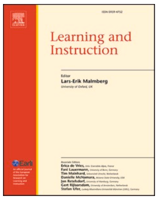
Learning and Instruction 92 (2024) 101922

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc



Improving (meta)comprehension: Feedback and self-assessment

Stephanie L. Hepner [a,*], Sophie Oudman [a], Trevor E. Carlson [b], Janneke van de Pol [a], Tamara van Gog [a]

[a] *Department of Education, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, the Netherlands*
[b] *School of Computing, National University of Singapore, 13 Computing Drive, 114717, Singapore*

ARTICLE INFO

Keywords:
Self-regulated learning
Text comprehension
Monitoring accuracy
Regulation
Feedback
Self-assessment

ABSTRACT

Background: Monitoring is important for self-regulated learning from text, but is often inaccurate. Completing causal diagrams after reading texts has been shown to improve monitoring accuracy.
Aims: We investigated whether providing one or two model answer diagrams and self-assessment instructions would improve learners' monitoring accuracy, regulation accuracy, and text comprehension. Because little is known about how accurately learners who are reading in a language other than their home language monitor their comprehension, we also explored whether effects differed between readers who have English or another language as their home language.
Sample: Participants were 258 secondary school students at international schools in Singapore and Spain; 103 spoke a language other than English at home.
Methods: Participants read 4 texts, completed diagrams on these texts, monitored comprehension, took a first comprehension test, self-assessed their diagram under one of 6 conditions resulting from a 3 (model answer: 0, 1, 2) x 2 (self-assessment instructions: yes, no) design, made restudy decisions, made monitoring judgments, and completed a final comprehension test.
Results: Comprehension benefitted most when learners had access to two model answers. There were no effects of model answers or self-assessment instructions on monitoring accuracy. Regulation accuracy improved with model answers combined with self-assessment instructions. There was no differential effect of home language.
Conclusions: This study supports prior research showing the benefit of model answer diagrams on comprehension. Yet, improvements in regulation accuracy suggest that model answers combined with self-assessment instructions support more effective self-regulated learning behaviors.

1. Introduction

Reading with comprehension is an essential skill for full access to a knowledge society (reading news/social media, voting, etc.). Moreover, text comprehension is an essential skill for learning in secondary education, and one that students reading in a foreign language often struggle with. Reading comprehension requires complex self-regulation skills: effective readers create an accurate mental model of the text as they interpret and integrate what they read with their prior knowledge (Kintsch, 1998, 2005; Kim, 2017), which requires that they monitor the quality of their mental model to ensure they understand what they read, and regulate their learning behaviors accordingly, by restudying (parts of) a text that are not yet understood well (Dunlosky & Rawson, 2012; Maki & Berry, 1984; Prinz et al., 2020a; Thiede et al., 2019).

However, students' comprehension monitoring (or metacomprehension; Dunlosky & Lipko, 2007) is often inaccurate, which results in suboptimal regulation (Thiede & Dunlosky, 1999) and consequently in remembering less information from the texts (Rawson et al., 2011). Inaccurate monitoring often results from making judgments based on surface-level cues that are not predictive of actual

[*] Corresponding author.
E-mail addresses: s.l.hepner@uu.nl (S.L. Hepner), v.s.oudman@uu.nl (S. Oudman), tcarlson@comp.nus.edu.sg (T.E. Carlson), j.e.vandepol@uu.nl (J. van de Pol), t.vangog@uu.nl (T. van Gog).

https://doi.org/10.1016/j.learninstruc.2024.101922
Received 10 February 2023; Received in revised form 25 March 2024; Accepted 1 April 2024
Available online 21 May 2024

comprehension test performance, such as interest in the text topic, or ease of reading (Koriat, 1997). Generative activities (Fiorella & Mayer, 2016), such as creating or completing a diagram that represents the causal relations described in the text (van de Pol et al., 2020; van Loon et al., 2014), give learners cues about the quality of their mental model (Kintsch, 1998) and thus improve monitoring accuracy (van de Pol et al., 2019; 2020; 2021; Van Loon et al., 2014; Yang et al., 2023).

While diagramming improves monitoring accuracy, its effects are only moderate (Prinz et al., 2020b). Feedback in the form of model answers (e.g., a correctly completed diagram) that students can use to self-assess their own diagram, providing more predictive cues about the quality of their own text representation, improves monitoring accuracy in other types of tasks (Follmer & Tise, 2021; Froese & Roelle, 2022; Rawson & Dunlosky, 2007; Van Loon & Roebers, 2017). A recent study suggests this also applies when diagramming immediately after reading a text (Braumann et al., 2024a). The present study builds on those findings to investigate if students' text comprehension, monitoring accuracy, and regulation accuracy would further benefit from receiving *multiple* standards (i.e., correct and partially correct) and explicit *self-assessment instructions* when diagramming at a *delay* (i.e., after reading several texts). We also explore if these effects differ for students whose home language is not the language of instruction.

1.1. Self-regulated learning from texts: monitoring and regulation

Self-regulated learning from texts requires readers to accurately monitor their comprehension of the text in order to inform their subsequent learning behaviors (i.e., regulation) (Thiede & Anderson, 2003). Readers show good comprehension of a text when they create an accurate mental model (*situation model*) of the gist of the text (Kintsch, 1998), based on their interpretation of the words and sentences, integrated with their prior knowledge.

However, students from elementary through tertiary education typically struggle to accurately monitor how well they understand what they read (Prinz et al., 2020a). *Monitoring accuracy*, measured by comparing learners' judgments of how well they understand a text (*Judgments of Learning,* or JoLs) with their actual performance on a comprehension test about that text, has been found to be rather low (Maki, 1998; $r = 0.24$ in Prinz et al., 2020a; *mean correlation* = 0.242 in Yang et al., 2023). Typically, learners overestimate their understanding (Follmer & Tise, 2021; Foster et al., 2016; Griffin et al., 2019; Lipko et al., 2009). Without accurate monitoring, learners struggle to make accurate regulation decisions.

Regulation refers to when learners adapt, or plan to adapt, their learning behavior or strategies, ideally based on the insights obtained through monitoring (Griffin et al., 2013). Regulation behavior may include the decision to terminate studying a text if it is well-understood or to restudy a text to create a stronger mental model if it is not yet well-understood. *Regulation accuracy* refers to how well learners' regulation decision matches their comprehension. Learners with high regulation accuracy make regulation decisions that match their (judgments of their) comprehension; that is, there are two measures of regulation accuracy, one that captures how well regulation decisions match actual comprehension (*comprehension-based regulation accuracy*) and one that shows how well regulation decisions match monitoring judgments of comprehension (*monitoring-based regulation accuracy*). The latter indicates if students use their monitoring judgments to inform their regulation decisions; the former indicates whether their regulation decisions match their actual learning needs. In case of fully accurate monitoring, the values of both regulation accuracy measures would be the same. Learners show low regulation accuracy when they (judge they) do not understand a text yet decide not to restudy it, or when they (judge they) do understand a text well but decide to spend valuable study time restudying that same text. Thus, for effective self-regulated learning, students also should have high regulation accuracy, and accurate monitoring seems a pre-condition for accurate regulation.

1.2. Improving monitoring accuracy

Understanding the theoretical basis of monitoring accuracy ensures interventions designed to improve monitoring accuracy target the relevant mechanisms. The *Situation-Model Approach to Metacomprehension* (Yang et al., 2023) relies on Koriat's (1997) cue utilization framework, which posits that learners do not have direct insight into their actual comprehension, but rather, rely on various cues when monitoring their comprehension. Some cues, such as a learner's ability to summarize the gist of a text, are highly predictive, or *diagnostic*, of their comprehension, whereas others, such as ease of reading, personal interest in a topic, size of font, are not (Griffin et al., 2019; Thiede et al., 2010). If learners' monitoring judgments are inferential and based on a variety of predictive and non-predictive cues, then some judgments will be more accurate than others, depending on the diagnosticity of the cues the learner used to make that specific judgment (Koriat, 1997). When learners have access to cues that give them insight into their mental model of the text (*situation model;* Kintsch, 1998), their metacomprehension judgments should be more accurate.

