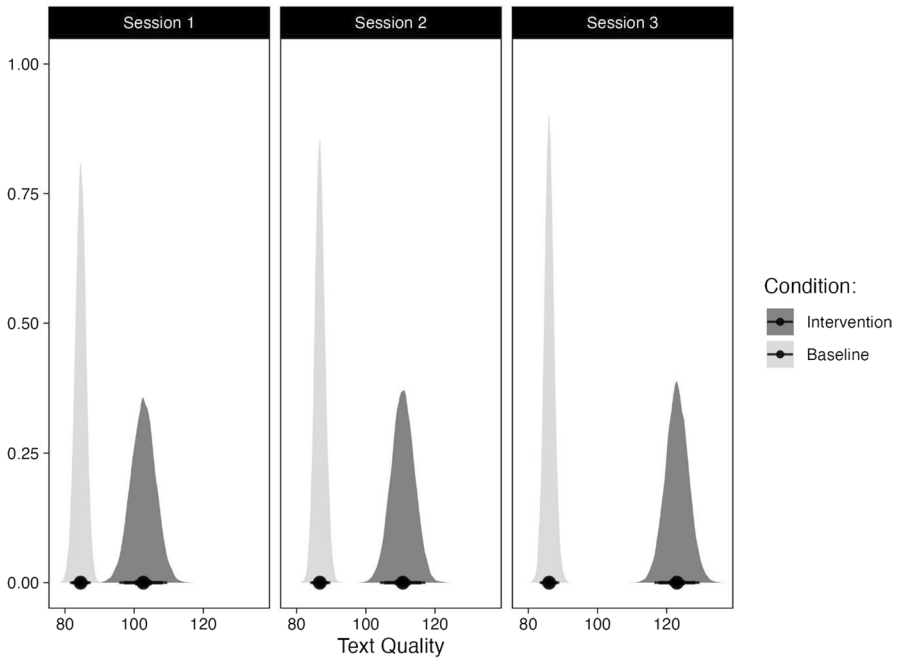
Fig. 6 Posterior probability distribution of the average text quality scores for both the national baseline and the students in the intervention study for the texts used in each of the three sessions

distribution of the intervention effect against the posterior probability distribution for the average text quality of similar tasks at the baseline study. The figure shows that already at Session 1 students in the intervention study outperformed the students in the national baseline. Nevertheless, we see that the difference between the students in the intervention and the national baseline increases in Sessions 2 and 3, compared to the difference in Session 1. Looking at the x-axis, we can observe that the gap between the distributions of the baseline scores and the distributions of the intervention scores is wider at Session 2 than at Session 1, and is wider at Session 3 than at Sessions 1 and 2. Based on the posterior probability distributions we can calculate the differences at each session and an 89% highest density interval. At the first session the 89% most probable values for the difference are situated between 12.1 and 24.5 points (with a median of 18.1 points). This difference is somewhat larger at Session 2 with the 89% most probable values for the difference being situated between 18.1 and 29.7 points (median of 23.9 points). As we can propagate the uncertainties, we can also calculate the strength and uncertainty around this increase in difference between baseline and intervention students between Session 1 and Session 2. The 89% most probable values for this increase in difference between Session 1 and Session 2 are situated between -1.6 and 13.24 (median of 5.8), meaning that we have no convincing evidence that there is an increase in difference between Session 1 and Session 2 (RQ1).

