RESEARCH



The implementation of peer assessment as a scaffold during computer-supported collaborative inquiry learning in secondary STEM education



Amber Van Hoe^{1*}, Joel Wiebe², Tijs Rotsaert¹ and Tammy Schellens¹

Abstract

Background Computer-supported collaborative inquiry learning (CSCiL) has been proposed as a successful learning method to foster scientific literacy. This research aims to bridge the knowledge gap surrounding the role of peers as scaffolding sources in CSCiL environments. The primary objective is to explicitly implement peer assessment as a scaffolding tool to enhance students' inquiry output in terms of research question, data, and conclusion. Additionally, students' perceptions of peer assessment within CSCiL are explored.

Results The study involved 9th and 10th-grade students from 12 schools (N= 382), exploring the effects of peer assessment with and without peer dialogue. The results highlight that while adjustments were more frequently made to the research question and data, adjustments to the conclusion showed significantly greater improvement. Furthermore, students' perceptions of peer assessment during CSCiL were examined, revealing that students generally perceive peer assessment as fair and useful, and they accept it while being willing to make improvements based on the feedback. While students did not report experiencing negative feelings, they also did not report positive emotions from the process. Additionally, the study found that including a peer dialogue in the peer assessment process did not significantly impact the abovementioned findings.

Conclusions This study enriches our understanding of peer assessment as a scaffolding tool in CSCiL, highlighting its potential to improve inquiry outputs and providing valuable insights for instructional design and implementation.

Keywords Collaborative learning, Inquiry-based learning, Technology-based learning, Scaffolding, Peer assessment

Introduction

Computer-supported collaborative inquiry learning is being advocated as an effective approach to promote scientific literacy among students (Barron & Darling-Hammond, 2010; de Jong, 2019). This collaborative learning strategy perfectly aligns with a social constructivist vision of learning that emphasizes active knowledge construction through interaction as students work actively together during group learning activities in a computersupported learning environment (Chen et al., 2018).

While the addition of active collaboration and technological tools positively impacts students' learning (Chen et al., 2018; de Jong, 2019), it is important to acknowledge that computer-supported collaborative inquiry learning can be demanding for students. As confirmed by multiple meta-analyses, providing support during the inquiry process is crucial for improving learning outcomes (Furtak et al., 2012; Lazonder & Harmsen, 2016).



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

^{*}Correspondence:

Amber Van Hoe

Amber.VanHoe@UGent.be

¹ Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium

² University of Toronto, 252 Bloor St W, Toronto, ON M5S 1V6, Canada

In response, integrating formative assessment as a scaffold within inquiry-based learning has been recommended (Linn et al., 2018; Mckeown et al., 2017; Xenofontos et al., 2019). In line with the current trend of collaborative learning, peers are increasingly considered a valuable source of formative assessment for one another. Peer assessment within various educational contexts has shown positive effects, such as improved academic performance (Double et al., 2020) and critical thinking (Jiang et al., 2022). Nevertheless, the application of peer assessment within computer-supported inquiry learning remains limited.

To address this research gap, this study focuses on implementing peer assessment as a scaffolding tool within a computer-supported inquiry learning environment designed for secondary students to investigate climate change. Specifically, this study aims to determine whether the quality of students' inquiry output improves through peer assessment while comparing two formats: one with a single peer assessment activity and another with peer assessment supplemented with peer dialogue.

Inquiry-based learning

Inquiry-based learning (IL) is an active pedagogical approach primarily used within school subjects focusing on sciences, technology, engineering, and mathematics, commonly referred to as STEM subjects. This approach connects science education to the outside world by introducing relevant and authentic scientific inquiry topics to students, encouraging them to construct knowledge using procedures and practices comparable to those used by professional scientists (Capps et al., 2012; Chu et al., 2017). During IL, student groups typically go through an inquiry cycle that divides the learning process into smaller parts. In a comprehensive literature review, Pedaste and colleagues (2015) identified five general inquiry stages, which can be found in Fig. 1. Each of the stages results in a well-defined output. The learning topic is introduced in the orientation phase, resulting in a problem statement that challenges and motivates students. Next, the conceptualization phase aims at comprehending central concepts related to the problem, leading to research questions and hypotheses. Observations or experiments are conducted in the following investigation phase, and the obtained data are interpreted. Consecutive, conclusions ought to be drawn in the conclusion phase. Lastly, after each inquiry phase, a discussion phase can be posed wherein findings are communicated to others (e.g., peers or teachers), and feedback can be collected. Although the previous order of inquiry phases appears to be the most likely, transitions between them are allowed throughout the inquiry (Pedaste et al., 2015).



Fig. 1 An adapted version of the inquiry cycle of Pedaste et al. (2015)

Furthermore, students' inquiry process is also structured and facilitated through technology (Matuk et al., 2019). Web applications and online learning environments are designed explicitly for inquiry and offer various advantages, including visualizing complex theoretical concepts and the ability to 'play' with scientific phenomena via simulations (de Jong, 2019). Logically, educational technology has since become the standard within IL. Consequently, IL was integrated into the broader research field of computer-supported collaborative learning (CSCL), leading to the emergence of computer-supported collaborative inquiry learning (CSCiL).

The research findings of CSCiL and IL, in general, are encouraging. First, it has been found that IL contributes to developing students' scientific conceptual knowledge (Furtak et al., 2012; Heindl, 2019). Second, IL has proven to enhance scientific inquiry skills (Mäeots, 2008; Raes et al., 2014; Sharples et al., 2015). This makes sense considering that students are doing science during IL. As students go through an inquiry cycle, they learn to identify problems, formulate research questions and hypotheses, plan and set up experiments, collect and analyse data, present results, draw conclusions, and communicate them (Constantinou et al., 2018). These skills are reflected in the research cycle of Pedaste et al. (2015). Third, IL also has been shown to stimulate affective learning outcomes. For example, according to research by Raes et al. (2014), CSCiL increases students' interest in science and can even bridge the interest gap between boys and girls, as girls typically show less interest in the subject. This is important since it was found that students who demonstrate a greater interest in scientific skills are more inclined to think about a STEM career (Blotnicky et al., 2018). In addition, Husnaini and Chen (2019) and Ketelhut (2007) discovered that IL also has a favorable impact on students' scientific inquiry self-efficacy. This means that IL stimulates students' beliefs about their ability to perform the competencies needed to do scientific inquiry. It is critical to promote scientific inquiry selfefficacy for two reasons. First, according to Blotnicky and colleagues' (2018) research, students with higher self-efficacy in mathematics were more aware of the demands of a STEM career and more inclined to pursue one. Second, it has been discovered that self-efficacy is a powerful predictor of academic success (Caprara et al., 2011). Lastly, IL also has been shown to nurture transferable and sustainable skills like communication, collaboration, and creativity (Barron & Darling-Hammond, 2010; Chu et al., 2017).

The importance of scaffolding during computer-supported collaborative inquiry learning (CSCiL)

Despite these positive learning outcomes, the fast rise of IL in research and school STEM curricula is not without controversy. IL is criticized the most by direct teaching proponents who claim that they are minimally or even unguided and cause cognitive overload in students, which hinders learning (e.g., Kirschner et al., 2006). However, social constructivist teaching methods, such as IL, fully embrace the need for guidance during the learning process (Hmelo-Silver et al., 2007). To do this, they refer to the concept of scaffolding, which originates in Vygotsky's sociocultural theory. This theory states that learning happens through interaction with adults or peers who are more informed (Shabani, 2016). Scaffolding itself refers to the customized support that helps learners perform tasks outside their independent reach and consequently develop the skills necessary for completing such tasks independently (Wood et al., 1976). It can be done in various ways, for example, by modelling and questioning (van de Pol et al., 2010). Reasonably, it is already wellestablished that scaffolding during IL is a prerequisite to attain the aforementioned learning outcomes (e.g., Alfieri et al., 2011; Lazonder & Harmsen, 2016). In computersupported learning environments, the need for scaffolding is often even more significant due to factors such as the complexity of the learning process (Pedaste et al., 2015) and the need for regulation (Dobber et al., 2017) and motivation (Raes & Schellens, 2016). Scaffolding provides guidance and support to learners in these environments. More precisely, during CSCiL, three potential scaffolding sources are available for learners: the teacher, technology, and peers (Kim & Hannafin, 2011).

Research has shown that *teachers* are an essential scaffolding resource during inquiry (e.g., Furtak et al., 2012; Matuk et al., 2015; Tissenbaum & Slotta, 2019). For instance, Raes and Schellens (2016) discovered that teacher interventions that provide structure and feedback are favourable as they lower students' frustration levels. Moreover, Dobber and colleagues (2017) found that teachers functionally support students to regulate their learning process throughout IL. They are essential for social (i.e., managing social processes, e.g., structuring

student collaboration), meta-cognitive (i.e., fostering an inquiry mindset, e.g., developing a culture of inquiry), and conceptual regulation (i.e., subject knowledge and rules, e.g., focusing on conceptual understanding). Next to that, a recent study by Pietarinen et al. (2021) observed that teachers spend much time during CSCiL assisting student groups with technology.

The second potential scaffolding resource during CSCiL, technology, could reduce teachers' workload (Dillenbourg, 2009) as online inquiry environments possess the ability to build in technology-enhanced scaffolding mechanisms (Matuk et al., 2019). Belland and colleagues' (2017) meta-analysis showed that computer-based scaffolding has a moderately positive effect on cognitive outcomes. This beneficial result persists irrespectively of the scaffolds' design (e.g., general or context-specific scaffolds). Additionally, Kim et al. (2020) found that computer-based scaffolding has the most significant effect on pairs of students compared to computer-based scaffolding for individual students or larger student groups. There is ongoing development to transform online inquiry environments into truly adaptive systems that provide students with timely and personalized guidance (de Jong, 2019).