Given the Situation Model Approach to Metacomprehension, diagramming, and especially delayed diagramming, should improve monitoring accuracy because completing a diagram of the causal relations described in the text gives learners access to diagnostic cues regarding their situation model of the text (e.g., when they cannot complete all boxes, they know they have not sufficiently understood the text). This is especially so when diagramming is done at a delay, because the diagram is completed based on the mental model held in long-term memory, not the surface features which would have degraded from short term memory due to the delay. Indeed, adolescents who read texts and then created diagrams of the underlying causal structure of texts at a delay (i.e., after reading all six texts) had improved monitoring accuracy compared to those who created diagrams immediately after reading a text and also compared to those who did not complete diagrams (van Loon et al., 2014). This finding has been replicated in a study by Van de Pol et al. (2019), which found that both drawing diagrams and completing diagrams at a delay improved monitoring accuracy over no diagram control conditions.

Delayed diagramming has a moderate impact on monitoring accuracy ($g = 0.72;$ Prinz et al., 2020b), and there is room for further improvement. An additional intervention that has been found to be effective with other types of generative activities to support monitoring accuracy is to provide feedback on the generated product in the form of standards, or model answers. Such feedback provides learners with more diagnostic cues about the quality of their own generated product, as they compare their response to the standard and use it to improve comprehension (Braumann et al., 2024a). Feedback has been shown to improve monitoring accuracy in several studies, for instance: Using correct answer feedback to evaluate middle school students' key term definitions (Lipko et al., 2009); providing correctness feedback on a reading comprehension test (Follmer & Tise, 2021); providing feedback on secondary students' self-assessment of their problem-solving skills (Baars et al., 2014); and using full definitions to self-assess key term learning (Dunlosky et al., 2011). Recently, Braumann et al. (2024a) found that immediate diagramming with a model answer standard improved monitoring accuracy and text comprehension in 18–23 year olds. This could be because feedback supports learners in engaging with their situation model. Given that delayed diagramming gives learners access to cues about their situation model and results in higher monitoring accuracy than immediate diagramming, it is important to know whether feedback further improves monitoring accuracy over and above the moderate improvement already provided by delayed diagramming.

To benefit from model answer feedback, learners should actively engage with the feedback and compare their answers with the model answers (Nicol, 2020). Yet not all studies explicitly instructed them to do so (exceptions: Dunlosky et al. (2011) and Lipko et al. (2009)). It is unclear to what extent explicit self-assessment comparison instructions impact the effectiveness of feedback interventions in improving monitoring accuracy.

There is evidence that having access to multiple examples of correct answer feedback may support learners in engaging better with the feedback (Nicol, 2020) and improving subsequent work (Lin-Siegler et al., 2015), particularly for generative writing activities. Learners are thought to engage with feedback through a comparison process (Butler & Winne, 1995; Nicol, 2020), and having multiple model answers allows them to compare their response to the correct standard and also the partially correct model to the correct standard. These multiple comparisons could result in more engagement with the standard and could generate more diagnostic cues as learners examine their own and the partially correct model for differences with the correct standard. However, there is no empirical evaluation of the differential impact of no, one or two different model answers on monitoring accuracy.

Moreover, little is known about the impact of feedback on regulation accuracy. As the desired outcome of improved monitoring accuracy is better regulation accuracy (which in self-regulated learning situations would ultimately lead to better comprehension), prior findings about the effect of generative activities on regulation accuracy are also relevant. Previous studies have either not calculated regulation accuracy, or shown readers to have high regulation accuracy which is not significantly impacted by delayed diagramming interventions (van de Pol et al., 2019; van Loon et al., 2014).

Finally, it is an open question whether such interventions are less, equally, or more effective for students who have a home language other than the language of the intervention. Very little research on monitoring and regulation has addressed this issue. There is evidence that language proficiency correlates with performance on language-based tasks (e.g., OECD, 2012) and there is evidence of a bilingual advantage for higher-order skills (Adesope et al., 2010; Grundy & Timmer, 2017). Like other higher-order skills, monitoring accuracy is often measured using verbal instructions (Buehler et al., 2021) and may therefore be affected by language proficiency (Ebert, 2015). Given the interplay between language skills and metacognition, it is not clear whether and how having a home language other than the language of instruction might impact monitoring accuracy. A prior study that did examine monitoring accuracy of learners reading in a second language showed that students answered more comprehension questions correctly when working in their home language, but found no differences in monitoring accuracy between students working in their home or a second language (Buehler et al., 2021).

1.3. *The present study*

Accurately monitoring comprehension of text is an important skill in secondary education, especially as students prepare to transition to tertiary education. We investigated to what extent self-assessment instructions and the number of model answers individually and jointly affect students' text comprehension, monitoring, and regulation accuracy. We used a typical metacomprehension accuracy paradigm (Griffin et al., 2019; Prinz et al., 2020a; van de Pol et al., 2020), specifically the delayed diagramming design (van de Pol et al., 2019; 2021; van Loon et al., 2014) and modified it by adding a novel comparative self-assessment phase that empirically assesses the impact of multiple model answers and comparison-based self-assessment instructions. Thus, this study builds on the literature on improving self-monitoring and self-regulation accuracy by means of generative activities, in particular, diagramming, and makes three new contributions to it, by evaluating whether*:* 1) feedback via model answers, and especially multiple model answers, improves text comprehension, monitoring accuracy and regulation accuracy with delayed diagramming; 2) explicit self-assessment instructions further improve comprehension, monitoring and regulation accuracy; and 3) alignment of home language and language of instruction has an impact on the efficacy of the intervention.

In the present study, all secondary school students read four different texts in English, completed diagrams of the relationships presented in the texts, judged their understanding, and completed a comprehension test of causal relations. In the self-assessment phase, participants then received no, one (correct), or two (correct and partially correct) model answers and were provided with their diagram along with the model answers (no self-assessment instructions) or were explicitly asked to compare their diagram to model answers (self-assessment instructions). Participants were then asked to indicate what updates they would make to their diagram based on their self-assessment and to make restudy decisions, after which they engaged in a second round of judgments of learning and a second test of inferences.

We addressed the following research questions (RQs):

> RQ1: To what extent does the availability of a correct model answer, or a combination of correct and partially correct model answers, and self-assessment instructions individually and jointly affect students' comprehension, monitoring, and regulation accuracy?

Based on McCrudden et al. (2007) and Braumann et al. (2024a), who found beneficial effects of receiving a (correct) model answer on monitoring accuracy and comprehension, we expected that availability of model answers would lead to higher text comprehension (H1a), higher monitoring accuracy (H2a), and higher regulation accuracy (H3a) compared to no model answers. Based on Nicol (2020), we furthermore expected that receiving two model answers of varying correctness would be more beneficial than one model answer for text comprehension (H1b), monitoring accuracy (H2b), and accuracy (H3b), as receiving one fully and one partially correct model answer provides more opportunities for learners to engage with the feedback and estimate the quality of their own diagram. We also expected an interaction effect with comparison instructions, such that explicitly prompting learners to compare their own and the model answers would lead to higher comprehension (H1c), monitoring accuracy (H2c) and regulation accuracy (H3c) compared to the other conditions.

> RQ2: How does home language (English or not English) impact how comparative self-assessment and the number of model answers individually and jointly impact comprehension, monitoring accuracy, and regulation accuracy?

Based on Buehler et al. (2021), we predicted that participants whose home language is English (the language of instruction at the school and the language of the intervention) have better comprehension than those whose home language is not English. We explored whether the intervention would support the monitoring and regulation accuracy of non-home language English students more than English home language speakers.

2. Methods

2.1. Participants and design

In total, 435 grade 11 and 12 students enrolled at English-language instruction international schools[1] in Singapore and Spain were invited to participate in the study through their English and/or psychology classes. The participating schools have highly diverse student populations, representing 100+ nationalities. An a priori power analysis for 3 × 2 ANOVA in G*Power (version 3.1.9.6; Faul et al., 2007) showed that 244 participants were needed to achieve a power of 0.80 with a medium-sized effect (0.20). The number of students who were present on the day of the study, who consented to participate in the study, and who completed the study during the class period was 258 (age: $M = 16.5$, $SD = 0.75$); we did not collect data on student absences and we did not collect data on how many students did not complete the study due to technical difficulties (e.g., wifi issues, page loading errors, etc.) vs other reasons. Because we did not expect gender differences in meta-comprehension accuracy based on the prior literature, we did not solicit information about participants' gender. 103 participants reported having a home language other than English. Students participated as part of their regular classes, so all students present participated in the study. However, active consent was obtained from students for use of their data, parents were informed at least one week in advance and gave passive consent for the use of their child's data (i.e., they could opt out). No parents opted out and three students did not consent to having their data used. Ethical approval for this study was obtained from the Ethics Committee of Utrecht University in October 2021.