At Session 3 the difference is even larger, with the 89% most probable values situated between 31.1 and 42.7 points (median of 36.9 points). This difference between
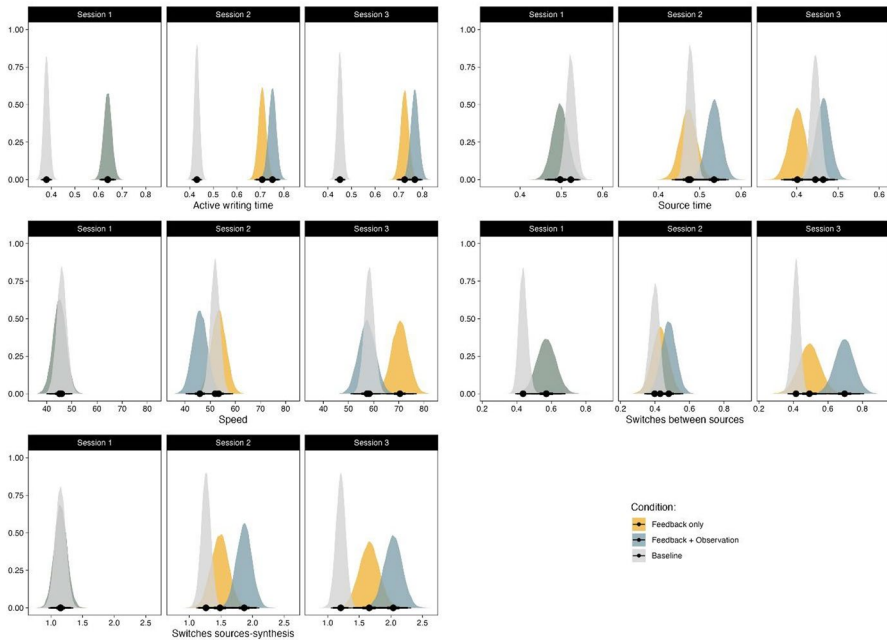
Fig. 7 Posterior probability distributions for the average scores on the behavioral measures for both the national baseline students and the students in the intervention study

national baseline students and intervention students is convincingly higher than in Session 1 and in Session 2: respectively the 89% most probable values for the increase in difference between Session 3 and Session 2 is situated between 5.6 and 20.1 points (median of 13.0) and the 89% most probable values for the increase in difference between Session 3 and Session 1 is situated between 11.2 and 26.7 points (median of 18.8 points). This brings us to the conclusion that students certainly improved the quality of their written texts in the intervention at Session 3 (RQ1).

Behavioral measures

The parameter estimates for the effects of the dummy variable condition show that F-condition together with FO-condition resulted in different scores for a set of behavioral measures (Online Supplementary Materials D, Table D2) (RQ2). Students in the FO-condition scored higher at Sessions 2 and 3 than students in the F-condition for the variables active writing time (CI: 0.01—0.08; Median: 0.04), source time (CI: 0.02—0.10; Median: 0.06) and number of switches between sources and synthesis (CI: 0.11—0.65; Median: 0.38). For the variable switches between sources, we also find that students in the FO-condition score higher than students in the F-condition, but only in Session 3 (CI: 0.07—0.34; Median: 0.20). Finally, we can also conclude that students in the FO-condition slowed down their writing speed compared to students in the F-condition, both at Session 2 (CI: −13.90—1.54;

Median: −7.47) and Session 3 (−20.25—5.77; Median: −13.11).When comparing the estimates from our intervention with estimates from the baseline study (Fig. 7) (RQ1), we can see that the intervention students clearly score higher on active writing time than the baseline students, at each of the three sessions. Concerning source time we see the following pattern: students in the FO-condition increased their time in the sources at Session 2 compared to the baseline students, while in Session 3 they score similar as the baseline students. But at the same time, students in the F-condition spend less time in the sources at Session 3 compared to the baseline. For the variable speed we see that only at Session 3 there are clear differences between students in the baseline study and the students in the F-condition, showing that the latter increased their speed compared to the former. Also for the variable switches between sources we can mainly see a difference at Session 3 between students in the FO-condition compared to the baseline student. Students in the FO-condition show more switches between the sources at Session 3. The same holds for the number of switches between sources and synthesis text, but also the students in the F-condition showed more switches between sources and synthesis than the baseline students at Session 3. At Session 2, students in the FO-condition started to show more switches between the sources and the synthesis than the students in the baseline study.

Discussion

Two questions were central to this study:

1. Do the two interventions (feed-forward process feedback and feed-forward process feedback followed by observation of a video model) have an impact on writing performance and writing process behaviors?
2. Does this impact differ between the two intervention conditions?