Lastly, CSCiL is predicated on the premise that peers as a scaffolding resource for one another because of the collaboration throughout learning activities. In addition to a substantial, moderate effect of collaboration on skill acquisition (e.g., critical thinking and problem-solving), Chen et al. (2018) also revealed a significant minor effect on knowledge acquisition and student perceptions. However, prior research has generally focused on individual or within-group learning during CSCL. How different student groups can work together (i.e., between-group collaboration) during CSCL is significantly understudied (Chen & Tan, 2021). Therefore, this study will investigate if different student groups, via between-group collaboration, can form a valuable scaffolding resource for one another during CSCiL. More specifically, peer assessment will be used to operationalize this between-group collaboration since already been shown that student groups require formative feedback during CSCiL (Barron & Darling-Hammond, 2010; Mckeown et al., 2017).

Peer assessment as underexplored scaffolding mechanism Although peer assessment as a formative assessment practice has already gained significant acceptance in educational research (Double et al., 2020), up to this point, it has not yet been widely investigated as a reliable scaffolding method within CSCiL (e.g., Tsivitanidou et al., 2011). Only a few research studies have implemented peer assessment within CSCiL in secondary STEM education. These studies' main focus was determining which

peer assessment format results in the most favorable outcomes.

For example, both Tsivitanidou et al. (2011) and Dmoshinskaia et al. (2020) investigated the effect of whether or not to provide assessment criteria to students when giving feedback on each other's inquiry products. Tsivitanidou et al. (2011) discovered that when students were asked to assess their peers but did not receive any instructions and assessment criteria to do this, they independently came up with the idea that they needed to formulate assessment criteria and provide suggestions to improve peers' inquiry products. However, the quality of these assessment criteria was poor. Dmoshinskaia et al. (2020) found that the quality of the provided peer feedback, finished inquiry products, and post-test knowledge acquisition did not significantly differ between students who received assessment criteria and those who did not.

Other researchers focused on the differences between quantitative (i.e., grading) and qualitative (i.e., commenting) peer feedback during CSCiL within secondary STEM courses. Dmoshinskaia et al. (2021) compared a group of students who gave quantitative peer feedback with a group who gave qualitative peer feedback. These two groups were the same in the number of adjustments made based on peer feedback. However, students of the qualitative feedback group scored significantly higher on the post-test that measured domain knowledge. Another study by Hovardas and colleagues (2014) compared quantitative and qualitative peer feedback with expert feedback. There were few similarities between the quantitative feedback of peers and experts. The structure of qualitative feedback of both parties did significantly overlap. While the qualitative peer feedback contained mostly scientifically accurate domain knowledge, critical evaluation was lacking, which led to an absence of improvement suggestions and mainly approval of inquiry products. When critical remarks did occur, they were not supported by sufficient arguments.

A common finding of Tsivitanidou et al. (2011) and Hovardas et al. (2014) is that students mostly ignore peer and expert feedback. What the previous research has in common is that only grades and comments are exchanged. As such, there is solely a medium interactivity between assessors and assessees as they do not have the opportunity to have an open dialogue (Deiglmayr, 2018). While peer assessment stems from a participatory learning culture (Kollar & Fischer, 2010), it is shaped mainly in a traditional one-way information flow, in which assessees only play a passive, receptive role. Tsivitanidou et al. (2012) previously tried to break through this passive role by having students actively ask for peer feedback. Previous findings, namely the absence of critical feedback and the ignorance of feedback as time progressed, were reciprocated in this study of Tsivitanidou and colleagues (2012). Although most of the students' questions for help were answered, the degree of interactivity between the assessor and assessee was not elevated to a higher level during this study. A more advanced understanding of peer assessment includes interaction, or even a sustained dialogue, establishing a partnership between the assessor and assessee (Winstone & Carless, 2020). Carless (2015) defines dialogic feedback as "iterative processes in which interpretations are shared, meanings negotiated, and expectations clarified to promote student uptake of feedback" (p. 196). Implementing a peer dialogue in the peer assessment process could thus respond to the pitfalls of peer assessment during CSCiL mentioned above. Such dialogue activates the assessee as they have to participate in a dialogue that is expected to result in a higher peer feedback uptake (Carless, 2016; Voet et al., 2018). Moreover, this dialogue could give both parties more room to explain the work produced and the feedback provided, and as a result, peer feedback may be taken, for example, more fairly. A deeper investigation of peer assessment perspectives during CSCiL is thus required.

The present study

By implementing peer assessment during a CSCiL lesson series, the contribution of this research is threefold. First, it answers the recent appeal to integrate formative assessment into inquiry-based STEM learning (Linn et al., 2018; Mckeown et al., 2017; Xenofontos et al., 2019). Contrary to the aforementioned studies, the main objective of this study is to do so by approaching peer assessment as a potential scaffolding tool during CSCiL that could take students' inquiry to a higher level. Second, this study contributes to the search for the best peer assessment format by investigating the potential benefits of peer dialogue, simultaneously advocating students' active role during the inquiry and feedback process (Carless, 2016). Third, since little is known regarding students' perspectives of peer assessment within CSCiL, this study aims to fill this knowledge gap.

In this study, peer assessment is defined as an interpersonal, collaborative learning arrangement in which student groups assess fellow student groups' inquiry output by providing feedback (i.e., between-group collaboration). Inquiry output refers to student groups' work through the conceptualization, investigation, and conclusion phases (i.e., resulting in a research question, data, or conclusion).

This leads to the following research questions (RQ):

1. What is the effect of the addition of a peer feedback dialogue on the number of adjustments that stu-

dents make to their inquiry output in terms of (a) a research question, (b) data, and (c) a conclusion?

- 2. What is the effect of peer assessment with or without peer feedback dialogue on the quality of students' generated inquiry output in terms of (a) a research question, (b) data, and (c) a conclusion?
- 3. What is the effect of peer assessment with or without peer feedback dialogue on students' perceptions of peer assessment?

To answer these RQs, a quasi-experimental study is established wherein two forms of formative peer assessment, namely peer feedback with or without peer dialogue, are compared. The peer feedback provided in this study comprises quantitative (i.e., grades or ratings across assessment criteria) and qualitative (i.e., written comments) components. In addition, a control condition was created to check the effectiveness of CSCiL.

Methods

Context

For this research, a lesson series about climate change was designed for the ninth and tenth grade, called Climate colLab. Climate colLab is developed in the web-based Scripting and Orchestration Environment (SCORE) which results from close collaboration among researchers and software developers at the University of Toronto and the University of Berkeley. It takes up four lesson periods of 50 min in total (200 min in total). Students engage in the lessons in randomly composed groups of two or three students sharing a single computer.

54 Master's students in Educational Sciences opted to support the implementation of Climate colLab, which was strictly protocolled. The graduate students were the actual teachers during the project, while the regular class teachers functioned as observers. This project was a required component of Ghent University's 7-credit course in Educational Technology for these Master's students. Every Master's student received rigorous training in advance. They knew the theoretical underpinnings of CSCiL and were acquainted with SCORE. Master's received a protocol about how they can scaffold student groups per exercise. The actual class teachers who served as observers were asked to complete a questionnaire about the Master students' scaffolding behaviour during the lessons. Based on this assessment by the teacher, data were included or excluded from the dataset.

Study participants

A quasi-experimental study with a pre-post test design was set up to answer the research questions. This study includes 28 ninth and tenth grade classes from 12 schools (N=506; Mage=15.11, SDage=0.69). These classes were randomly split into one control and two experimental conditions (i.e., peer assessment with or without peer dialogue). While all classes took a pre- and post-test, only the classes in the experimental conditions took Climate colLab. Only data from the students who took the pre- and post-test in each of the three conditions and who were present for each of the four class periods in the experimental conditions were used. Total data from 382 students (Mage=15.04, SDage=0.67) were deemed valid. This resulted in the following distribution: control (n=69), peer assessment (PA; n=160), and PA+Dialogue (n = 153). To answer RQ1 and RQ2, data from the student groups in the SCORE system were used. This dataset consists of 187 student groups that were divided between two conditions: PA (n=93) and PA + Dialogue (n = 94).

The Ethics Committee of the Faculty of Psychology and Educational Sciences of Ghent University approved the research. Before the start of the study, informed consent forms were distributed online to all participants and their parents, as well as the responsible teachers and school principals. In these informed consent forms, the design of the research, as well as the collection and processing of data, was outlined. Informed consent was obtained from all the involved parties to use the data for this study.

Procedure

Design of Climate colLab

The specific lesson topics of Climate colLab were determined by the curriculum standards that students must meet by the end of the second grade of secondary education in Flanders, Belgium. A teacher and subject expert examined the lessons and gave feedback to ensure that the content was accurate and the difficulty level was suitable for students of the ninth and tenth grade. A pilot trial of the lessons was held with 24 ninth graders. Based on the results of this pilot research, specific exercises were eliminated (e.g., too challenging, unclear, or redundant), certain concepts were rephrased, and so on.