The study had a 3 × 2 between-subjects design with factors Model Answer (0/1 (correct)/2 (correct, partially correct)) and Self-Assessment Instruction (yes/no). Participants were randomly assigned to one of the six conditions: No model answer and no self-assessment instructions (control) (n = 51); One model answer and no self-assessment instructions (n = 43); Two model answers and no self-assessment instructions (n = 42); No model answer and self-assessment instructions (n = 39); One model answer and self-assessment instructions (n = 47); Two model answers and self-assessment instructions (n = 36). The study consisted of two phases. The first phase followed the typical 'generation paradigm': Participants read texts; completed a causal diagram; judged their understanding; and completed a comprehension test. In the second phase, participants reviewed their diagram along with (depending on assigned condition) 0, 1, or 2 model answers and with or without explicit instructions to self-assess their diagram; identified updates for their diagram; made restudy decisions; judged their understanding; and again completed a comprehension test.

2.2. Materials

The materials were based on those used in van Loon et al. (2014) and van de Pol et al. (2019; 2021). All materials were presented in the Gorilla Experiment Builder online platform in English (http://www.gorilla.sc).

2.2.1. Initial instructions

The initial instructions presented participants with a visual overview of the steps of the study and a practice trial with all steps and instructions of their assigned condition.

2.2.2. Prior knowledge test

The prior knowledge test consisted of four questions, one open-answer question based on the topic of each text (e.g.,"In the United States, subway cars have been sunk into the ocean. What are some consequences of this?"). Each test question was presented on a separate screen.

2.2.3. Texts

Four texts on various science and social science topics were presented, each on a separate screen. The texts contain causal connections so readers can make causal inferences (Wiley et al., 2005). Three of the texts (Suez Canal, Botox, Subway Cars) were based on those used in previous delayed diagramming studies (van de Pol et al., 2019, 2021; van Loon et al., 2014); they were translated into English and the translations were checked by an expert on text comprehension and a home language English speaker. One additional text (Pesticides) was created for this study to ensure that there was a balance of diagram types represented (i.e., two texts required linear diagrams, two texts required serial diagrams; see Fig. 1). Average length of the four texts was 172.25 words ($SD = 4.2$). The Gorilla software randomized the order of the four texts for each participant (but individual participants received the same order in each phase of the study, i.e., the diagrams were in the same order as the texts). Two sample texts (corresponding to the diagrams in Fig. 1) along with the relevant comprehension test questions are provided in the Supplementary Materials.

2.2.4. Diagrams

Participants were provided with diagrams that included the correct number of nodes in the correct layout (linear or serial) based on the underlying causal structure of the text. In addition, the information for one node (either the first or final) was always provided. Each diagram required participants to complete four empty nodes with causally connected information from the relevant text. Participants were instructed to fill in the information in the remaining nodes and were also encouraged to include causal transition words or phrases in the diagram such as 'because' or 'therefore'. Each diagram was presented on a separate screen.

2.2.5. Judgments of learning

Participants were asked to judge their understanding of the causal relations presented in the texts (response options ranging from 0 to 4), using the following prompt: "How many points do you think you will receive on a test question assessing your comprehension of the connections between the different ideas in the text "TEXT TITLE"?" The judgment of learning for each text was presented on a separate screen and required a response. Judgments were made after completing the

[1] International schools refer to English-medium schools which offer internationally-recognized exit qualifications (e.g., International Baccalaureate Diploma, Advanced Placement). International Schools can be classified as Type A (traditional), which are those which were founded to educate the children of expatriates around the world; Type B (ideological), which are committed to education for global peace; and Type C (non-traditional), which are for-profit and serve the local (host country) upper and middle class families (Bunnell et al., 2016). International schools currently serve almost 7 million students around the world (ISC Research, 2024). The international schools involved in the study were traditional, established to provide an English-language education primarily to children of expatriates (or dual nationals), with one included school also educating students from other countries with a mission of educating for peace. The majority of students attending the schools are considered Third Culture Kids, defined as "a person who spends a significant part of his or her first 18 years of life accompanying parent(s) into a country that is different from at least one of the parent's passport country(/countries) due to a parent's choice of work or advanced training" (Pollock et al., 2017; cited in Tan et al., 2021). All included schools offered the same internationally-recognized qualification (International Baccalaureate Diploma Programme) and also accepted students with limited English proficiency as part of their admissions process. Given the similarities in terms of school type, student backgrounds, and educational program, the participants from the different schools are expected to be similar.

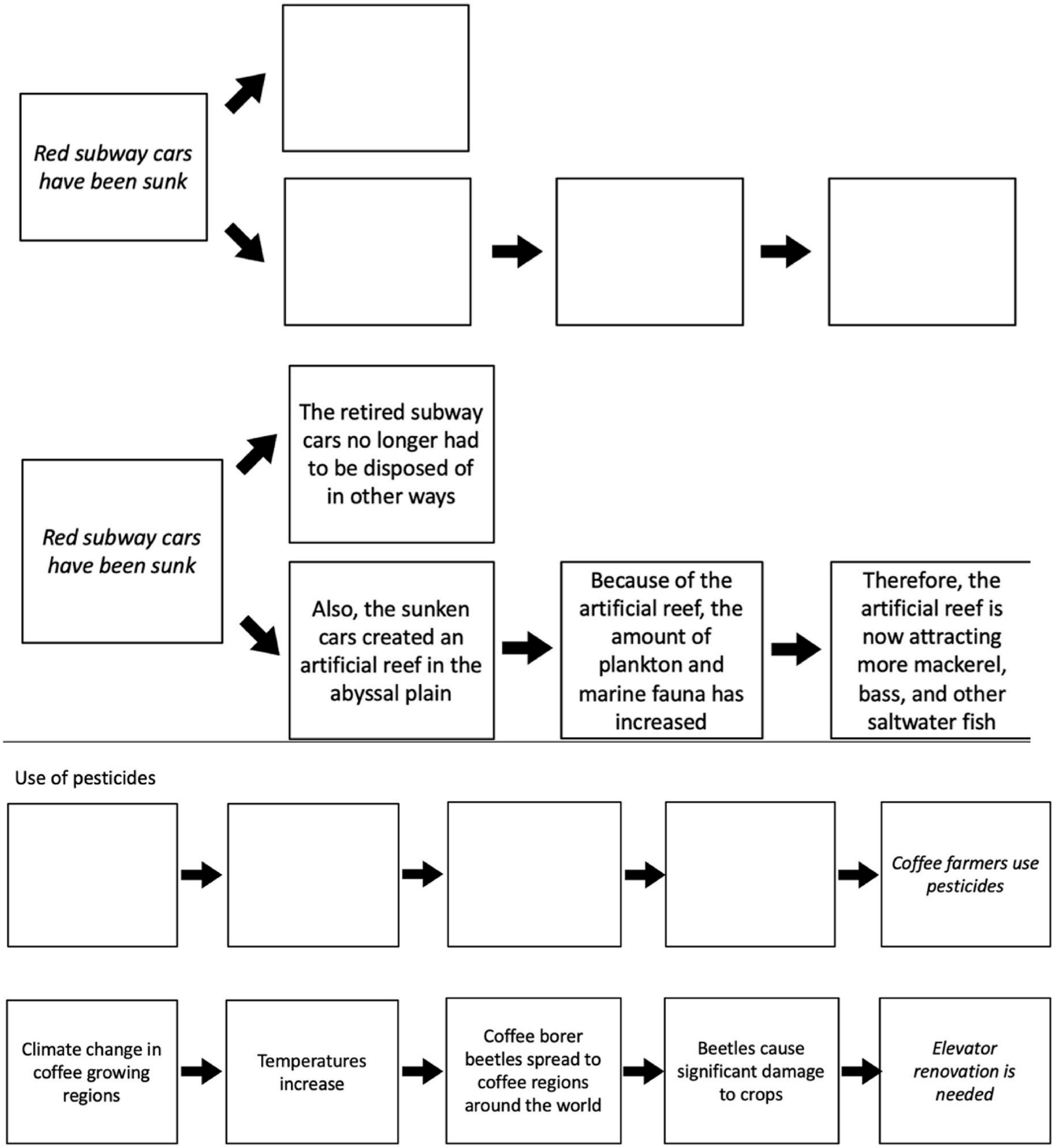

Red subway cars have been sunk
Red subway cars have been sunk
The retired subway cars no longer had to be disposed of in other ways
Also, the sunken cars created an artificial reef in the abyssal plain
Because of the artificial reef, the amount of plankton and marine fauna has increased
Therefore, the artificial reef is now attracting more mackerel, bass, and other saltwater fish
Use of pesticides
Coffee farmers use pesticides
Climate change in coffee growing regions
Temperatures increase
Coffee borer beetles spread to coffee regions around the world
Beetles cause significant damage to crops
Elevator renovation is needed

Fig. 1. Example serial and linear diagram.

diagram in phase 1 (JoL1) and after making restudy decisions in phase 2 (JoL2).