We investigated the first research question using a research design that partially replicated Vandermeulen et al.'s (2023) design, in which students were exposed to the process feedback intervention twice. In that intervention, students received feedback about their task behavior via a report with keystroke logging data, and compared their writing behavior to exemplar processes. As Vandermeulen et al.'s (2023) study showed a positive effect of the feedback on text quality when students compared their processes to exemplar processes of higher-scoring students (feed-forward), the current study also included comparisons with process behaviors related to higher-scoring texts. The difference between the previous study and the current study is in the design of the exemplars. While Vandermeulen et al.'s (2023) study used two annotated process graphs (visualizations of the complete writing process) relative to the participants' performance (exemplars of 1 SD and 2 SD higher), in the present study, insight via detailed automatic process feedback into students' processes is conducted via a holistic comparison with three empirically validated exemplars. These three exemplars show three different approaches for the first process phase, at the moment when task and source information must be comprehended and

the student starts to set up a synthesis text. The selection of these exemplars is based on an analysis of data available from a national assessment study (Vandermeulen et al., 2020a). Besides replicating the previous study, we also tested whether the subsequent observation of a video model added to the effect of feedback. Effects were determined for text quality and five writing process behavior measures in the first of three intervals of the writing process. We compared the effects of both conditions to a national baseline, which contains data from a nationally representative sample of students of the same age and educational level.

The results of this study into effects of process feedback and model observation on writing performance and process behaviors can be summarized in the following way.

Replication: impact of process feedback

The feed-forward process feedback had a positive effect on text quality at the end of the intervention program (i.e., after two feedback intervention instances), compared to the national baseline. Students in the experiment already wrote better on average from the beginning of the experiment than students in the national baseline, but after a second process feedback intervention, the gap widened significantly. Students in the experiment received no feedback on the quality of the text they wrote. They were only told how they scored on the national performance scale (Vandermeulen et al., 2020a). The score may in many cases have been a sign that there was still room for growth, but without indicating in which features of the texts that growth could be achieved. The increase in text quality could be due to practice: after all, students wrote three synthesis texts in a short period of time. In the national baseline study in which students each wrote four texts such an effect could, however, not be determined based on a secondary analysis (ES at most 0.14, $F(3,2197) = 2.314$, $p = 0.074$). Consequently, we could infer that the feed-forward process feedback, both with and without observation, had a significant effect on text quality compared to a national baseline.

On four of the five behavioral measures, the experimental conditions had an effect, compared to the baseline data. In the process of the first text, we saw no differences between the conditions and the national baseline except for Active Writing Time. Active Writing Time was higher in the experiment than in the baseline, at all measurement moments, and the gap did not change across the sessions. On the four remaining behavioral variables, that is, Keystrokes per Minute, Time in Sources, Switches between Sources and Switches between Sources and Synthesis, we did observe an effect, and it always occurred at the third session, after two rounds of process feedback. The feedback conditions showed higher writing speed, more switches between sources, more switches between sources and text-produced-so-far, and less time in sources. If students spend less time in sources while switching more between sources, this may indicate a more active search for information integration. Such an activity is then accompanied by a higher writing speed, which may indicate

that students are utilizing more information from their own memory in the text (Baaijen & Galbraith, 2018).

Extension: observation of the enactment of an exemplar video model

For text quality, the addition of an activity to observe a process model in action had no additional contribution to the impact of the feed-forward process feedback. We observed no differences between the two conditions; both had similar positive effects compared to the baseline. We did observe an additional effect on all five process variables, however, all of which were already observable after a single round of feedback. Compared to the F-condition, students spent more Active Writing Time, but also more Source Time, indicating that less time was left unused. Students in the FO-condition also switched more frequently between sources and between sources and the synthesis text. Writing Speed, on the other hand, was lower than in the F-condition. Two of these effects are amplified after the second round of feedback: even less rapid text production and even more switches between sources. More time in sources, more time in writing, more switches between sources and sources and text, but slower text production: we interpret these tendencies as reflecting a focus on coherence and integration in the first stage of the synthesis process. This development, however, does not yet show itself in higher overall text quality as a result of the observation task.