The lesson content of Climate colLab is structured according to the inquiry cycle of Pedaste et al. (2015), mentioned in Fig. 2. There are four inquiry cycles in total. During the *orientation phase*, student groups are introduced to the concepts necessary to master the research questions presented in the following conceptualization phase. This is accomplished through various activities and text resources (e.g., simulations, informative



Fig. 2 Structure of the inquiry cycles during Climate colLab based on Pedaste et al. (2015)

images, multiple-choice or drag-and-drop questions, external links, and newspaper articles...). In the conceptualization phase, student groups were introduced to the research question. In the first inquiry cycle, groups must investigate how the sun provides heat to Earth. In the following two inquiry cycles, students investigate the albedo and greenhouse effect. Student groups are asked to jointly develop a hypothesis before moving on to the investigation phase. During this investigation phase, various resources (e.g., simulations, internet links to external scientific sources, interactive exercises, ...) are provided to students to gather research data in order to form an answer to the research question in the *conclusion phase*. Since students work in groups, the discussion phase is continuously realized through within-group discussions. These first three inquiry cycles aim to give students a firm knowledge base about climate change and take up the two first lesson periods. To get acquainted with the inquiry cycle, an instructional video was shown in the classroom at the start of the first lesson, explaining the inquiry cycle through an exemplary research subject.

Whereas student groups are provided with research questions, simulations, and exercises throughout the abovementioned three inquiry cycles, they are expected to set up their own research on sustainability during the fourth and last inquiry cycle during the last two lesson periods. The three sustainability principles, People Planet Profit (3Ps), are introduced during the orientation phase. Subsequently, student groups need to choose the theme of their sustainability research out of five proposed themes (i.e., energy, nutrition, transport/travel, climate refugees, and fast fashion). News article titles are provided to inspire students about possible sustainability topics within these themes. Afterward, student groups proceed to the conceptualization phase, wherein they must formulate a research question about sustainability in their chosen theme. Students are given five hints about how to draft a good research question (e.g., 'Formulate your research question clearly and concretely'), and six types of research questions are illustrated (e.g., evaluative). After that, student groups proceed to the investigation phase, wherein they select a maximum of 5 internet sources that could contribute to their research. Again, four hints are given to support the students in their search for information (e.g., 'Who created the source?'). Web links and useful information need to be pasted into SCORE. When student groups think they have collected enough data, they proceed to the conclusion phase, wherein they must formulate a comprehensive conclusion to their posed research question. They are reminded to incorporate information about each of the 3Ps into their conclusion.

Quasi-experimental research design

A quasi-experimental study with a pretest–posttest design was set up to answer the RQs. Figure 3 shows three conditions, which are one control condition and two experimental conditions.

The control condition and both experimental conditions took a pre- and post-test. The time interval between the tests for the control and experimental conditions was the same, namely two weeks.



Fig. 3 Quasi-experimental research design

During this 2-week interval, the educational content was taught to the students. For those in the control group, this encompassed receiving instruction on key topics from the Climate colLab project (i.e., the mechanisms by which the sun provides warmth to Earth, the albedo effect, and the greenhouse effect). Their subject teachers taught these topics through their regular teaching approach.

For students in the two experimental conditions, this meant receiving the lesson content via Climate colLab. Whereas the two experimental conditions (i.e., PA and PA+Dialogue) are similar during the first three inquiry cycles, they differ from one another during the fourth inquiry cycle (see Fig. 3). This fourth inquiry cycle is also structured differently than the first three, as it addresses the discussion phase in two different ways. First, similarly to the first three inquiry cycles (see Fig. 2), withingroup discussion occurs during each inquiry phase as students work together in a group. Second, given that student groups share their results with another group during peer assessment moments, a between-group discussion is added in the fourth inquiry cycle (see Fig. 4). This between-group discussion, operationalized as a peer assessment episode (see Fig. 5), occurs at three fixed points in the fourth inquiry cycle, namely after the conceptualization, investigation, and conclusion phases. This between-group discussion focuses on the inquiry output of these phases: the research question, data, or conclusion.

The difference between the two experimental conditions in the fourth inquiry cycle lies in operationalizing the between-group discussion. While quantitative and qualitative peer feedback is given in both experimental



Fig. 4 Structure of the fourth inquiry cycle during Climate colLab based on Pedaste et al. (2015)



Fig. 5 Peer assessment design according to the experimental condition

conditions, solemnly in the PA+Dialogue condition, peer feedback is accompanied by a peer dialogue (see Fig. 5). The following section elaborates on the operationalization of the peer feedback process.

Peer assessment design

The peer assessment during Climate colLab is reciprocal, meaning that a student group gives peer feedback to another student group and receives peer feedback from that group. These so-called peer review groups are randomly composed. As mentioned before and shown in Fig. 5, both experimental conditions receive quantitative (i.e., grades or ratings across assessment criteria) and qualitative feedback (i.e., written comments). This is only accompanied by a peer feedback dialogue in the PA + Dialogue condition.

The quantitative peer feedback is operationalized via rubrics with preformulated criteria, as it has been found that self-made up criteria by learners are of low quality (Tsivitanidou et al., 2011). The rubrics are defined on a 5-point Likert scale ranging from very good (5) over sufficient (3) to insufficient (1). Three rubrics were developed as the three inquiry phases each result in different inquiry outputs (i.e., a research question, data, or conclusion) (see Appendix A). These rubrics were developed in consultation with experts and then pilot tested with a tenth-grade class. Based on this pilot test, the rubrics were finetuned (e.g., the phrasing was simplified). The first inquiry output is the research question assessed on three assessment criteria: scope of the research question, importance of sustainable living, and language use. Next, the three assessment criteria for the research data are the choice of information resources, the occurrence of the 3Ps, and the accuracy of the selected information. Finally, the research conclusion is tested against two assessment criteria, the occurrence of the 3Ps and language use.

Each time, quantitative feedback is accompanied by qualitative feedback. Student groups are asked to clarify the scores (i.e., 'What works well and why?; What might be improved and why?') and offer suggestions for improvement.'). Through these question prompts, students are urged to provide constructive criticism and suggestions for improvement (Gan & Hattie, 2014) rather than just confirming one another's work as was found to happen in earlier studies (Tsivitanidou et al., 2011, 2012).

In the PA+Dialogue condition, peer review groups engage in dialogue immediately after shortly reviewing their received peer feedback. Six question starters are provided to stimulate the peer dialogue (e.g., 'We do (not) agree with the feedback about... because...').

After reviewing the received peer feedback in the PA condition or when the peer dialogue is wrapped up in the PA + Dialogue condition, student groups in both conditions have the opportunity to revise their inquiry output (i.e., research question, data, or conclusion dependent on inquiry phase) before moving on to the following inquiry phase where this process is repeated.

An instructional video was shown at the start of the last inquiry cycle to get acquainted with the peer assessment procedure. This video explains the peer assessment procedure and trains the participants to use the assessment rubrics through a worked example. In the PA + Dialogue condition, this video was expanded by explaining and demonstrating a peer dialogue.

Peer assessment is conducted via an embedded tool in SCORE. This tool automatically exchanges the student work that needs to be assessed and the peer feedback provided. Likewise, the peer dialogue was facilitated within SCORE via an embedded chat tool. The peer assessment was not anonymized as SCORE showed to whom feedback needed to be given, from whom feedback was received, and with whom they were chatting within the chat tool. This was done to minimize spam messages and maximize the collaboration quality (Velamazán et al., 2023).

Measures and data analysis *Pre- and post-test*

To determine whether students gained scientific knowledge about climate change by following the lesson series, the first section of the pre- and post-test included five questions that tested their scientific knowledge of climate change. The first three questions entailed a combination of a multiple-choice and an open-ended question component in which they needed to complete the correct answer and receive space for explaining the scientific idea behind their chosen answer. These questions were scored on a total from 0 to 4. The multiple-choice component was scored 0 (false) or 1 (correct), and the open-choice component was scored on a rubric from 0 to 3. Additionally, the fourth question was an open-ended knowledge question scored on the same rubric from 0 to 3. More precisely, this rubric is a modified version of the knowledge integration rubric (Liu et al., 2008). It contains several competence levels, whereas higher proficiency levels correspond to more sophisticated abilities to solve scientific problems. The fifth and last question of the test was a closed question in which participants could earn a minimum of 0 points and a maximum of 6 points. The knowledge test was scored on a total of 21 points which was converted to a total of 20 points. An example of these test items can be found in Appendix B, accompanied by a rubric example in Appendix C. All questions were assessed by two independent raters, the first author and an independent rater trained to use the rubric. The independent rater coded 30 percent of the answers to check the inter-rater reliability. Regarding all items, Cohen's Kappa ranged from 0.63 to 0.87 (see Table 1), which indicates substantial (0.61-0.80) to almost perfect (0.81-1.00) inter-rater agreement (Cohen, 1960).

To capture the expectations of students about peer feedback as well as their perceptions of the peer feedback they had received during the intervention, the questionnaire of Strijbos and colleagues (2010) was included in the pre- and post-test. Statements were measured on a bi-polar scale ranging from 0 (=fully disagree) to 10 (=strongly agree). Table 2 summarizes the six scales, each with a sample item and Cronbach's alphas. Each scale could be formatted as reliability analysis generates acceptable to good Cronbach's alphas.