2.2.6. *Comprehension tests*

Participants completed two comprehension tests, one at the end of each phase (Test1 and Test2, respectively). These two comprehension tests were identical to each other and consisted of four questions in total, one question per text, each of which required participants to recall four causal relations (see Supplementary Materials). Each question appeared on a separate screen.

2.2.7. *Diagram review, model answers, and self-assessment instructions*

At the start of the second phase, all participants were presented with their own filled-out diagram to review. Depending on their assigned condition, they also received model answer diagrams (0,1,2) along with their own diagram, and received instructions (or not) to self-assess their diagram. When participants received zero model answers, they saw only their own completed diagram on the screen. When participants saw one model answer, it was a correct standard model answer presented alongside their own diagram. Participants were told “Here is a sample diagram for [TITLE] text. This diagram would receive 4 points (full marks).” When participants saw two model answers, one was the correct standard model answer and the other contained one commission error; these were again presented alongside their own diagram. In the model answer that contained an error, the text in the remaining nodes was correct, but was not identical to the correct standard diagram, thus requiring close reading of the diagrams to identify the commission error. They were told “Here are two sample diagrams for [TITLE] text. This diagram would receive 4 points (full marks). This diagram would receive 3 points.” The diagrams were presented in the same order between conditions, such that participants’ own diagrams were always displayed at the top of the screen (0, 1, and 2 model answer), the correct model answer was displayed below the participant response (1 and 2 model answers), and the partially correct model was displayed at the bottom of the screen (2 model answers).

Participants in the self-assessment instruction conditions received an additional prompt to self-assess their own diagram. Those who did not receive any model answers were prompted with "Please study your diagram. Be sure to compare your diagram to the task instructions." Those who received one or two model answer(s) were told "Please study the diagram(s). Be sure to compare your own diagram to the diagram that is provided." See Supplementary Materials for sample correct and partially correct diagrams.

2.2.8. Diagram updates

Participants were provided with their diagram and the model answers according to their assigned condition and were asked "What is one change you would make to this diagram?" followed by "Why would you make this change?", all on one screen. They had an option to make more changes, which would bring them to the same (blank) screen again. The maximum number of changes any participant made was 4.

2.2.9. Restudy decisions

Participants were told there would be a second test on the texts and "You will now have the chance to decide if you would like to restudy any of the texts before taking the test." On the next screen, they were provided with all four text titles and a yes/no selection for whether they would like to restudy the text. Then they were told "In the interest of time, you will not restudy the selected texts during this session," and were taken to the final judgments of learning and test.

2.3. Procedure

Students participated at their schools, as part of their regular classes. The study had the duration of one lesson period (ca. 75 min.) The participants accessed the study materials from their laptops in class. Participants were provided a link, provided (or withheld) consent for the use of their data, read the initial instructions and then completed a practice trial aligned with their assigned condition, which was randomly assigned in balanced mode (random without replacement) by the Gorilla software. They were given the opportunity to ask questions at this point if anything in the procedure was unclear to them. Then, all participants completed the prior knowledge test and read all four texts (order randomized between participants); they could not go back to review a previously-read text. Then participants completed the diagramming task, filling out the diagrams for each of the four texts (in the same order) without having access to the texts. After completing all four diagrams, participants were asked to make a judgment of learning on each of the four texts. Subsequently, they completed the comprehension test on all four texts. This was followed by the diagram review, during which, depending on assigned condition, participants were presented with zero, one, or two model answers and were instructed (or not) to self-assess their diagram. All participants were then asked, for each diagram, to indicate what updates they would make to their own diagrams and provide a reason for that update. Next, participants had the option to select texts for restudy, were asked to make judgments of learning again on each of the four texts, and finally, completed the comprehension test again.

2.4. Measures

For all measures, 10% of the participant responses were scored by two coders independently. Interrater agreement was good (see **α** values below; Krippendorff, 2004), so the remainder were coded by a single rater (first author).

2.4.1. Performance on tests

The prior knowledge tests were scored by assigning one point per correct idea unit from the diagram/comprehension test scoring guide (i.e., possible range: 0–4 points). We did not subtract points when incorrect information was provided (Krippendorff α = 0.80).

Comprehension tests were scored using coding schemes from other studies that used the same materials (e.g., van de Pol et al., 2021). On each test question, participants could receive 0–4 points, one point per correct idea unit that was causally connected to preceding ideas. Test responses were segmented into idea units and each segment was scored as correct, a commission error (incorrect information or repeated information), or an omission error (omitted information). Participants got credit for correct idea units and no credit for errors; we did not subtract points for errors (segments α = 0.85, correct elements α = 0.95, commission errors α = 0.87, correct sequences α = 0.88).

2.4.2. Monitoring accuracy

Students' absolute monitoring was computed by taking the absolute (i.e., unsigned) difference between a participant's judgment of learning about a text and their actual performance on the comprehension test on that text; absolute monitoring was computed twice per participant, once in each phase. Values range from 0 to 4; scores of 0 indicate perfect monitoring accuracy.

2.4.3. Regulation accuracy

Regulation accuracy can be defined in two different ways: monitoring-based or comprehension-based, as described in section 1.1. Monitoring-based regulation accuracy reflects the correlation between students' judgment of learning about a text and their restudy decision for that text (i.e., whether they select less-well understood texts for restudy). This measure shows to what extent the learner accurately uses their monitoring judgment to inform subsequent study activities. However, it does not reflect whether the restudy decision was accurate in the sense that it is what the learner should be doing, based on their actual performance. Thus, comprehension-based regulation accuracy reflects the correlation between students' actual performance on the comprehension test about a text and their decision to restudy that text.

We calculated absolute measures of both measures of regulation accuracy based on the second set of judgments of learning and restudy decisions (monitoring-based) and the second comprehension test scores and restudy decisions (comprehension-based) per text. We used the approach of Baars et al. (2014; also used in van de Pol et al., 2021): Low judgments of learning or test scores combined with the (desirable) decision to restudy that text resulted in higher accuracy (closer to 1), and higher judgments of learning or test scores combined with the (desirable) decision *not* to restudy that text resulted in higher accuracy (see Table 1).

2.5. Data analysis strategy

To analyze the effects of self-assessment instructions and model answers on students' test performance, monitoring and regulation accuracy, we fitted linear mixed models using the lme4 package and obtained the result using the anova () function, which provides *F* statistics based on sequentially decomposing the contributions of the fixed effects (Bates et al., 2015). The linear mixed models accounted for the nested data structure with texts (level 1) clustered in students (level 2), with number of model answers (0, 1, 2) and self-assessment instructions (yes, no) as fixed effects, as well as an interaction between model answers and self-assessment instructions. The adjusted intra-class coefficients (ICCs)

Table 1
Regulation accuracy conversion.

JoL2 or Test2 score:	Restudy Decision: No	Restudy Decision: Yes
0	0	1
1	0.25	0.75
2	0.5	0.5
3	0.75	0.25
4	1	0

Table 2
Descriptive statistics of main variables.