Comparing the effects of the two conditions, it seems that adding a short observation task reinforces the effect of process feedback on the writing process. Both conditions generate an effect on Text Quality, but the observation condition reinforces the effects on writing process behaviors. We can only qualitatively interpret those effects on writing behaviors: indeed, no correlations were observed between writing behaviors and text quality in this study. It is not the case that increases in certain activities (source time) and decreases in others (speed) are associated with text quality. Spending more source time in the initial phase of the synthesis process, as well as switching between sources and between sources and the text-so-far are often suggested in intervention studies (Van Ockenburg et al., 2019). But in the present study, we found no support for the effect of such strategies on text quality. From studies on learning by examples, it is known that such interventions lower the cognitive load during task execution (Van Gog & Rummel, 2010; Renkl, 2014). If this would have been the case in the present study, participants could have paid more attention to the quality of task execution, which might have resulted in better text quality, which is, however, not demonstrated in the present study.

The relation between task behaviors and output quality is intriguing. When learning conditions affect both task behaviors as well as output quality, and no relation between these variables is observed, some other variables might be in play. In the present study, students also got insight in the text quality of their paper related to the national scale; in many cases, there must have been room for improvement. Even if no information was provided about the particular aspect(s) which could have been improved, the mere fact that other texts were rated higher might have motivated and incited students to invest in writing. Future studies might include indicators of motivation and effort in the

design to study the mediating effect of motivation on text quality and task processes. Another perspective on this matter might be the compensatory relations between writing behaviors on an individual level (Rijlaarsdam & Van de Bergh, 1996). Although on a group level an effect on text quality and writing behaviors was observed, the individual patterns of investment in writing behaviors must have been different, provided the low correlations within the set of writing behaviors. Van Steendam et al. (2022) showed that students varied their writing behaviors quite frequently from task to task. In the present study, the exemplars and models students read and observed varied clearly, and yet they all focused on comprehending the source materials and starting the text production process. If students indeed differed in the way they combined these processes, scores on single behaviors may indeed not predict text quality. Note that in the present study, the process feedback was offered in a non-normative approach, as three different exemplars were shown, and students were free to choose whether or not to set goals to adapt their writing behaviors and to decide which goals suited them best. In terms of effects, this approach seems to have worked well, and at this point, writing process research cannot offer more.

Limitations

We aimed to carry out an intervention study with a high ecological validity. The intervention was situated in a classroom setting, and materials and tools are freely available for teachers to use in their classes (see website www.liftwritingresearch. wordpress.com). However, a number of reservations can be made about this study. We believe that the intervention's ecological validity would increase if future studies could address the following limitations.

That the students in the experiment were found to be better writers than the national sample may indicate a selection effect of the schools that volunteered to participate. It could indicate that the condition effects we report relate only to students who are among the better writers. Future (replication) studies could shed light on the further generalizability of the study's results to the population in question, for example, to initially weaker writers.

Secondly, the process feedback and video models in this intervention study were oriented towards one specific genre of writing, namely source-based writing. This implies that we cannot generalize findings for other writing task genres.

Another limitation that can be commented is the short duration of the intervention. Though the brevity of the intervention makes its deployment in classroom practice feasible, we currently do not know what the long-term effect of the intervention might be. Following up with a delayed post-test would be interesting. Moreover, more extensive intervention programs could be explored, e.g. devoting more time to certain activities, and involving teacher instruction and peer collaboration.

Further research

On a group level, the interventions were effective on writing behaviors and text quality in a non-specific content domain. In that respect, the findings contribute to our understanding of the effect of process feedback and additional observation on synthesis writing with secondary school students in such a non-specific content domain. Synthesis tasks, however, are often set in disciplinary domains. A replication of this study with content knowledge as an outcome, in different disciplines, could shed light on the effects of such a kind of neutral process feedback on source-based writing processes that are so different across disciplines, even within individuals (Holdinga et al., 2021). In line with Van Ockenburg et al. (2021), further research could explore the effects of moderator variables, such as writing process approach, writing attitude and self-efficacy, since students in the present study had to invest quite some effort in understanding the feedback that was offered.

Additionally, a future study could include not only the effect of self-efficacy and writing process approach as moderator variables in both feedback conditions but could also study the effect of the feedback on students' appreciation. We know from previous studies that feedback appreciation is an important factor for students' feedback literacy (Carless & Boud, 2018). A study that focuses on how students evaluate the intervention could shed light on the relation between the evaluation and writing process and product. That way the study's contribution to both the feedback and writing domain would be significantly heightened. Additionally, an analysis of students' feedback appreciation would highlight the strong and weaker points of the intervention and would thus provide valuable insights to further develop the intervention activities and materials.