Quality of students' inquiry output

SCORE saved the version of the inquiry output before the peer assessment and the possibly reworked version after receiving the peer assessment. To capture the difference between the two conditions on the quality of students' generated inquiry output in terms of (a) a research question, (b) data, and (c) a conclusion (RQ2), it was first required to record if the student groups made any adjustments after the peer assessment episode (RQ1). A dichotomous variable was developed: one denotes making

 Table 1
 Overview of Cohen's Kappas for each question in the pre- and post-test

	Pretest	Post-test
Question 1	0.63	0.72
Question 2	0.87	0.78
Question 3	0.72	0.70
Question 4	0.73	0.73
Question 5	1.00	1.00

Scale	Example item	Items	α ^a
Fairness	I expect that I will be satisfied with the feedback that I will receive from my classmates	3	0.70
Usefulness	I expect that the feedback from my classmates will be useful	3	0.88
Acceptance	I expect that I will accept the feedback from my classmates	3	0.70
Willingness to improve	I expect that I will be willing to improve my performance	3	0.80
Affect			
Positive affect	I expect that the feedback from my classmates will satisfy me	3	0.85
Negative affect	I expect that the feedback from my classmates will offend me	3	0.89

Table 2 Overview of the six scales of Strijbos et al.'s (2010) questionnaire regarding students' perceptions of peer assessment, including example items, number of items, and Cronbach's alpha coefficients

^a Based on pre-test

changes in response to peer assessment, and zero denotes not making changes. Only the inquiry output with value one (i.e., adjustments were made after peer assessment) was included in further analyses as the goal is to differentiate between the two experimental conditions, and comparison can only be made when adjustments are made after peer assessment.

To examine the potential impact of peer assessment, the inquiry output was assessed twice, namely before and after. The same rubric was used as the one student groups received in class. Following evaluation, scores on each assessment criterion were added per inquiry product. This resulted in a total of 15 points for defining the research question and data and a total of 10 points for the conclusion. To make comparisons more straightforward, the latter was converted to a total of 15 points. The first author of this article assessed all of the inquiry output. A second independent rater was trained to use the rubric and assessed 30 percent of the inquiry output. To verify inter-rater reliability, Cohen's Kappa was calculated. Cohen's Kappa was 0.81, which indicates almost perfect inter-rater agreement (0.81–1.00; Cohen, 1960).

Data-analysis

Descriptive analytics were used to get insight into the data in general. Multilevel analysis was performed to answer the aforementioned RQs as the gathered data are organized hierarchically. The pre- and posttest involves pupils who are nested in classes who are nested in schools. In the case of inquiry output, student groups are nested in classes that are nested in schools. Thus in both cases, a three-level model was considered. The analyses included two independent variables: the inquiry phase (i.e., research question, data, and conclusion) and experimental condition (i.e., PA versus PA + Dialogue). An identical approach was used each time when performing the analyses. In the first phase, an unconditional three-level null model is created without the independent variables. This null model indicates whether a multilevel analysis is needed to analyse the data and investigates the amount of variance at each of the three levels. If a level does not contribute to explaining variance, a new null model was created where only the relevant levels were included. In the next phase, the two independent variables (i.e., inquiry phase and experimental condition) are added to the fixed part of the model. Regarding the frequency of the adjustments, a generalized linear mixed model was fitted with the aforementioned binary variable as a dependent variable. In the case of the pre- and post-test and quality of the inquiry output, linear mixed models were fitted with as a dependent variable either the score difference between the two test moments or the score difference between the two assessment moments. Tukey post hoc tests were used each time to examine if there were significant differences between the three inquiry phases. The statistics software R was used to perform the analyses.

Table 3 Descriptive statistics concerning students' results on the pre- and post-test

	Control condition		PA condition		PA + Dialogue condition	
	PRE	POST	PRE	POST	PRE	POST
Knowledge	8.61 (2.45)	10.34 (2.75)	7.28 (3.35)	12.74 (3.98)	7.36 (3.20)	13.39 (3.59)

Note Scores need to be interpreted on a total of 20 points

Results

Pre- and post-test results: students' scientific knowledge about climate change

Table 3 summarizes, per condition, the descriptive results of the knowledge pre- and post-test.

Multilevel analysis was carried out to determine whether Climate colLab contributes to knowledge acquisition and whether there is a difference between the two experimental conditions. In Table 4, a threelevel model is presented. In the intercept-only model, Model 0, the estimated intercept in the fixed part of the model is 5.28 (SE = 0.65), representing the overall mean knowledge gain between the two test moments. In Model 1, the predictor 'condition', with the control condition as the reference category, is added and was found to be significant. The results reveal that students in the control condition do not show a significant increase in knowledge between the pre- and post-test, as the estimated intercept 1.79 (SE = 1.27) does not significantly differ from zero.

Post hoc analyses were carried out to compare the knowledge increase between the three conditions with one another. Tukey's post hoc tests indicate that students in the PA and PA + Dialogue conditions show a significantly higher knowledge gain in comparison to students of the control condition. No significant difference in knowledge gain between the PA and PA + Dialogue conditions was found. Hence, these results indicate that Climate colLab is effective regarding knowledge acquisition as the participants in both experimental conditions show a significant gain in

 Table 4
 Summary of the model estimates for the three-level analysis of students' knowledge acquisition

	Model 0		Model 1			
	Coefficient	SE	Coefficient	SE		
Fixed effect						
Intercept	5.28***	0.65	1.79	1.30		
PA			3.72*	1.44		
PA + Dialogue			4.33**	1.45		
Random effect						
School variance	3.24	1.80	1.16	1.08		
Class variance	1.46	1.21	1.61	1.27		
Student variance	8.97	3.00	8.99	3.00		
Model fit						
Log-likelihood	- 983.54		- 979.98			
χ^2 , df = 2			7.1121			
p			0.03			
* <i>p</i> < 0.05						
** <i>p</i> < 0.01						

Table 5	Number	of adjus	stments	according	to	the	experim	iental
condition	n							

Adjustments	Conditio	Condition		
	PA	PA + Dialogue		
Yes	112	124	236	
No	166	141	307	
Total	278	265	543	

knowledge between the two test moments compared to students of the control condition who did not participate in Climate colLab.

RQ1: Number of adjustments

To answer RQ1, Table 5 summarizes the number of adjustments made within the experimental conditions. Across all student groups, 236 times (43.46%), it was chosen to make changes to the inquiry output after the peer assessment episode, and 307 times (56.54%), it was decided not to. A Chi-square test showed that the proportion of adjustments in the PA condition (40.29%; n=112) and in the PA+Dialogue condition (46.79%; n=124) do not differ significantly from each other ($\chi^2 = 2.08$, df = 1, p = 0.15).

Additionally, Table 6 shows the number of adjustments across the three types of inquiry output. Of all the student groups, 94 (50.27%) and 96 (51.61%) student groups adjusted their research question and data after peer assessment. In contrast, only 46 (27.01%) student groups decided to adapt their research conclusion after peer assessment.

Via multilevel analysis, shown in Table 7, it was examined if the number of adjustments differs across experimental conditions and type of inquiry output. Model 1 supports the trends identified in the aforementioned descriptive analyses because it indicates no significant effect of the predictor 'condition' and a significant effect of the predictor type of 'inquiry output' ($\chi^2 = 25.99$, df = 2, p < 0.001). Additionally, there was no evidence of an interaction effect between condition and inquiry output.

Table 6 Number of adjustments according to the type of inquiry output

Adjustments	Inquiry outp	Total		
	Research question	Data	Conclusion	
Yes	94	96	46	236
No	93	90	124	307
Total	187	186	170	543

^{***} p < 0.001

Table 7	Summary	of	the	model	estimates	for	the	one-level
analysis o	f the numb	er d	of ad	justmer	nts			

	Model 0		Model 1			
	Coefficient	SE	Coefficient	SE		
Fixed effect						
Intercept	- 0.27**	0.09	- 0.12	0.19		
Data			0.06	0.22		
Conclusion			- 1.06***	0.24		
PA + Dialogue			0.26	0.20		
Random effect						
Group variance	0.17	0.41	0.29	0.53		
Model fit						
Log-likelihood	- 371.20		- 355.61			
χ^2 , df = 3			31.19			
р			<i>p</i> < 0.001			
*** p < 0.01						

^{***} *p* < 0.001

Further post hoc analyses were conducted to compare the number of adjustments of each type of inquiry output with one another. The Tukey test for post hoc comparisons indicates a significant difference between the number of adjustments of the research question and conclusion. Likewise, a significant difference was found between the data and the conclusion. More specifically, the research conclusion is substantially less frequently adjusted than the research question and data. The adjustment frequency of the research question and data does not differ significantly.

RQ2: Quality of the inquiry output

To answer RQ2, only student work to which adjustments were made after the peer assessment episodes were included, as the goal is to detect any differences in outcomes between the two experimental conditions. This leads to 236 observations (as found in Tables 6 and 7) of 147 student groups, from which 68 were in the PA condition and 79 were in the PA + Dialogue condition. Table 8 shows the descriptive results of the inquiry output scores

 Table 9
 Summary of the model estimates for the two-level analysis of students' pre-test scores

	Model 0		Model 1			
	Coefficient	SE	Coefficient	SE		
Fixed effect						
Intercept	9.06***	0.23	11.12***	0.25		
Data			- 1.22***	0.29		
Conclusion			- 8.06***	0.36		
PA + Dialogue			- 0.08	0.14		
Random effect						
School variance	0.00	0.00	0.13	0.37		
Group variance	12.83	3.58	3.91	1.98		
Model fit						
Log-likelihood	- 635.45		- 496.15			
χ^2 , $df = 3$			278.6			
р			p<0.001			

^{***} p < 0.001

before, and after making adjustments per experimental condition.

Multilevel analysis was applied to the quality scores of the different types of inquiry before the peer assessment to find out whether there was a difference in the starting level of the students depending on the experimental conditions and the type of inquiry output. Model 1 demonstrates, as shown in Table 9, that there is no significant influence of 'condition', meaning the groups' beginning levels in both experimental conditions are the same. Nevertheless, once more, a significant impact of the type of 'inquiry output' on the pre-scores was found ($\chi^2 = 530.63$, df=2, p<0.001). More specifically, it was found using Tukey's post hoc test that the starting level for developing a research question is significantly higher than the starting level of both research data and conclusion. The starting level for searching research data is significantly higher than for formulating a research conclusion.