			No Self-Assessment Instructions			Self-Assessment Instructions		
		Overall	No Model Answers	1 Model Answer	2 Model Answers	No Model Answers	1 Model Answer	2 Model Answers
Number of Participants		258	51	43	42	41	47	36
Variable	Range	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Prior Knowledge	0–4[a]	0.18 (0.47)	0.19 (0.46)	0.20 (0.47)	0.20 (0.53)	0.15 (0.46)	0.13 (0.42)	0.21 (0.51)
JoL:Time One	0–4[a]	2.21 (1.11)	2.05 (1.12)	2.21 (1.10)	2.14 (1.11)	2.09 (1.10)	2.40 (1.04)	2.38 (1.19)
Test Score:Time One	0–4[a]	1.94 (1.44)	1.96 (1.41)	1.86 (1.45)	1.90 (1.38)	1.96 (1.49)	2.05 (1.51)	1.89 (1.36)
Word Count:Test 1	0–281	51.42 (30.37)	50.74 (30.88)	52.58 (36.05)	52.51 (27.33)	47.14 (29.44)	52.63 (30.16)	52.81 (26.73)
Absolute Monitoring Accuracy:Time 1	0–4[b]	1.09 (0.95)	0.95 (0.93)	1.20 (0.99)	1.10 (0.96)	1.08 (0.86)	1.17 (1.03)	1.04 (0.88)
JoL:Time Two	0–4[a]	2.40 (1.19)	2.18 (1.13)	2.47 (1.22)	2.40 (1.24)	2.15 (1.20)	2.53 (1.11)	2.68 (1.19)
Test Score:Time Two	0–4[a]	2.01 (1.56)	1.78 (1.45)	1.93 (1.68)	2.05 (1.57)	1.69 (1.50)	2.24 (1.62)	2.46 (1.42)
Word Count:Test 2	0–148	40.47 (26.68)	42.19 (29.91)	37.51 (28.26)	35.86 (22.67)	39.62 (26.94)	41.82 (26.58)	46.12 (22.74)
Absolute Monitoring Accuracy:Time 2	0–4[b]	1.13 (1.05)	1.11 (0.97)	1.32 (1.27)	1.17 (1.05)	1.09 (0.93)	1.06 (1.04)	1.01 (0.98)
Monitoring Based Regulation	0–1[c]	.57 (.31)	.53 (.28)	.61 (.31)	.59 (.31)	.54 (.30)	.50 (.31)	.68 (.29)
Comprehension Based Regulation	0–1[c]	.48 (.39)	.50 (.37)	.45 (.42)	.49 (.39)	.44 (.38)	.43 (.40)	.59 (.36)

[a] Higher scores indicate better comprehension (4 indicates perfect comprehension).
[b] Lower scores indicate better monitoring accuracy (0 indicates perfect monitoring accuracy).
[c] Higher scores indicate better regulation (1 indicates perfect regulation).

for all outcome variables indicate 24%–62% of the random variance in monitoring accuracy, comprehension, and regulation were due to differences between students. We used the maximum likelihood estimation with robust standard errors (lmer default, REML = true) which is robust to non-normality and which excludes missing data. To explore the effects of home language on our outcome variables, we re-ran the same models, adding home language as a fixed effect and as a third interaction term. We used the check_model () function (Luedecke et al., 2021) to confirm the data met the assumptions for all models. The dataset is available on OSF at https://osf.io/4dhxm/?view_only=6010f8c601ab478dac52d7faf6579026.

3. Results

3.1. Preliminary analyses

To understand how accurate participants' monitoring was overall, and whether they tended towards over- or underestimation, we conducted exploratory preliminary analyses on monitoring accuracy on the first test. Accuracy was relatively high in this sample. Of the monitoring judgments made at time 1, 30% were perfectly accurate, and 71% were highly accurate (monitoring accuracy of 0 or 1). In addition, almost 10% of participants (24 participants) had high monitoring accuracy (0 or 1) on all four texts at time 1. Nevertheless, there was a deviation of 1 or more on 70% of the judgments, and a mixed-effects model with the calculated difference between judgments of learning and performance as fixed effects at time 1 showed that on average, participants tended to over-estimate their understanding, meaning that their judgment of learning scores exceeded their test performance scores, $t(257) = 5.10$, $p < 0.001$. Table 2 provides descriptive statistics of all main variables, per condition; descriptive statistics for participants with a home language other than English are available in Supplementary Materials.

3.2. Effect of model answers and self-assessment instructions on text comprehension (H1)

There was a significant main effect of model answers on students' comprehension test performance in phase 2, F (2,252) = 3.45, p = .03. There was no significant main effect of Self-Assessment Instruction, F (1,252) = 1.52, p = .22, and no significant interaction effect, F (2,252) = 0.83, p = .44. As for the effects of model answers, pairwise comparisons showed that students who received two model answers performed significantly better on the second comprehension test than students who received no model answers, t (252) = −2.56, p = .03. No other comparisons were significant.

3.3. Effect of model answers and self-assessment instructions on monitoring accuracy (H2)

We found no significant main effect of model answers, F (2,252) = 0.43, p = .65, or self-assessment instructions, F (1,252) = 2.24, p = .14, nor a significant interaction effect, F (2,252) = 0.50, p = .61, on monitoring accuracy in phase 2.

3.4. Effect of model answers and self-assessment instructions on regulation accuracy (H3)

There was a small, significant main effect of model answers on monitoring-based regulation accuracy, F (2,252) = 5.14, p = .007, partial η^2 = .04, and on comprehension-based regulation accuracy, F (2,252) = 3.04, p = .05, partial η^2 = .02. There was no significant main effect of self-assessment instructions on either monitoring-based, F (1,252) = 0.07, p = .80, or comprehension-based, F (1,252) = 0.02, p = .90, regulation accuracy. However, there was a small but significant interaction effect of model answers and self-assessment instructions on monitoring-based regulation accuracy, F (2,252) = 5.05, p = .007,

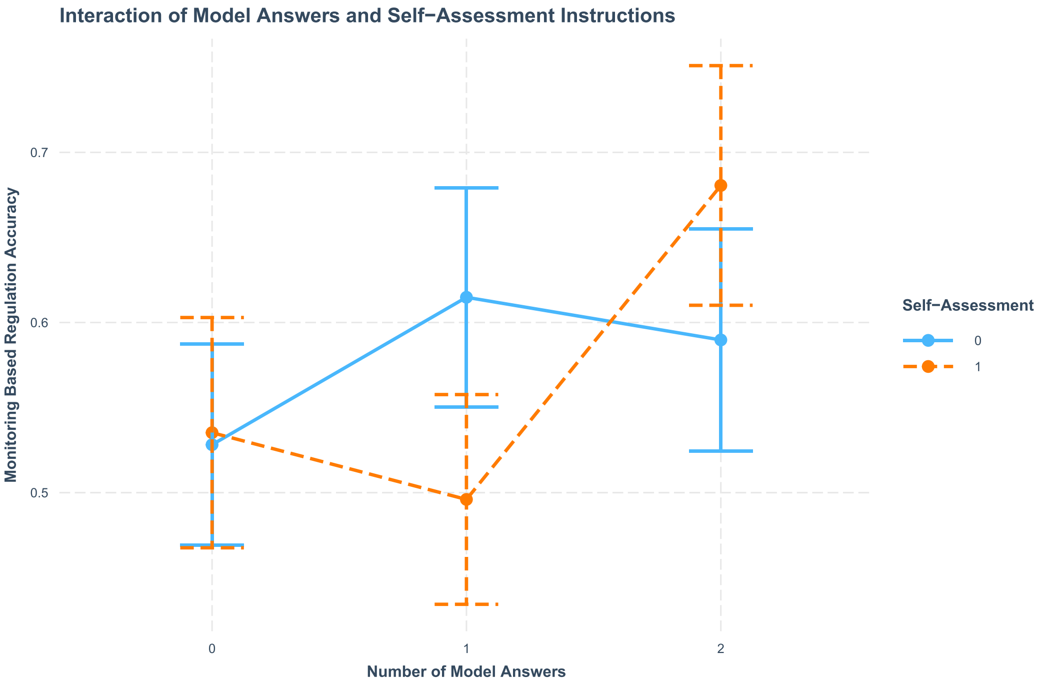


Fig. 2. Plot of model answer and self-assessment instructions on monitoring-based regulation means.

partial $\eta^2 = .04$, but not on comprehension-based regulation accuracy, $F(2,252) = 2.07$, $p = .13$ (H3c).