A deeper analysis of the rationale behind students' model selection and an analysis of their written responses on the goal-setting questions, could provide us with an insight into the extent to which changes were in alignment with the goals students set themselves. Moreover, a detailed analysis of the keystroke logging data for a number of selected cases could shed light on the complexity of the changes in writing process behaviors. Forming a more integrative view of the writing process by taking into account various process behaviors and their interactions is essential to understand changes in writing process behaviors, and their relations with writing performance.

Nevertheless, despite the fact that deeper and more fine-grained qualitative analyses are needed to further explain the study's results, the study contributes to a rapidly expanding field of process feedback studies in writing research. To a certain extent, the study, along with Van Steendam et al. (2022), illustrates that "feedback aiming to expose students to different writing approaches by for example, having them contrast them or compare them to their own writing approach, could benefit awareness-building and reflection and ultimately result in learning" (Van Steendam et al., 2022, p. 1792). Additionally, the fact that it does so by relating its results to a larger national baseline study, enables researchers to position and interpret the findings in the light of general performance in both synthesis writing process and product of higher-secondary school students.

As this effective intervention is relatively short, requires limited effort from teachers, and the tools (i.e., user-friendly keystroke logging report generator) and materials are available, it can be implemented in regular classroom practice. There are many options for further research to build on this study and to continue exploring writing process feedback and its implementation in the classroom. Some options include exploring the long-term effects, adding additional instructional components, and combining process- with product-oriented feedback.

Appendix A: Check for differences between conditions at pretest

See Table 4.

Table 4 Parameter estimates for the differences between the FO-condition and F-condition at Session 1 for all dependent variables based on a multivariate model (fitted within the Bayesian framework in a similar way as the main analyses) and results of independent t-tests

Parameter	Med	CI_l	CI_h	t-test results
Text Quality	−1.62	−13.74	10.25	$t = 0.60$; df $= 164.85$; $p = 0.55$
Active writing time	0.03	−0.02	0.07	$t = -0.86$; df $= 131.06$; $p = 0.40$
Source time	−4.79	−11.82	2.34	$t = 0.$-88; df $= 140.73$; $p = 0.38$
Switches between sources	0.03	−0.11	0.18	$t = -0.19$; df $= 123.29$; $p = 0.85$
Switches sources-synthesis	−0.10	−0.33	0.14	$t = 0.40$; df $= 138.38$; $p = 0.69$
Speed	0.03	−0.02	0.08	$t = 0.95$; df $= 135.48$; $p = 0.35$

Med. = Median of posterior distribution; CI_l = lower limit of an 89% highest density interval; CI_h = upper limit of an 89% highest density interval; S1 – S3 = Session1 – Session3

References

Baaijen, V. M., & Galbraith, D. (2018). Discovery through writing: Relationships with writing processes and text quality. *Cognition and Instruction, 36*(3), 1–25. https://doi.org/10.1080/07370008.2018.1456431

Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ, 1986 (23–28).

Bowen, N. E. J. A., Thomas, N., & Vandermeulen, N. (2022). Exploring feedback and regulation in online writing classes with keystroke logging. *Computers and Composition*. https://doi.org/10.1016/j.compcom.2022.102692

Braaksma, M. A. H., Rijlaarsdam, G., Van Den Bergh, H., & Van Hout-Wolters, B. H. A. M. (2004). Observational learning and its effects on the orchestration of writing processes. *Cognition and Instruction*. https://doi.org/10.1207/s1532690Xci2201_1

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs the method of paired comparisons. *Biometrika, 39*, 324–345. https://doi.org/10.2307/2334029

Breetvelt, I., Van den Bergh, H., & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: When and how? *Cognition and Instruction, 12*(2), 103–123. https://doi.org/10.1207/s1532690xci1202_2

Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*, 1315–1325.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Couzijn, M. (1999). Learning to write by observation of writing and reading processes: Effects on learning and transfer. *Learning and Instruction, 9*(2), 109–142. https://doi.org/10.1016/s0959-4752(98)00040-1