In a subsequent stage, the difference scores are analysed to determine how much the different types of inquiry output have been enhanced in response to the peer assessment episodes. A descriptive analysis

Table 8 Descriptive results of students' scores who made adjustments per type of inquiry output according to the experimental condition

	PA condition			PA + Dialogue condition			
	PRE (M, SD)	POST (M, SD)	Improvement (M, SD)	PRE (M, SD)	POST (M, SD)	Improvement (M, SD)	
Research Question	11.00 (2.31)	13.70 (1.64)	+ 2.70 (2.06)	11.19 (2.07)	13.47 (1.72)	+2.28 (1.70)	
Data	9.98 (2.30)	12.32 (2.07)	+ 2.34 (1.55)	9.85 (2.11)	11.98 (2.01)	+2.13 (1.41)	
Conclusion	2.81 (0.81)	6.10 (1.78)	+ 3.29 (1.39)	3.36 (0.70)	6.36 (1.50)	+ 3.00 (1.62)	

Note Scores need to be interpreted on a total of 15 points



of the progress achieved for each experimental condition and type of inquiry output is previously shown in Table 8. It reveals that in both experimental conditions, the progress made appears to be most significant for the conclusion. Figure 6 depicts the distribution of the inquiry output's difference scores. Descriptive analysis reveals that when groups made revisions, their inquiry outputs received an average score increase of 2.51 points. The most considerable improvement is 7.5

points, but regression is also noted by 1.5 points.

Multilevel analysis was used to investigate if difference scores vary across conditions and the type of inquiry output. As Model 1 in Table 10 shows, no significant effect of 'condition' was found. However, a significant effect of the type of 'inquiry output' $(\chi^2 = 9.971, df = 2, p < 0.01)$ was observed, meaning that the made progress size varies according to the inquiry output. Using the Tukey test for post hoc comparison, it can be determined that the mean progress made when adjusting the conclusion is, on average, significantly 0.95 (SD = 0.30) and 0.74 (SD = 0.31) points higher than the progress achieved after adjusting the research question and data, respectively. The progress achieved when adjusting the research question and data does not differ significantly.

Table 10	Summary	of th	ne	model	estimates	for	the	two-level
analysis of	f student gi	oups	′ di	fference	e scores			

	Model 0		Model 1			
	Coefficient	SE	Coefficient	SE		
Fixed effect						
Intercept	2.51***	0.11	2.63***	0.26		
Data			- 0.21	0.24		
Conclusion			0.74	0.31		
PA + Dialogue			- 0.48	0.25		
Random effect						
School variance	0.00	0.00	0.21	0.45		
Group variance	3.00	1.73	2.79	1.67		
Model fit						
Log-likelihood	- 463.92		- 458.28			
χ^2 , df=3			11.271			
p			0.01			
**** n < 0.001						

^{***} *p* < 0.001

RQ3: Peer assessment perceptions

To answer RQ3, the questionnaire of Strijbos et al. (2010) was registered during the pre- and post-test. The results in Table 11 demonstrate that students generally agreed with the statement regarding fairness, usefulness, accept-ance, and willingness to improve, with scores ranging

Table 11	Descriptive	results	of	students'	questionnaires
regarding	peer assessme	ent perce	ptior	ns (Strijbos e	et al., <mark>2010</mark>)

	PA condition	ı	PA + Dialogue condition		
	PRE (M, SD)	POST (M, SD)	PRE(M, SD)	POST (M, SD)	
Fairness	6.80 (1.73)	7.18 (2.17)	6.95 (1.47)	7.00 (2.69)	
Usefulness	6.39 (2.28)	6.43 (2.60)	6.33 (2.17)	6.58 (2.82)	
Acceptance	7.21(1.87)	7.51(2.25)	7.18 (1.82)	7.05 (2.51)	
Willingness	6.40 (1.93)	6.71(2.01)	6.47 (1.85)	7.18 (1.85)	
Positive affect	5.31 (2.12)	5.88 (2.45)	5.56 (2.03)	6.06 (2.69)	
Negative affect	2.27 (2.30)	1.80 (2.33)	2.36 (2.36)	1.86 (2.41)	

Note Measured on a bi-polar scale ranging from 0 (= fully disagree) to 10 (= strongly agree)

between 6.33 and 7.51 before and after Climate colLab. Regarding the negative affect scale, students' scores range between 1.80 and 2.36, thus indicating that they disagree with the assertions that peer assessment during Climate colLab would elicit or provoke negative emotions. Finally, given that their ratings ranged from 5.31 to 6.06, students demonstrated a neutral stance regarding the notion that the peer assessment would stimulate or provoke positive emotions.

To explore whether students' expectations about peer assessment during Climate colLab would be fulfilled and if perceptions about peer assessment would differ between the PA and PA+Dialogue condition after Climate colLab, multilevel analysis was conducted for each of the six self-reported scales of perceptions about peer assessment which can be found in Table 12. Regarding

Table 12 Summary of the model estimates for the three-level or two-level analysis of students' difference scores regarding peer assessment perceptions

	Fairness			Usefulness				Willingness to improve				
	Model 0		Model 1		Model 0		Model 1		Model 0		Model 1	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed effect												
Intercept	0.07	0.27	0.25	0.33	- 0.18	0.28	- 0.31	0.34	0.33	0.26	0.06	0.33
PA + DIALOGUE			- 0.38	0.37			0.25	0.37			0.48	
Random effect												
School variance	0.27	0.52	0.28	0.53	0.35	0.59	0.35	0.59	0.20	0.45	0.32	0.20
Class variance	0.19	0.44	0.20	0.44	0.02	0.13	0.04	0.20	0.34	0.58	0.26	0.34
Student variance	6.47	2.54	6.47	2.54	8.28	2.88	8.28	2.88	5.17	2.27	5.17	5.17
Model fit												
Log-likelihood	- 739.93		- 739.39		- 775.63		- 775.37		- 707.57		- 706.84	
$\chi^2 (df = 1)$			1.09				0.51				1.46	
р			0.30				0.47				0.23	
	Acceptance				Positive affe	ect			Negative af	fect ¹		
	Model 0		Model 1		Model 0		Model 1		Model 0		Model 1	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed effect												
Intercept	0.03	0.23	0.23	0.28	0.40	0.23	0.41	0.30	- 0.49**	0.16	- 0.47*	0.22
PA + DIALOGUE			- 0.39	0.31			- 0.04	0.39			- 0.02	0.31
Random effect												
School variance	0.20	0.45	0.22	0.47	0.09	0.31	0.10	0.31				
Class variance					0.26	0.51	0.30	0.55				
Student variance	6.35	2.52	6.33	2.52	6.86	2.62	6.86	2.62				
Model fit												
Log-likelihood	- 733.23		- 732.43		- 748.19		- 748.18					
$\chi^2 (df=1)$			1.62				0.02					
р			0.20				0.90					

¹ Linear regression was applied as variance equals 0

* *p* < 0.05

** *p* < 0.01

fairness, usefulness, acceptance of peer assessment, willingness to improve, and positive affect, the intercepts in the fixed part of the intercept-only models (Model 0) do not differ significantly from zero. This means no significant difference is found between the pre- and post-test scores on these five perceptions scales.

In the case of the negative affect perception scale, the intercept in the fixed part of the unconditional null significantly differs from zero. This means a significant difference exists between the expectations regarding and experiences with peer assessment during Climate colLab. The significant slope is negative, suggesting that peer assessment does not provoke as many negative emotions during Climate colLab as students initially expected.

To address RQ3, the variable 'condition', with PA condition serving as the reference category, was later added to each model. None of the six models showed a significant effect of condition, as shown in Table 11.

Discussion

This research aimed to examine the impact of peer assessment as a specific scaffolding mechanism during CSCiL. To accomplish this, a lesson series called Climate colLab was developed in the web-based learning environment SCORE. Students were grouped in pairs, and classes were randomly divided into three conditions: one control condition and two experimental conditions that incorporated peer assessment, with or without additional peer dialogue. During the first half of the lesson series, students familiarized themselves with the different stages of an inquiry cycle and gained a solid understanding of fundamental concepts related to climate change. In the second half of the lesson series, students focused on conducting their own sustainability research. Recognizing the importance of scaffolding for successful IL (Alfieri et al., 2011; Lazonder & Harmsen, 2016), this study implemented peer assessment during students' research to improve their inquiry output in terms of research question, data, and conclusion. Specifically, it examined the influence of peer assessment as a scaffolding tool on three aspects: (1) the extent to which students made adjustments to their inquiry output, (2) the quality of students' inquiry output, and (3) students' perceptions of peer assessment during CSCiL. Notably, this study stands out as the first to explicitly implement peer assessment as a scaffolding tool within CSCiL, and it does so using a large sample size.

Concerning the effectiveness of the Climate colLab lesson series, it was discovered that the intervention significantly enhanced students' scientific conceptual knowledge of climate change compared to students in the control group, who showed no significant improvement. This finding aligns with previous research indicating that IL promotes scientific conceptual understanding (Furtak et al., 2012; Heindl, 2019). Moreover, this improvement was consistent across both experimental conditions, as anticipated. The knowledge assessed was acquired during the first half of the lessons, which was identical in both experimental conditions.