The interaction effect is visualized in Fig. 2. The figure suggests that receiving model answers positively affected students' monitoring-based regulation accuracy when they were not instructed to compare their answers to the model answer, whereas for students who did receive such instructions, accuracy only improved when they received two model answers. Pairwise comparisons were conducted using the emmeans

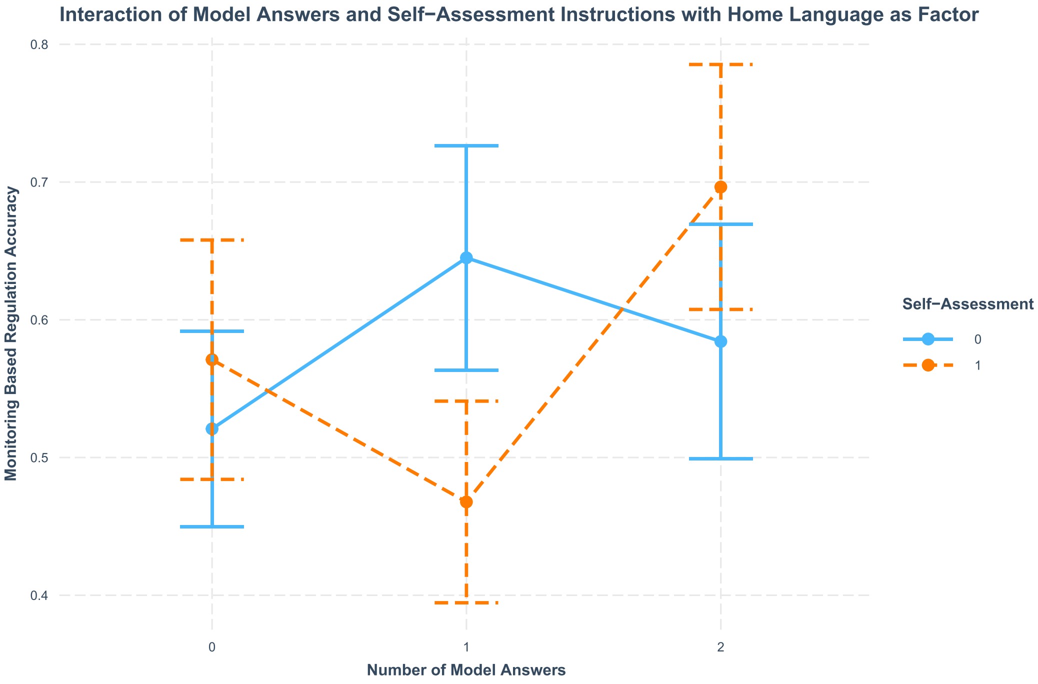


Fig. 3. Plot of model answer and self-assessment instructions on monitoring-based regulation means with home language as a factor.

function in R (Lenth, 2023) to explore the differences in monitoring-based regulation accuracy across the interaction number of model answers and self-assessment instructions (yes or no), using the Kenward-Roger degrees of freedom and the Tukey p-value adjustment method. The greatest difference in monitoring-based regulation accuracy was between those with no model answers and no self-assessment instructions compared to those with two model answers and self-assessment instructions (estimated mean difference = -0.15, t (252) = −3.25, p = .02), while those who had access to self-assessment instructions also did significantly better when they had two model answers instead of one (estimated mean difference = -0.15, t (252) = −2.92, p = .04).

3.5. Impact of home language on comprehension, monitoring accuracy and regulation (RQ2)

The models that added home language not English as a factor showed no main effect of home language (monitoring accuracy: F (1,237) = 0.06, p = .8; comprehension: F (1,237) = 2.95, p = .09, partial η^2 = .01; monitoring-based regulation accuracy: F (1,237) = 0.02, p = .90; comprehension-based regulation accuracy: F (1,237) = 0.19, p = .66), and otherwise yielded a rather similar pattern of results as the models without home language: A main effect of model answers on comprehension test performance at time 2, F (2,237) = 4.37, p = .01, partial η^2 = .04, with a significant difference between 0 and 1 model answers, t (237) = −2.41, p = .04, and a significant difference between 0 and 2 model answers, t (237) = −2.66, p = .02, but not between 1 and 2 model answers, t (237) = −0.32, p = .95. In addition, there was a main effect of model answers on monitoring-based regulation accuracy, F (2,237) = 7.18, p = .001, partial η^2 = .06, and on comprehension-based regulation accuracy, F (2,237) = 3.50, p = .03, partial η^2 = .03. For monitoring-based regulation accuracy there was a significant difference between 0 model answers and 2 model answers, t (237) = −3.59, p = .001, and between 1 and 2 model answers, t(237) = −2.93, p = .01. For comprehension-based regulation accuracy, there was only a significant difference between 1 and 2 model answers, t (237) = −2.38, p = .048. As with the analysis without the home language variable, there was an interaction effect for monitoring-based comprehension accuracy, F (2,237) = 3.63, p = .03 (Fig. 3).

4. Discussion

We investigated whether text comprehension, monitoring, and regulation accuracy could be further improved by providing feedback in the form of one (correct) or two (correct and partially correct) model answers and explicit self-assessment instructions after diagramming the causal relations in a text. We also explored whether the impact of model answers and self-assessment instructions was different for speakers of a home language other than English.

4.1. Do model answers and self-assessment instructions affect text comprehension (H1), monitoring accuracy (H2) and regulation accuracy (H3)?

With regard to text comprehension, the results partially support our hypothesis that model answers would improve comprehension (H1a) and that two model answers would be better than one or none (H1b), especially when self-assessment instructions were provided (H1c). Indeed, we found that students who received two model answers outperformed students who did not receive model answers on the final comprehension test. However, based on prior studies (McCrudden et al., 2007; Braumann et al., 2024a, 2024b) we would have expected that receiving one correct diagram would also be beneficial for text comprehension, as a diagram is thought to make implicit causal relations from the text visually available to learners and students report using diagrams to prepare for comprehension tests (Braumann et al., 2024b). We should note that in those studies, the correct diagram was studied immediately after reading or reading + diagram completion, whereas we applied a delayed design. Possibly, benefits of diagram study for text comprehension only arise when initial reading and studying a visualization of the important relations in the text in a correct diagram are more closely connected in time. Moreover, while an overall beneficial effect of model answers compared to no model answers was found by Braumann et al. (2024a), exploratory post-hoc analyses suggest that this effect was mainly driven by the diagramming + correct diagram condition compared to the no diagramming/no model answer control condition. Similarly, a recent study (which appeared after our study was conducted) showed that immediate diagramming with a correct model answer improved comprehension more than delayed diagramming with a correct model answer (Braumann et al., 2024b). Our findings suggest that delayed diagramming followed by two model answers (a correct and partially correct one) was effective for improving text comprehension. Perhaps it was the opportunity to examine both a correct and partially correct model answer, allowing participants to examine the differences between the diagrams, and therefore get a better understanding of the important features of the correct causal relations, which further improved comprehension over the diagramming intervention alone and provided the expected benefit of model answers. This finding adds to the literature on the effects of feedback on students' performance (e.g., Koenka et al., 2021; Nicol, 2020).

As for monitoring accuracy, the finding from the preliminary analyses that students tend to overestimate their performance is in line with findings from many prior studies (e.g., Griffin et al., 2019; Lipko et al., 2009). With regard to our hypotheses, in contrast to our expectations, we found no significant effects of model answers (H2a&b) alone or combined with self-assessment instructions (H2c). It is hard to explain why receiving one or two model answers would not improve students' monitoring accuracy, especially when they also received explicit instructions to compare their own diagram to the model answer(s), yet this finding is in line with the recent study by Braumann et al. (2024b), who similarly found no significant effect of receiving a model answer in delayed diagramming. One possible reason is that on average, students in the present study were quite accurate, with the preliminary analyses showing that 71% of the monitoring judgments deviated not at all or only by 1 point from their actual performance on the first test (before the model answers were presented). It could also be that the delayed diagramming intervention all participants engaged in gave them access to effective cues regarding their situation model (i.e. strong control condition) and perhaps the model answer intervention was not strong enough to further improve metacomprehension accuracy.

Finally, we expected the hypothesized beneficial effects of model answers and self-assessment instructions on monitoring accuracy and text comprehension to translate into improved regulation accuracy as well. Interestingly, despite the fact that we found no significant effects of our interventions on monitoring accuracy, we did find a significant main effect of model answers on monitoring-based and comprehension-based regulation accuracy, as well as an interaction effect of model answers and self-assessment instructions on monitoring-based regulation accuracy. These findings suggest that students who received model answers used their monitoring judgments when making restudy decisions, and more so when they received two model answers of varying correctness combined with self-assessment instructions. Note, that this measure shows they used their monitoring judgments when making regulation judgments, which is important in self-regulated learning, because when monitoring is accurate, we want learners to rely on these judgments to make regulation decisions. However, since monitoring accuracy did not increase, monitoring-based regulation decisions are not necessarily in line with students' actual needs for restudy. It is also somewhat surprising in light of the lack of significant effect on monitoring accuracy, that the effect of model answers on comprehension-based regulation accuracy shows that students who received model answers made restudy decisions that were in line with their actual test performance. Although many studies on improving self-monitoring and self-regulation only use one measure of regulation accuracy, our results highlight the importance of evaluating the effect of metacomprehension interventions on both aspects of regulation accuracy to unravel the impact of the interventions on students' self-regulated learning (see also Van de Pol et al., 2020).