De Smedt, F., Van Keer, H., & Merchie, E. (2016). Student, teacher and class-level correlates of Flemish late elementary school children's writing performance. *Reading and Writing, 29*, 833–868. https://doi.org/10.1007/s11145-015-9590-z

Dixon, H., Hawe, E., & Hamilton, R. (2020). The case for using exemplars to develop academic self-efficacy. *Assessment and Evaluation in Higher Education, 45*(3), 460–471. https://doi.org/10.1080/02602938.2019.1666084

Dux Speltz, E., & Chukharev-Hudilainen, E. (2021). The effect of automated fluency focused feedback on text production. *Journal of Writing Research, 13*(2), 231–255. https://doi.org/10.17239/jowr-2021.13.02.02

Dux Speltz, E., Roeser, J., & Chukharev-Hudilainen, E. (2022). Automating individualized, process-focused writing instruction: A design-based research study. *Frontiers in Communication*. https://doi.org/10.3389/fcomm.2022.933878

Gelman, A., Hill, J. & Vehtari, A. (2021). *Regression and other stories. Analytical methods for social research.* Cambridge University Press. Doi: https://doi.org/10.1017/9781139161879.

Graham, S., Gillespie, A., & McKeown, D. (2013). Writing: Importance, development, and instruction. *Reading and WritIng, 26*, 1–15. https://doi.org/10.1007/s11145-012-9395-2

Graham, S., & Harris, R. K. (2000). The role of self-regulation and transcription skills in writing and writing development. *Educational Psychologist, 35*(1), 3–12. https://doi.org/10.1207/S15326985EP3501_2

Graham, S., Harris, K. R., & Mason, L. (2005). Improving the writing performance, knowledge, and self-efficacy of struggling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology, 30*, 207–241. https://doi.org/10.1016/j.cedpsych.2004.08.001

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104*(3), 879–896. https://doi.org/10.1037/A0029185

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*(3), 445–476. https://doi.org/10.1037/0022-0663.99.3.445

Groenendijk, T., Janssen, T., Rijlaarsdam, G., & van den Bergh, H. (2013). The effect of observational learning on students' performance, processes, and motivation in two creative domains.

British Journal of Educational Psychology, 83(1), 3–28. https://doi.org/10.1111/j.2044-8279.2011.02052.x

Hayes, J. R., & Flower, L. S. (1980). Identifying the organisation of the writing process. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Lawrence Erlbaum Associates Inc.

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

Hendry, G. D., Bromberger, N., & Armstrong, S. (2011). Constructive guidance and feedback for learning: The usefulness of exemplars, marking sheets and different types of feedback in a first year law subject. *Assessment and Evaluation in Higher Education, 36*(1), 1–11. https://doi.org/10.1080/02602930903128904

Holdinga, L., Janssen, T., & Rijlaarsdam, G. (2021). The relationship between students' writing process, text quality, and thought process quality in 11th-grade history and philosophy assignments. *Written Communication, 38*(4), 544–586. https://doi.org/10.1177/07410883211028853

Lambert, Ben. *A student's guide to Bayesian statistics.* SAGE, 2018.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: using inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Comparative judgment as a promising alternative to score competences. In E. Cano & G. Ion (Eds.), *Innovative practices for higher education assessment and measurement* (pp. 119–138). IGI Global. https://doi.org/10.4018/978-1-5225-0531-0.

Lindgren, E. (2004). The uptake of peer-based intervention in the writing classroom. In G. Rijlaarsdam, H. Van den Bergh, & M. Couzijn (Eds.), *Effective Learning and Teaching of Writing* (pp. 259–274). Kluwer Academic Publishers. https://doi.org/10.5117/9789053561980

Lindgren, E., & Sullivan, K. P. H. (2003). Stimulated recall as a trigger for increasing noticing and language awareness in the L2 writing classroom: A case study of two young female writers. *Language Awareness, 12*(3–4), 172–186. https://doi.org/10.1080/09658410308667075

Lindgren, E., Sullivan, K. P. H., Deutschmann, M., & Steinvall, A. (2009). Supporting learner reflection in the language translation class. In M. Chang & C.-W. Kuo (Eds.), *Handbook of Research on Computer Enhanced Language and Culture Learning* (pp. 21–40). IGI Global. https://doi.org/10.4018/jicthd.2009070102