As for the first research question, which examines the number of adjustments made to the three types of inquiry output, it was found that the number of adjustments did not differ between both experimental conditions. However, interestingly, the number of adjustments corresponds to the number of times the feedback was ignored, which is unexpected considering previous studies that focused on finding the most optimal peer assessment format within CSCiL reported students predominantly ignoring peer feedback (Hovardas et al., 2014; Tsivitanidou et al., 2011). Notably, although the frequency of adjustments remained consistent across experimental conditions, it did vary depending on the type of inquiry output. Specifically, the findings revealed that although the number of adjustments made to the research question and data was similar, it was significantly higher than the number of adjustments made to the conclusion. Students tend to make fewer adjustments as they progress through the inquiry cycle. This is consistent with the research of Tsivitanidou and colleagues (2012), who found a regression in the use of peer assessment by students over time when final inquiry products were assessed. The content of peer feedback is often cited as the primary explanation for this outcome in previous studies. For instance, students primarily offered praise rather than critical remarks that would assist in adjusting and enhancing their inquiry output (Hovardas et al., 2014). Alternatively, students could fail to provide compelling arguments, resulting in less persuaded assessees to make changes (Tsivitanidou et al., 2012). Since the content of peer feedback was not examined in this study, it is impossible to determine if these factors were at play. However, this study addressed these previous research findings by including prompts to encourage students to offer constructive criticism of their peers' work and support their arguments. Based on our findings, the prompts only positively affect the number of adjustments made to the research question and data. A second explanation for the significantly lower adjustments to the conclusion is its position as the last step of the inquiry cycle. As the preceding steps demanded significant time and effort from students, the fewer adjustments to the conclusion could be attributed to a probable time limitation or work overload for learners. Peer assessment and CSCiL are learning activities that impose a heavy workload on students (Hovardas et al., 2014; Raes & Schellens, 2016), possibly leading students to choose the "shortcut" of making no adjustments to complete the task faster and with less effort. Therefore, the conclusion phase may be negatively impacted by its position at the end of the inquiry cycle.

Regarding the second research question, which focuses on the quality of students' inquiry output, it was observed that when students chose to make adjustments, there was a significant improvement in the quality of all three types of inquiry output. This indicates that peer assessment as a scaffolding technique can enhance inquiry output, specifically improving the quality of a research question, data, and conclusion. This finding confirms previous evidence that peer assessment as a formative assessment practice is an effective instructional strategy across various contexts (Double et al., 2020). However, the extent of the improvement of inquiry output varied across the three types of inquiry output but not according to experimental conditions. Specifically, the progress made after adjusting the conclusion was more substantial than the improvement made after adjusting the research question and data. A possible explanation for this may be that the scores on the pre-test for the quality of the conclusion were considerably lower before the peer assessment, indicating a greater potential for improvement. Moreover, prior to reaching a conclusion, students must engage in challenging learning tasks that demand specific skills. One such task involves accurately interpreting gathered data, a known hurdle for students (de Jong & Van Joolingen, 1998), but a vital step preceding formulating conclusions. Wu and Hsieh (2006) also identified the evaluation of scientific explanations as a difficult inquiry skill necessary for drawing conclusions. Hence, the inherent complexity of formulating conclusions might explain the generally lower quality scores observed in research conclusions.

Regarding the last research question, it was found that students were more inclined to expect that peer assessment during Climate colLab would be fair and valuable. They were willing to accept the peer assessment and modify their research based on it. Additionally, they anticipated that the peer assessment would contribute to a positive emotional experience. Students' expectations regarding fairness, usefulness, acceptance, willingness to improve, and positive affect were confirmed throughout Climate colLab. These rather positive expectations regarding peer assessment are in accordance with the findings of, for example, Rotsaert and colleagues (2018) and Loretto and Demartino (2016), who found that students overall possess a positive predisposition towards the use of peer assessment to optimize their learning process. Specifically for this study, this could be attributed to the scaffolds (i.e., training with rubrics and worked examples and question prompts) implemented within the peer assessment process.

Page 15 of 21

Adding peer dialogue to the peer assessment process did not influence the number of adjustments (RQ1), progress made (RQ2), or perceptions (RQ3) following peer assessment. Despite theoretical propositions advocating the inclusion of peer dialogue in peer assessment suggesting potential benefits like improved attitudes and increasing the use of peer feedback by enabling explanatory discussions and seeking consensus (Carless, 2016; Tsivitanidou et al., 2012), the empirical evidence from this study does not strongly validate these assertions but do not undermine these claims either. A potential explanation is that while peer assessment is widespread in schools in the studied research context, it predominantly relies on quantitative grading systems (Double et al., 2020; Rotsaert et al., 2017). Thus, students might be familiar with grading peers but need more experience engaging in feedback dialogues with peers (Double et al., 2020; Planas Lladó et al., 2014). Doing so, this feedback practice becomes a key strategy within the whole classroom feedback culture.

Implications for practice

Acknowledging the significance of scaffolding in fostering effective IL, this research employed peer assessment within students' research to enhance their inquiry output in terms of research question, data, and conclusion. Based on the results, this study shows evidence of the effectiveness of peer assessment as a scaffolding tool within CSCiL. Educators and instructional designers can reflect on their current instructional practices and consider how peer assessment and scaffolding mechanisms align with their teaching methods. Exploring opportunities to integrate similar strategies into their teaching context allows for enhancing collaborative IL experiences.

From this research, strategies can be gleaned to leverage peer assessment effectively, enhancing students' inquiry outputs and refining their quality. The findings of this study shed light on tailoring feedback strategies across different phases of the inquiry cycle. Scaffolding via peer assessment during different phases of the inquiry cycles should take different forms targeted at specific areas where students commonly require support. This can be done by adapting the general question prompts for formulating peer feedback to concentrate on the specific challenges students face with each inquiry output (e.g., interpreting data or evaluating scientific evidence). This approach directs student assessors' attention to the particular difficult challenging aspects of each inquiry output, consequently providing student assessees with the necessary assistance.

Finally, understanding students' perceptions regarding peer assessment is pivotal. Our study showed that adding training, question prompts, and peer dialogue to peer assessment in a CSCiL environment positively influenced students' perceptions toward peer assessment.

Implications for future research

Based on the findings of this study, future research should focus on guiding students to systematically outline positive and negative aspects while proposing improvements in qualitative peer assessment messages. Integrating an AI-driven intelligent tutor that monitors and analyses the data generated throughout the peer assessment activities could actively scaffold the peer feedback dialogue via the chat tool and prompt students for elaborate arguments and provide question prompts for high-quality feedback (Gan & Hattie, 2014). Subsequently, the effectiveness of the AI-driven tutor could be measured through content analysis of the peer dialogues and mapping challenges students face during these interactions.

Regarding the low post-test scores of the conclusion phase in both experimental conditions, further research is needed to deepen our understanding of the possible impact of differences in peer assessment task complexity, and associated cognitive load for the subphases of the inquiry cycle (de Jong & van Joolingen, 1998; Wu & Hsieh, 2006).

Limitations

This study captured students' perceptions of peer assessment through a brief questionnaire, leaving certain aspects of these perceptions unexplored. Future research could employ qualitative methods such as focus groups or interviews to delve deeper into students' perspectives to enhance our understanding. This qualitative approach would offer a more comprehensive exploration of their experiences, allowing for a closer examination of the potential actions they suggest to influence their perceptions positively in future instances.

Although this study includes an intervention period that was already considerably long, it is worth considering the possibility of extending the intervention further in future studies. When implementing CSCiL in the classroom, students must be given sufficient time to get acquainted with peer assessment as a scaffolding tool within CSCiL. It can be expected that this is a learning process for students, and they need the proper time to develop the necessary skills to take full advantage of the benefits peer scaffolds can offer during the inquiry process. Additionally, it is crucial to consider the teacher's role and how they might contribute to CSCiL both as a participant in peer assessment and as a scaffolding support.

Conclusion

In summary, this study aimed to examine the impact of peer assessment as a scaffolding mechanism in the context of CSCiL. The results showed that the frequency of adjustments (RQ1) varied depending on the type of inquiry output, with more adjustments to the research question and data compared to the conclusion. Regarding the quality of students' inquiry output (RQ2), it was observed that when students made adjustments, the quality significantly improved across all types of inquiry output. Notably, the most substantial improvements were seen in the conclusion. Students' perceptions of peer assessment (RQ3) indicated positive expectations regarding fairness, usefulness, acceptance, willingness to improve, and limited negative affect. Students generally accepted peer assessment and were willing to adapt based on feedback. Lastly, no additional impact of including a peer dialogue in the peer assessment process was found on the outcomes mentioned above. Overall, this study enhances our understanding of peer assessment as a scaffolding tool in CSCiL, highlighting its potential to improve the quality of inquiry outputs and providing valuable insights for instructional design and implementation.