4.2. Do effects differ for students whose home language is not English?

In our exploration of how home language interacted with the effects of model answers and self-assessment instructions, we expected that model answers would more positively support text comprehension of speakers of other home languages (Buehler et al., 2021). However, we found no effects of home language; thus, it seems that the beneficial effects of model answers and self-assessment instructions on text comprehension and regulation accuracy did not significantly differ for students with and without English as a home language. In line with the findings of Buehler et al. (2021) and the results in the overall sample, the intervention did not result in differences in monitoring accuracy for speakers of home languages other than English (the language of the intervention). Buehler et al. (2021) suggest that making monitoring judgments may be language independent, and that, provided students have adequate language proficiency to understand the task and texts, they can make metacomprehension judgments in any language. Although we did not measure proficiency (only home language), our students' were exposed to highly academic language through their educational program, and their comprehension test performance scores (and the word count data) suggest our students were quite proficient in English. So, our results that learners' ability to make monitoring and regulation judgments does not differ based on home language seem to lend some support to this theory.

4.3. Limitations and future research

One limitation of this study is that the number of participants in our conditions was not perfectly balanced, as a substantial number of participants did not manage to complete the study in time, which resulted in unequal drop-out across conditions. Another limitation is that the number of speakers of home languages other than English was relatively low, so the results of the analyses including home language as a factor should be interpreted with caution. These students were also not equally distributed across conditions. We intentionally asked demographic questions (including home language) at the end of the study to avoid stereotype threat (Steele & Aronson, 1995); however, this meant that we were not able to randomize participants into conditions based on their language background.

Moreover, even though our sample included students from a variety of language backgrounds, all participants were attending English medium of instruction international schools and participating in university-preparatory academic programs of study (e.g., the International Baccalaureate). The participants were therefore exposed to high levels of academic English through their coursework, so perhaps even those who have home languages other than English had the academic language proficiency in English to perform well on the prescribed tasks which were in English; however, we did not collect standardized language proficiency data. Future studies examining the impact of home language on the efficacy of metacomprehension interventions should take language into account when allocating participants to conditions, and also explore to what extent language proficiency plays a role, for instance by using standardized language assessments to more accurately differentiate the impact of monitoring accuracy interventions of language learners at various levels of language proficiency.

Instructing students to self-assess their diagrams did not improve their monitoring accuracy in our study. We cannot rule out that this is due to the prompts students received. In line with Nichol's (2020) argument that self-assessment relies on comparison, in this study participants in the self-assessment and model answers conditions were prompted to compare their own diagram to the model answers. However, it is likely that students also engaged in such comparisons without explicit instructions, and as such, our self-assessment instruction may not have been specific enough to yield significant differences. Future studies could examine the impact of other self-assessment prompts, for example, asking students explicitly to focus on identifying their commission errors.

Another potential limitation is that the study materials may have been slightly too easy for this population. We used the same materials as previous diagramming studies. Although the participants in this study were of a similar age to those in other studies (e.g., van de Pol et al., 2021), their mean performance scores were higher across the board (current study test 1 overall mean: 1.94; van de Pol et al., 2021: range from 1.13 to 1.45) and their responses were often quite thorough (average word count at test 1: 51; average word count of texts: 174). Moreover, as mentioned, 71% of initial judgments were quite accurate (no or only one point deviation between their judgment and actual performance at time one). This raises the question whether the short causal texts used in this intervention were too easy for this specific population. As most studies use these short texts, an interesting question for future studies to address is whether diagramming, receiving model answers as feedback, and/or receiving self-assessment instructions would improve monitoring and regulation accuracy and text comprehension also for longer, more complex texts, that are more similar to the type of academic text upper secondary students encounter in their studies. In addition, our study design required participants to complete two comprehension tests and the first test could also be seen as a generative activity that might have provided participants with some cues regarding their performance. However, they did not get any feedback on the comprehension test and since all participants completed the first test, we have no reason to assume that it would explain (potential) effects of condition. Finally, as our findings on model answers seem to differ somewhat from studies in which diagramming and feedback were engaged with immediately after reading each text rather than after reading several texts (but see Braumann et al., 2024b), future research should compare whether effects of model answers indeed differ with immediate and delayed diagramming, both with these shorter text materials and in the context of longer, more complex texts, and in studies with a similarly strong control.

4.4. Implications for instruction

In the classroom, students read for various purposes and are assessed in a variety of ways on their knowledge. When assessing for comprehension of knowledge gained from text, our results suggest that ensuring students have the opportunity to create a diagram (at a delay after reading) and providing them with correct and partially correct model answers would lead to higher scores, both for students working in their home language as well as those working in another academic language. This may be especially effective when students are reading shorter texts which explain specific phenomena. While our intervention did not improve monitoring accuracy over the effect of delayed diagramming, it also did not reduce monitoring accuracy. Our results suggest that asking students to decide whether they need to restudy texts based on their judgment of their understanding may also be helpful, as in our sample students used their monitoring judgments to make effective regulation decisions.

5. Conclusion

In contrast to our expectations, providing feedback in the form of multiple model answers and explicit self-assessment instructions after delayed diagramming (which improves monitoring accuracy; Van Loon et al., 2014; Van de Pol et al., 2019; 2021) did not further improve monitoring accuracy. We did find that students who received two model answers scored higher on the final comprehension test. Interestingly, students who received two model answers and self-assessment instructions also showed highest monitoring-based regulation accuracy, meaning they used their monitoring judgments when making restudy decisions more strongly than other students.

This study is one of the few studies we know of that examined the impact of such interventions on learners whose home language is not that of the school. Taking home language (English or not) into account as a factor did not change the findings, suggesting that in our sample, home language did not moderate the effects of the interventions; yet due to the low sample sizes when split by language, this result has to be interpreted with caution and needs to be substantiated in future research. The results of this study may also be helpful for teachers; even if a combination of correct and partially correct model answers did not improve *meta* comprehension in our sample, it did not hurt either, and fostered students' comprehension compared to delayed diagramming alone.

CRediT authorship contribution statement

Stephanie L. Hepner: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sophie Oudman:** Writing – review & editing, Writing – original draft. **Trevor E. Carlson:** Software, Data curation. **Janneke van de Pol:** Resources, Methodology, Conceptualization. **Tamara van Gog:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2024.101922.

References

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*(2), 207–245. https://doi.org/10.3102/0034654310368803

Baars, M., Vink, S., van Gog, T., de Bruin, A. B. H., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction, 33*, 92–107. https://doi.org/10.1016/j.learninstruc.2014.04.004

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi:10.18637/jss.v067.i01.