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review, 66*, 81. https://doi.org/10.1037/h0043178

MacArthur, C. A., Philippakos, Z. A., & Ianetta, M. (2015). Self-regulated strategy instruction in college developmental writing. *Journal of Educational Psychology, 107*, 855–867. https://doi.org/10.1037/edu0000011

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of effect existence and significance in the bayesian framework. *Frontiers in Psychology, 10*, 2767. https://doi.org/10.3389/fpsyg.2019.02767

Masui, C., & De Corte, E. (2005). Learning to reflect and to attribute constructively as basic components of self-regulated learning. *British Journal of Educational Psychology, 75*, 351–372. https://doi.org/10.1348/000709905X25030

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in r and stan* (2nd ed.). CRC Press. https://doi.org/10.1201/9780429029608

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

Nicol, D., & MacFarlane-Dick, D. (2006). Formative assessment and selfregulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking Criteria in peer and Self-assessment. *Assessment and Evaluation in Higher Education, 27*(4), 309–323. https://doi.org/10.1080/0260293022000001337

Panadero, E., Lipnevich, A., & Broadbent, J. (2019). Turning Self-Assessment into Self-Feedback. In *The Impact of Feedback in Higher Education: Improving Assessment Outcomes for Learners*, edited by Henderson, M.D. , Ajjawi, R., Boud, D., & Molloy, E. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-25112-3_9

Quinton, S., & Smallbone, T. (2010). Feeding forward: Using feedback to promote student reflection and learning – a teaching model. *Innovations in Education and Teaching International, 47*(1), 125–135. https://doi.org/10.1080/14703290903525911

Raedts, M., Rijlaarsdam, G., Van Waes, L., & Daems, F. (2007). Observational learning through video-based models: Impact on students' accuracy of self-efficacy beliefs, task knowledge and writing performances. In G. Rijlaarsdam (Series Ed.), & P. Boscolo & S. Hidi (Vol. Eds.), *Studies in Writing. Vol. 19. Writing and Motivation* (pp. 219–238). Elsevier. https://doi.org/10.1163/9781849508216_013

Ranalli, J., Feng, H. H., & Chukharev-Hudilainen, E. (2018). Exploring the potential of process-tracing technologies to support assessment for learning of L2 writing. *Assessing Writing, 36*, 77–89. https://doi.org/10.1016/j.asw.2018.03.007

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science, 38*(1), 1–37. https://doi.org/10.1111/cogs.12086

Rijlaarsdam, G., Braaksma, M., Couzijn, M., Janssen, T., Raedts, M., van Steendam, E., Toorenaar, A., & van den Bergh, H. (2008). Observation of peers in learning to write: practice and research. *Journal of Writing Research, 1*(1), 53–83. https://doi.org/10.17239/jowr-2008.01.01.3

Rijlaarsdam, G. C. W., & van den Bergh, H. (2006a). Writing process theory: A functional dynamic approach. In C. A. Macarthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 41–53). The Guilford Press.

Rijlaarsdam, G., & van den Bergh, H. (1996). The dynamics of composing—An agenda for research into an interactive compensatory model of writing: Many questions, some answers. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 107–125). Lawrence Erlbaum Associates Inc.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*, 191–209. https://doi.org/10.1080/0305498870130207

Scheiter, K. (2020). Embracing complexity in research on learning from examples and from problem solving. *Applied Cognitive Psychology, 34*(4), 906–911. https://doi.org/10.1002/acp.3702

Sullivan, K., & Lindgren, E. (2002). Self-assessment in autonomous computer-aided second language writing. *ELT Journal*. https://doi.org/10.1093/elt/56.3.258

Rijlaarsdam, G. C. W., & van den Bergh, H. (2006b). Writing process theory: A functional dynamic approach. In C. A. Macarthuur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 41–53). The Guilford Press.

van den Bergh, H., Rijlaarsdam, G., & van Steendam, E. (2016). Writing process theory: A functional dynamic approach. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 57–71). The Guilford Press.