Appendix A

Scoring rubric for each type of inquiry output Research question (RQ)

	Very good	±	Sufficient	±	Insufficient
Scope Check if the RQ is not for- mulated too broadly or too narrowly	Your RQ is limited to one specific topic (Who/ What, Where, When, How, To What Extent). It is an open- ended RQ to which the answer is not appar- ent		Your RQ is an open question to which the answer is not appar- ent. However, the RQ is still worded too broadly. Try to delineate the ques- tion even more (Who/ What, Where, When, How, To what extent)		The answer to your RQ is obvious and requires no research OR the RQ is phrased too broadly and impos- sible to answer with online information

	Very good	±	Sufficient	±	Insufficient
Importance of sustainable living Consider whether the RQ is important within sustain- able living	What you want to explore is vital for building a sustainable life		What you want to explore is relevant to building a sustainable life. However, the link to sustain- ability needs to be clarified from your question		What you want to explore is not essen- tial in designing a sustain- able life
Language Review the wording of the RQ	Your RQ begins with a ques- tion word. The question is clearly written and contains no language errors		Your RQ begins with a ques- tion word. The question is clearly written but con- tains some language errors		Your RQ does not begin with a ques- tion word. In addition, the RQ reads stiffly and con- tains several language errors

Data

	Very good	±	Sufficient	±	Insufficient
Choice of informa- tion source Check whether the cho- sen informa- tion sources are reliable and whether data have been searched in differ- ent places	The sources of informa- tion you chose are reliable. Moreover, you looked for infor- mation in different places		The sources of information you chose are reliable. You can, however, broaden your informa- tion search by looking for informa- tion in even more different places		The sources of informa- tion you chose are not at all reli- able. Moreo- ver, you only looked for informa- tion in one place
Occurrence of 3Ps Check whether enough information has been found on each of the 3 Ps (people, planet, profit)	You searched extensively for infor- mation about each of the 3 Ps to answer your research question		You have looked up infor- mation about each of the 3 Ps. There is still room for possible expansion of the infor- mation, though		You did not look up informa- tion on one or more P's

	Very good	±	Sufficient	±	Insufficient
Accuracy Check if the selected information is substantively correct	All the infor- mation you found is correct		Errors do not imme- diately sur- face. But some of the infor- mation you found is untrace- able online and therefore unverifiable		The informa- tion you found con- tains several errors

Conclusion

	Very good	±	Sufficient	±	Insufficient
Occurrence of 3Ps Check whether suf- ficient attention has been paid to each of the 3 Ps in answering the RQ (people, planet, profit)	The conclusion addresses each of the 3 Ps in detail. Conse- quently, the RQ is answered in detail		The conclusion addresses each of the 3 Ps. The RQ is answered succinctly. There is an oppor- tunity for expansion		One or more P's are not addressed in the conclu- sion
Language Check whether enough nformation nas been found on each of the 3 Ps (people, olanet, profit)	Your conclusion is very clear, coherent, and flaw- lessly written		Your conclu- sion is clearly written but contains some lan- guage errors. Sometimes there needs to be more consistency between sen- tences		Your conclu- sion is written stiffly, contains multiple lan- guage errors, and needs coherence. Moreover, you copied the answer from the inter- net

Appendix B

Exemplary question of the knowledge test

- Compare the two figures below: Which part of Figure A resembles the glass in Figure B? 1. Indicate what you think is the correct answer. 2. Then describe in as much detail as possible what happens to the energy coming from the Sun. Use the correct scientific terms when doing so.

Figure A



Figure B



Which part of Figure A resembles the glass in Figure B? 1. Indicate what you think is the correct answer.

- O The Universe
- O Earth
- The Atmosphere
- O The Sun
- O I don't know

Describe in as much detail as possible what happens to the energy coming from the Sun. Use the correct scientific terms when doing so.

Appendix C Scoring rubric of the knowledge test

Scores	Response description
0	Students have no or incorrect and irrelevant ideas in the given context
1	Correct multiple choice answer, but without further explanation
2	Correct multiple choice answer with further explanation, but rather isolated and still some incorrect and irrelevant ideas are included
3	Students have correct and relevant ideas but do not fully elaborate links between them in the given context. They still fail to connect the relevant ideas

4 Students recognize connections between scientific concepts and understand how they interact. They have a systematic understanding and apply this in their explanation and argumentation

Abbreviations

3Ps	People, planet, profit (three principles of sustainability)
CSCiL	Computer-supported collaborative inquiry-based learning
CSCL	Computer-supported collaborative learning
IL	Inquiry-based learning
PA	Peer assessment
RQ	Research question
SCORE	Scripting and Orchestration Environment
STEM	Science, Technology, Engineering, and Mathematics

Acknowledgements

We would like to thank the ENCORE lab of the University of Toronto for programming the peer assessment tool in the SCripting and ORchestration Environment (SCORE) to make this study possible.

Author contributions

AVH: conceptualization, methodology, formal analysis, investigation, resources, data curation, writing—original draft, writing—review and editing, and visualization; JW: data curation, and writing—review and editing; TR: writing—review and editing, and supervision; TS: writing—review and editing, and supervision.

Funding

This study was funded by the Special Research Fund of Ghent University.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 17 July 2023 Accepted: 5 January 2024 Published online: 25 January 2024

References

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, 103(1), 1–18. https://doi.org/10.1037/a0021017

Barron, B., & Darling-Hammond, L. (2010). Prospects and challenges for inquiry-based approaches to learning. In H. Dumont, D. Istance, & F. Benavides (Eds.), *The nature of learning: using research to inspire practice* (pp. 199–225). OECD.

- Belland, B. R., Walker, A. E., Kim, N. J., & Lefler, M. (2017). Synthesizing results from empirical research on computer-based scaffolding in STEM education: A meta-analysis. *Review of Educational Research*, 87(2), 309–344. https://doi.org/10.3102/0034654316670999
- Blotnicky, K. A., Franz-odendaal, T., French, F., & Joy, P. (2018). A study of the correlation between STEM career knowledge, mathematics self- efficacy, career interests, and career activities on the likelihood of pursuing a STEM career among middle school students. *International Journal of STEM Education*, 5, 1–15. https://doi.org/10.1186/s40594-018-0118-3
- Capps, D. K., Crawford, B. A., & Constas, M. A. (2012). A review of empirical literature on inquiry professional development: Alignment with best practices and a critique of the findings. *Journal of Science Teacher Education, 23*(3), 291–318. https://doi.org/10.1007/s10972-012-9275-2
- Caprara, G. V., Vecchione, M., Alessandri, G., Gerbino, M., & Barbaranelli, C. (2011). The contribution of personality traits and self-efficacy beliefs to academic achievement: A longitudinal study. *British Journal of Educational Psychology*, 81(1), 78–96. https://doi.org/10.1348/2044-8279.002004
- Carless, D. (2015). Exploring learning-oriented assessment processes. *Higher Education*, 69(6), 963–976. https://doi.org/10.1007/s10734-014-9816-z
- Carless, D. (2016). Feedback as dialogue. *Encyclopedia of Educational Philosophy* and Theory. https://doi.org/10.1007/978-981-287-532-7_389-1
- Chen, J., Wang, M., Kirschner, P. A., & Tsai, C.-C. (2018). The role of collaboration, computer use, learning environments, and supporting strategies in CSCL: A meta-analysis. *Review of Educational Research*, 88(6), 799–843. https:// doi.org/10.3102/0034654318791584
- Chen, W., & Tan, J. S. H. (2021). The spiral model of collaborative knowledge improvement : An exploratory study of a networked collaborative classroom. *International Journal of Computer-Supported Collaborative Learning*, 16, 7–35. https://doi.org/10.1007/s11412-021-09338-6
- Chu, S. K. W., Reynolds, R. B., Tavares, N. J., Notari, M., & Lee, C. W. Y. (2017). 21st century skills development through inquiry-based learning: from theory to practice. Springer International Publishing. https://doi.org/10.1007/ 978-981-10-2481-8
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational* and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/ 001316446002000104
- Constantinou, C. P., Tsivitanidou, O. E., & Rybska, E. (2018). Professional development for inquiry-based science teaching and learning. In O. E. Tsivitanidou, P. Gray, E. Rybska, L. Louca, & C. P. Constantinou (Eds.), *Professional development for inquiry-based science teaching and learning* (Vol. 5, pp. 1–26). Springer. https://doi.org/10.1007/978-3-319-91406-0
- de Jong, T. (2019). Moving towards engaged learning in STEM domains; there is no simple answer, but clearly a road ahead. *Journal of Computer Assisted Learning*, *35*(2), 153–167. https://doi.org/10.1111/jcal.12337
- de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, 68(2), 179–201. https://doi.org/10.3102/00346543068002179
- Deiglmayr, A. (2018). Instructional scaffolds for learning from formative peer assessment: Effects of core task, peer feedback, and dialogue. *European Journal of Psychology of Education*, 33(1), 185–198. https://doi.org/10. 1007/s10212-017-0355-8
- Dillenbourg, P. (2009). Exploring neglected planes: Social signals and class orchestration. In A. Dimitracopoulou, C. O'Malley, D. Suthers, & P. Reimann (Eds.), Proceedings of the 9th international conference on Computer supported collaborative learning (Vol. 2, pp. 6–7).
- Dmoshinskaia, N., Gijlers, H., & de Jong, T. (2021). Learning from reviewing peers' concept maps in an inquiry context: Commenting or grading, which is better? *Studies in Educational Evaluation, 68*, 100959. https://doi.org/10.1016/j.stueduc.2020.100959
- Dmoshinskaia, N., Gijlers, H., & Jong, T. D. (2020). Giving feedback on peers' concept maps in an inquiry learning context: the effect of providing assessment criteria. *Journal of Science Education and Technology, 30*(3), 420–430. https://doi.org/10.1007/s10956-020-09884-y
- Dobber, M., Zwart, R., Tanis, M., & van Oers, B. (2017). Literature review: The role of the teacher in inquiry-based education. *Educational Research Review*, 22, 194–214. https://doi.org/10.1016/j.edurev.2017.09.002
- Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group

studies. Educational Psychology Review, 32(2), 481–509. https://doi.org/10. 1007/s10648-019-09510-3

- Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A metaanalysis. *Review of Educational Research*, 82(3), 300–329. https://doi.org/10. 3102/0034654312457206
- Gan, M. J. S., & Hattie, J. (2014). Prompting secondary students' use of criteria, feedback specificity and feedback levels during an investigative task. *Instructional Science*, *42*(6), 861–878. https://doi.org/10.1007/s11251-014-9319-4
- Heindl, M. (2019). Inquiry-based learning and the pre-requisite for its use in science at school : A meta-analysis. *Journal of Pedagogical Research*, 3(2), 52–61.
- Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, 42(2), 99–107. https://doi.org/10.1080/00461520701263368
- Hovardas, T., Tsivitanidou, O. E., & Zacharia, Z. C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers and Education*, 71, 133–152. https://doi.org/ 10.1016/j.compedu.2013.09.019
- Husnaini, S. J., & Chen, S. (2019). Effects of guided inquiry virtual and physical laboratories on conceptual understanding, inquiry performance, scientific inquiry self-efficacy, and enjoyment. *Physical Review Physics Education Research*, 15(1), 10119. https://doi.org/10.1103/PhysRevPhysEducRes.15. 010119
- Jiang, J. P., Hu, J. Y., Zhang, Y. B., & Yin, X. C. (2022). Fostering college students' critical thinking skills through peer assessment in the knowledge building community. *Interactive Learning Environments*. https://doi.org/10. 1080/10494820.2022.2039949
- Ketelhut, D. J. (2007). The impact of student self-efficacy on scientific inquiry skills: An exploratory investigation in river city, a multi-user virtual environment. *Journal of Science Education and Technology, 16*(1), 99–111. https://doi.org/10.1007/s10956-006-9038-y
- Kim, M. C., & Hannafin, M. J. (2011). Scaffolding problem solving in technologyenhanced learning environments (TELEs): Bridging research and theory with practice. *Computers and Education*, 56(2), 403–417. https://doi.org/ 10.1016/j.compedu.2010.08.024
- Kim, N. J., Belland, B. R., Lefler, M., & Andreasen, L. (2020). Computer-based scaffolding targeting individual versus groups in problem-centered instruction for stem education: Meta-analysis. *Educational Psychology Review*, 32, 415–461. https://doi.org/10.1007/s10648-019-09502-3
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75–86. https://doi.org/10.1207/s1532 6985ep4102_1
- Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction*, 20(4), 344–348. https://doi.org/ 10.1016/j.learninstruc.2009.08.005
- Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research, 86*(3), 681–718. https://doi.org/10.3102/0034654315627366
- Linn, M. C., Eylon, B. S., Kidron, A., Gerard, L., Toutkoushian, E., Ryoo, K. K., Bedell, K., Swearingen, A., Clark, D. B., Virk, S., Barnes, J., Adams, D., Acosta, A., Slotta, J., Matuk, C., Hovey, C., Hurwich, T., Sarmiento, J. P., Chiu, J. L., & Laurillard, D. (2018). Knowledge integration in the digital age: Trajectories, opportunities and future directions. *Proceedings of International Conference of the Learning Sciences, ICLS, 2*, 1259–1266.
- Liu, O. L., Lee, H. S., Hofstetter, C., & Linn, M. C. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational* Assessment, 13(1), 33–55. https://doi.org/10.1080/10627190801968224
- Loretto, A., & Demartino, S. (2016). Secondary students' perceptions of peer review of writing. Research in the Teaching of English, 51(2), 134–161.
- Mäeots, M., Pedaste, M., & Sarapuu, T. (2008). Transforming students' inquiry skills with computer-based simulations. In: Proceedings - The 8th IEEE International Conference on Advanced Learning Technologies, ICALT 2008, pp. 938–942. https://doi.org/10.1109/ICALT.2008.239
- Matuk, C., Linn, M., & Eylon, B.-S. (2015). Technology to support teachers using evidence from student work to customize technology-enhanced

inquiry units. Instructional Science, 43, 229–257. https://doi.org/10.1007/ s11251-014-9338-1

- Matuk, C., Tissenbaum, M., & Schneider, B. (2019). Real-time orchestrational technologies in computer-supported collaborative learning: An introduction to the special issue. *International Journal of Computer-Supported Collaborative Learning*, 14(3), 251–260. https://doi.org/10.1007/ s11412-019-09310-5
- Mckeown, J., Hmelo-Silver, C. E., Jeong, H., Hartley, K., Faulkner, R., & Emmanuel, N. (2017). A meta-synthesis of CSCL literature in STEM education. In CSCL 2017 Proceedings, pp. 439–446. Retrieved 19 June 2023, from https:// repository.isls.org/bitstream/1/262/1/59.pdf
- Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E. (2015). Phases of inquirybased learning: Definitions and the inquiry cycle. *Educational Research Review*, 14, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003
- Pietarinen, T., Palonen, T., & Vauras, M. (2021). Guidance in computer-supported collaborative inquiry learning: Capturing aspects of affect and teacher support in science classrooms. *International Journal of Computer-Supported Collaborative Learning*, 16, 261–287. https://doi.org/10.1007/ s11412-021-09347-5
- Planas Lladó, A., Feliu Soley, L., Fraguell Sansbelló, R. M., Arbat Pujolras, G., Pujol Planella, J., Roura-Pascual, N., Suñol Martínez, J. J., & Montoro Moreno, L. (2014). Student perceptions of peer assessment: An interdisciplinary study. Assessment and Evaluation in Higher Education, 39(5), 592–610. https://doi.org/10.1080/02602938.2013.860077
- Raes, A., & Schellens, T. (2016). The effects of teacher-led class interventions during technology-enhanced science inquiry on students' knowledge integration and basic need satisfaction. *Computers & Education*, 92–93, 125–141. https://doi.org/10.1016/J.COMPEDU.2015.10.014
- Raes, A., Schellens, T., & De Wever, B. (2014). Web-based collaborative inquiry to bridge gaps in secondary science education. *Journal of the Learning Sciences*, 23(3), 316–347. https://doi.org/10.1080/10508406.2013.836656
- Rotsaert, T., Panadero, E., Estrada, E., & Schellens, T. (2017). Studies in educational evaluation how do students perceive the educational value of peer assessment in relation to its social nature ? A survey study in Flanders. *Studies in Educational Evaluation, 53*, 29–40. https://doi.org/10.1016/j. stueduc.2017.02.003
- Rotsaert, T., Panadero, E., & Schellens, T. (2018). Anonymity as an instructional scaffold in peer assessment: Its effects on peer feedback quality and evolution in students' perceptions about peer assessment skills. *European Journal of Psychology of Education, 33*(1), 75–99. https://doi.org/10.1007/s10212-017-0339-8
- Shabani, K. (2016). Applications of Vygotsky's sociocultural approach for teachers' professional development. *Cogent Education*, 3(1), 1252177. https:// doi.org/10.1080/2331186X.2016.1252177
- Sharples, M., Scanlon, E., Ainsworth, S., Anastopoulou, S., Collins, T., Crook, C., Jones, A., Kerawalla, L., Littleton, K., Mulholland, P., & O'malley, C. (2015). Personal inquiry: Orchestrating science investigations within and beyond the classroom. *Journal of the Learning Sciences*, 24(2), 308–341. https://doi. org/10.1080/10508406.2014.944642
- Strijbos, J.-W., Pat-El, R., & Narciss, S. (2010). Validation of a (peer) feedback perceptions questionnaire comics for vocational education-CoforVE view project serena view project. In L. Dirckinck-Holmfeld, V. Hodgson, C. Jones, M. de Laat, D. McConnell, & T. Ryberg (Eds.), *Proceedings of the 7th International Conference on Networked Learning 2010* (pp. 378–386).
- Tissenbaum, M., & Slotta, J. (2019). Supporting classroom orchestration with real-time feedback: A role for teacher dashboards and real-time agents. *International Journal of Computer-Supported Collaborative Learning*, 14(3), 325–351. https://doi.org/10.1007/s11412-019-09306-1
- Tsivitanidou, O., Zacharia, Z. C., & Hovardas, T. (2011). Investigating secondary school students' unmediated peer assessment skills. *Learning and Instruction*, 21(4), 506–519. https://doi.org/10.1016/j.learninstruc.2010.08.002
- Tsivitanidou, O., Zacharia, Z. C., Hovardas, T., & Nicolaou, A. (2012). Peer assessment among secondary school students: Introducing a peer feedback tool in the context of a computer supported inquiry learning environment in science. *Journal of Computers in Mathematics and Science Teaching*, 31(4), 433–465.
- van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacherstudent interaction: A decade of research. *Educational Psychology Review*, 22(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6

- Velamazán, M., Santos, P., Hernández-Leo, D., & Vicent, L. (2023). User anonymity versus identification in computer-supported collaborative learning: Comparing learners' preferences and behaviors. *Computers & Education*, 203, 1–16. https://doi.org/10.1016/j.compedu.2023.104848
- Voet, M., Gielen, M., Boelens, R., & De Wever, B. (2018). Using feedback requests to actively involve assesses in peer assessment: Effects on the assessor's feedback content and assesses' agreement with feedback. *European Journal of Psychology of Education, 33*(1), 145–164. https://doi.org/10. 1007/s10212-017-0345-x
- Winstone, N., & Carless, D. (2020). Designing effective feedback processes in higher education: A learning-focused approach. Routledge.
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry, 17(2), 89–100.
- Wu, H. K., & Hsieh, C. E. (2006). Developing sixth graders' inquiry skills to construct explanations in inquiry-based learning environments. *International Journal of Science Education*, 28(11), 1289–1313. https://doi.org/10.1080/ 09500690600621035
- Xenofontos, N. A., Hovardas, T., Zacharia, Z. C., & Jong, T. (2019). Inquiry-based learning and retrospective action: Problematizing student work in a computer-supported learning environment. *Journal of Computer Assisted Learning*, 36(1), 12–28. https://doi.org/10.1111/jcal.12384

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.