Braumann, S., van Wermeskerken, M. M., van de Pol, J., Pijeira-Díaz, H. J., de Bruin, A. B., & van Gog, T. (2024a). The role of feedback on students' diagramming: Effects on monitoring accuracy and text comprehension. *Contemporary Educational Psychology, 76*, Article 102251. https://doi.org/10.1016/j.cedpsych.2023.102251

Braumann, S., van Wermeskerken, M. M., van de Pol, J., Pijeira-Díaz, H., De Bruin, A. B. H., & van Gog, T. (2024b). Causal diagramming to improve students' monitoring accuracy and text comprehension: Effects of diagram standards and self-scoring instructions. *Applied Cognitive Psychology*. https://doi.org/10.1002/acp.4170

Buehler, F. J., van Loon, M. H., Bayard, N. S., Steiner, M., & Roebers, C. M. (2021). Comparing metacognitive monitoring between native and non-native speaking primary school students. *Metacognition and Learning*, 1–20. https://doi.org/10.1007/s11409-021-09261-z

Bunnell, T., Fertig, M., & James, C. (2016). What is international about International Schools? An institutional legitimacy perspective. *Oxford Review of Education, 42*(4), 408–423. https://doi.org/10.1080/03054985.2016.1195735

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *Quarterly Journal of Experimental Psychology, 64*(3), 467–484. https://doi.org/10.1080/17470218.2010.502239

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*(4), 228–232. https://doi.org/10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Ebert, S. (2015). Longitudinal relations between theory of mind and metacognition and the impact of language. *Journal of Cognition and Development, 16*(4), 559–586. https://doi.org/10.1080/15248372.2014.926272

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review, 28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Follmer, D. J., & Tise, J. (2021). Across-task relations among monitoring judgments: Differential effects of item feedback on monitoring bias during reading. *Learning and Individual Differences, 88*, Article 102007. https://doi.org/10.1016/j.lindif.2021.102007

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2016). Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. *Metacognition Learning, 12*, 1–19. https://doi.org/10.1007/s11409-016-9158-6

Froese, L., & Roelle, J. (2022). Expert example standards but not idea unit standards help learners accurately evaluate the quality of self-generated examples. *Metacognition and Learning, 17*(2), 565–588. https://doi.org/10.1007/s11409-022-09293-z

Griffin, T. D., Mielicki, M. K., & Wiley, J. (2019). Improving students' metacomprehension accuracy. In *The cambridge handbook of cognition and education* (pp. 619–646). New York, NY, US: Cambridge University Press. https://doi.org/10.1017/9781108235631.025.

Griffin, T. D., Wiley, J., & Salas, C. R. (2013). Supporting effective self-regulated learning: The critical role of monitoring. In R. Azevedo, & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 19–34). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-5546-3.

Grundy, J. G., & Timmer, K. (2017). Bilingualism and working memory capacity: A comprehensive meta-analysis. *Second Language Research, 33*(3), 325–340. https://doi.org/10.1177/0267658316678286

Kim, Y.-S. G. (2017). Why the simple view of reading is not simplistic: Unpacking component skills of reading using a Direct and Indirect Effect Model of Reading (DIER). *Scientific Studies of Reading, 21*(4), 310–333. https://doi.org/10.1080/10888438.2017.1291643

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY, US: Cambridge University Press.

Kintsch, W. (2005). An overview of top-down and bottom-up effects in comprehension: The CI perspective. *Discourse Processes, 39*(2–3), 125–128. https://doi.org/10.1080/0163853X.2005.9651676

Koenka, A. C., Linnenbrink-Garcia, L., Moshontz, H., Atkinson, K. M., Sanchez, C. E., & Cooper, H. (2021). A meta-analysis on the impact of grades and comments on academic motivation and achievement: A case for written feedback. *Educational Psychology, 41*(7), 922–947. https://doi.org/10.1080/01443410.2019.1659939

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An r package for assessment, comparison and testing of statistical models. *Journal of Open Source Software, 6*(60), 3139. https://doi.org/10.21105/joss.03139

Lenth, R. (2023). Emmeans: Estimated marginal means, aka least-squares means. R package version 1.8.5 https://CRAN.R-project.org/package=emmeans.

Lin-Siegler, X., Shaenfield, D., & Elder, A. D. (2015). Contrasting case instruction can improve self-assessment of writing. *Educational Technology Research & Development, 63*, 517–537. https://doi.org/10.1007/s11423-015-9390-9

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied, 15*(4), 307–318. https://doi.org/10.1037/a0017599

Maki, R. H. (1998). Test predictions over text material. In *Metacognition in educational theory and practice* (pp. 131–158). Routledge.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(4), 663–679. https://doi.org/10.1037/0278-7393.10.4.663

McCrudden, M. T., Schraw, G., Lehman, S., & Poliquin, A. (2007). The effect of causal diagrams on text learning. *Contemporary Educational Psychology, 32*(3), 367–388. https://doi.org/10.1016/j.cedpsych.2005.11.002

Nicol, D. (2020). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education, 46*(5), 756–778. https://doi.org/10.1080/02602938.2020.1823314

OECD. (2012). *Untapped skills: Realising the potential of immigrant students*. Paris: OECD Publishing. https://doi.org/10.1787/9789264172470-en

Pollock, D. C., Van Reken, R. E., & Pollock, M. V. (2017). *Third culture kids: The experience of growing up among worlds* (Third edition). Nicholas Brealey Publishing.

Prinz, A., Golke, S., & Wittwer, J. (2020b). To what extent do situation-model-approach interventions improve relative metacomprehension accuracy? Meta-Analytic insights. *Educational Psychology Review, 32*(4), 917–949. https://doi.org/10.1007/s10648-020-09558-6

Prinz, A., Golke, S., & Wittwer, J. (2020a). How accurately can learners discriminate their comprehension of texts? A comprehensive meta-analysis on relative metacomprehension accuracy and influencing factors. *Educational Research Review, 31*, Article 100358. https://doi.org/10.1016/j.edurev.2020.100358

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19* (4–5), 559–579. https://doi.org/10.1080/09541440701326022

Rawson, K. A., O'Neil, R., & Dunlosky, J. (2011). Accurate monitoring leads to effective control and greater learning of patient education materials. *Journal of Experimental Psychology: Applied, 17*(3), 288–302. https://doi.org/10.1037/a0024749

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69* (5), 797–811. https://doi.org/10.1037/002-3514.69.5.797

Tan, E. C., Wang, K. T., & Cottrell, A. B. (2021). A systematic review of third culture kids empirical research. *International Journal of Intercultural Relations, 82*, 81–98. https://doi.org/10.1016/j.ijintrel.2021.03.002

Thiede, K. W., & Anderson, M. C. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*(2), 129–160. https://doi.org/10.1016/So361-476X(02)00011-5

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 1024. https://doi.org/10.1037/0278-7393.25.4.1024

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*(4), 331–362. https://doi.org/10.1080/01638530902959927

Thiede, K. W., Wright, K. L., Hagenah, S., & Wenner, J. (2019). Drawings as diagnostic cues for metacomprehension judgment. In N. Feza (Ed.), *Metacognition in learning. IntechOpen.* https://doi.org/10.5772/intechopen.86959

Van de Pol, J., de Bruin, A. B. H., van Loon, M. H., & van Gog, T. (2019). Students' and teachers' monitoring and regulation of students' text comprehension: Effects of comprehension cue availability. *Contemporary Educational Psychology, 56*, 236–249. https://doi.org/10.1016/j.cedpsych.2019.02.001

van de Pol, J., van den Boom-Muilenburg, S. N., & van Gog, T. (2021). Exploring the relations between teachers' cue-utilization, monitoring and regulation of students' text learning. *Metacognition and Learning, 16*, 769–799. https://doi.org/10.1007/s11409-021-09268-6

van de Pol, J., van Loon, M. H., van Gog, T., Braumann, S., & de Bruin, A. B. H. (2020). Mapping and drawing to improve students' and teachers' monitoring and regulation of students' learning from text: Current findings and future directions. *Educational Psychology Review, 32*(4), 951–977. https://doi.org/10.1007/s10648-020-09560-y

van Loon, M. H., de Bruin, A. B. H., van Gog, T., van Merriënboer, J. J. G., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143–154. https://doi.org/10.1016/j.actpsy.2014.06.007

Van Loon, M. H., & Roebers, C. M. (2017). Effects of feedback on self-evaluations and self-regulation in elementary school. *Applied Cognitive Psychology, 31*(5), 508–519. https://doi.org/10.1002/acp.3347

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *The Journal of General Psychology, 132*(4), 408–428. https://doi.org/10.3200/GENP.132.4.408-428

Yang, C., Zhao, W., Yuan, B., Luo, L., & Shanks, D. R. (2023). Mind the gap between comprehension and metacomprehension: Meta-analysis of metacomprehension accuracy and intervention effectiveness. *Review of Educational Research, 93*(2), 143–194. https://doi.org/10.3102/00346543221094083

2024 ISC Research. (2024). What data tells us about the international schools market [White Paper]. go.iscresearch.com/data_whitepaper_2024.

Stephanie L. Hepner is a PhD candidate, Dr. Sophie Oudman is assistant professor, Dr. Janneke van de Pol is associate professor, and Prof. Dr. Tamara van Gog is professor of Educational Sciences at Utrecht University, The Netherlands. Their research focuses on improving monitoring and regulation accuracy of students. Dr. Trevor E. Carlson is assistant professor of Computer Science at National University of Singapore. His research focuses on computer architecture, and computer architecture education.