van den Bergh, D., Vandermeulen, N., Lesterhuis, M., De Maeyer, S., Van Steendam, E., Rijlaarsdam, G., & Van den Bergh, H. (2023). How prior information from national assessments can be used when designing experimental studies without a control group. *Journal of Writing Research, 14*(3), 447–469. https://doi.org/10.17239/jowr-2023.14.03.05

Vandermeulen, N., De Maeyer, S., Van Steendam, E., Lesterhuis, M., van den Bergh, H., & Rijlaarsdam, G. (2020a). Mapping synthesis writing in various levels of Dutch upper-secondary education. A national baseline study on text quality, writing process and students' perspectives on writing. *Pedagogische Studiën, 97*(3), 187–236

Vandermeulen, N., Leijten, M., & Van Waes, L. (2020b). Reporting writing process feedback in the classroom: Using keystroke logging data to reflect on writing processes. *Journal of Writing Research, 12*(1), 109–140. https://doi.org/10.17239/jowr-2020.12.01.05

Vandermeulen, N., van den Broek, B., Van Steendam, E., & Rijlaarsdam, G. (2020c). In search of an effective source use pattern for writing argumentative and informative synthesis texts. *Reading and Writing, 33*(2), 239–266. https://doi.org/10.1007/s11145-019-09958-3

Vandermeulen, N., Van Steendam, E., De Maeyer, S., & Rijlaarsdam, G. (2023). Writing process feedback based on keystroke logging and comparison with exemplars: Effects on the quality and process of synthesis texts. *Written Communication, 40*(1), 90–144. https://doi.org/10.1177/07410883221127998

Van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22*, 155–174. https://doi.org/10.1007/s10648-010-9134-7

van Ockenburg, L., van Weijen, D., & Rijlaarsdam, G. (2019). Learning to write synthesis texts: A review of intervention studies. *Journal of Writing Research, 10*(3), 401–428. https://doi.org/10.17239/jowr-2019.10.03.01

van Ockenburg, L., van Weijen, D., & Rijlaarsdam, G. (2021). Choosing how to plan informative synthesis texts: Effects of strategy-based interventions on overall text quality. *Reading and Writing*, pp. 1–27.

Van Steendam, E., Vandermeulen, N., De Maeyer, S., Lesterhuis, M., van den Bergh, H., & Rijlaarsdam, G. (2022). How students perform synthesis tasks: an empirical study into dynamic process configurations. *Journal of Educational Psychology, 114*(8). https://doi.org/10.1037/edu0000755

Winston, N., & Carless, D. (2020). *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge.

Zimmerman, B. J., & Kitsantas, A. (1999). Acquiring writing revision skill: Shifting from process to outcome self-regulatory goals. *Journal of Educational Psychology, 91*(2), 241–250. https://doi.org/10.1037/0022-0663.91.2.241

Zimmerman, B. J., & Kitsantas, A. (2002). Acquiring writing revision and self-regulatory skill through observation and emulation. *Journal of Educational Psychology, 94*(4), 660–668. https://doi.org/10.1037/0022-0663.94.4.660

Authors and Affiliations

Nina Vandermeulen[1,2] · **Elke Van Steendam[3]** · **Sven De Maeyer[4]** · **Marije Lesterhuis[5]** · **Gert Rijlaarsdam[6,7]**

✉ Nina Vandermeulen
   nina.vandermeulen@umu.se

   Elke Van Steendam
   elke.vansteendam@kuleuven.be

   Sven De Maeyer
   sven.demaeyer@uantwerpen.be

   Marije Lesterhuis
   mlesterh@umcutrecht.nl

   Gert Rijlaarsdam
   g.c.w.rijlaarsdam@uva.nl

[1] Department of Language Studies, Umeå University, Umeå, Sweden

[2] Department of Management, University of Antwerp, Antwerp, Belgium

[3] Department of Linguistics, Louvain, KU, Belgium

[4] Department of Training and Education Sciences, University of Antwerp, Antwerp, Belgium

[5] Center for Research and Development of Health Professions Education, UMC Utrecht, Utrecht, The Netherlands

[6] Department of Social and Behavioral Sciences, University of Amsterdam, Amsterdam, The Netherlands

[7] Department of Teacher Education, Norwegian University of Science and Technology, Trondheim, Norway