Hämmerle, Linda; Möller, Andrea; Bergmann-Gering, Alexander; u. a.

Investigating students' awareness of their own and others' deviations from controlled science experiments

Date of secondary publication: 07.08.2026

Version of Record (Published Version), Article

Persistent identifier: urn:nbn:de:bvb:473-irb-116617x

Primary publication

Hämmerle, Linda; Möller, Andrea; Bergmann-Gering, Alexander; u. a. (2026): Investigating students' awareness of their own and others' deviations from controlled science experiments, in: International Journal of Science Education, London [u.a.]: Taylor & Francis, Vol. 48, No. 8, pp. 1197–1219, doi: 10.1080/09500693.2024.2447556.

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION
2026, VOL. 48, NO. 8, 1197–1219
https://doi.org/10.1080/09500693.2024.2447556

Investigating students’ awareness of their own and others’ deviations from controlled science experiments

Linda Hämmerle [a], Andrea Möller [a,b], Alexander Bergmann-Gering [c], Theresa Krause-Wichmann [d] and Judith Lederman [e]

[a]Austrian Educational Competence Centre for Biology, Centre for Teacher Education, University of Vienna, Vienna, Austria; [b]Faculty of Life Sciences, Department of Evolutionary Biology, University of Vienna, Vienna, Austria; [c]Faculty of Life Sciences, Department of Biology, University of Leipzig, Leipzig, Germany; [d]Department of Education, Saarland University, Saarbrücken, Germany; [e]Mathematics & Science Education Department, Illinois Institute of Technology, Chicago, IL, USA

ABSTRACT

Despite an emphasis on experimentation in science curricula worldwide as part of efforts to improve scientific literacy, students encounter challenges, especially with designing unconfounded experiments as part of the control-of-variables strategy (CVS). Becoming aware of experimental design errors is a potential starting point for learners to enhance their experimental skills. Few studies investigated whether learners can accurately assess experiments and are aware of errors. This experimental study, conducted during the COVID-19 pandemic, investigates the accuracy of students’ assessment of experiments and their awareness of self-generated and others’ (vicarious) design errors. 127 students (grade 7–8) were randomly split into two groups. One group conducted an experiment themselves. The other examined an erroneous example of a fictitious peer. Afterwards, both received the same instructions on the CVS and prompts to assess the experiments. Data were collected via worksheets and photos of their experiments. Analysis reveals difficulties controlling all variables in an experiment, especially if they were continuous. Interestingly, while self- and peer assessment accuracy was generally high, students were significantly more aware of vicarious errors than of self-generated ones. This highlights the potential of using assessment of experimental design errors as learning opportunity for experimentation skills, especially when using vicarious errors.

ARTICLE HISTORY
Received 22 March 2024
Accepted 20 December 2024

KEYWORDS
Experimentation skills; control-of-variables strategy; error awareness

Introduction

An understanding of how a scientific consensus is formed can enhance trust in science (Chinn & Colan Duncan, 2018). It is thus unsurprising that an understanding of scientific

CONTACT Andrea Möller andrea.moeller@univie.ac.at Austrian Educational Competence Centre for Biology, University of Vienna, Porzellangasse 4/2, 1090 Vienna, Austria
This article has been corrected with minor changes. These changes do not impact the academic content of the article.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/09500693.2024.2447556.

practices – such as experiments, which are part of what constitutes consensus building in science – is integral to scientific literacy (OECD, 2007, 2018), and is consequently reflected in both science standards and curricula worldwide (e.g. BMBWF, 2019; KMK, 2005; National Research Council, 2012, 2013). According to the Austria curriculum, experimentation, as a key scientific method and component of scientific literacy, plays a significant role in biology education and should be incorporated into every grade level of Austrian middle schools (BMBWF, 2019, 2024). Controlled, unconfounded experiments in which there is only a single difference made between individual treatments are the basis for argumentations about causality (Chen & Klahr, 1999). The skills needed to design and interpret such experiments are crucial to the understanding of core scientific concepts (Schwichow et al., 2022; Siler & Klahr, 2012). These skills, including planning, interpreting and distinguishing unconfounded experiments from confounded ones, are subsumed under the 'control-of-variables strategy' (CVS) (Chen & Klahr, 1999).

However, students face numerous challenges during experimentation such as not stating causal research questions or altering conditions during the experiment (Kranz et al., 2022). Students struggle especially with the CVS, such as controlling for variables (Kranz et al., 2022). This leads to 'design errors' in experiments (Schwichow et al., 2022). While defining errors is challenging, Oser et al. (1999) and Gartmeier et al. (2008) describe them as deviations from the norm that may hinder the achievement of desired goals. In this context, design errors in experiments are deviations from the CVS as a norm, which guides users towards controlled experiments to answer research questions. However, these errors should not be inherently avoided; instead, they can help to gain 'negative knowledge', which provides an understanding of how something does not work. In turn, this allows for an understanding of the correct concepts or processes (Gartmeier et al., 2008; Oser et al., 1999). Allchin (2012) underscores the potential of errors as learning opportunities in science education, stating, 'If the goal is to teach 'how science works' […], then it seems equally important to teach, on some occasions, how science *does not* work […]. Every error helps profile (by contrast) an element of how science works […]' (Allchin, 2012, pp. 904–905). However, learning from errors does not happen spontaneously. Research has shown that students need to become aware (after making an error and receiving additional information) that a solution is not correct to engage cognitive processes involved in learning (Loibl & Leuders, 2019); this necessitates an accurate self-assessment of their errors. Typically, working with self-generated errors can be demanding, requiring a strong sense of self-regulation to persist after making errors and reflect on their causes (Tulis et al., 2016). One potential means of avoiding this issue is to use the errors of others, also known as 'vicarious errors', as valuable learning opportunities (Kapur, 2014). Switching from self- to peer assessment allows individuals to avoid engaging with their own errors, removing the inherent challenges, which could affect their error awareness. Additionally, it allows for plannable and scalable learning activities, as are needed in science classrooms (Morris et al., 2024). So far, few studies have looked at students' self- and peer assessment in the context of experimentation (Schreiber, 2016; Schreiber & Theyssen, 2019), much less their error awareness in the design of their own experiments or experiments conducted by someone else.

While there is a lack of research on students' error awareness in experiments, several studies have investigated the challenges that students encounter when conducting experiments (e.g. Dasgupta et al., 2016; García-Carmona et al., 2016). However, a recent review by Kranz et al. (2022) revealed that most of these studies do not use practical performance assessments, but rather assess the students' experimentation skills following the experimentation process, using methods such as multiple-choice tests or interviews. This approach may be problematic, as Hammann et al. (2008) have shown that students' performance in multiple-choice tests does not appear to predict their performance in actual experiments. Consequently, the measurement of practical performance requires authentic hands-on activities (Schreiber & Gut, 2022).

Against this background, this study aims to deepen our understanding of how middle school students deal with experimental design errors. We explore common design errors in a practical (hands-on) setting and assess the accuracy of students' self- and peer assessment. Additionally, we examine students' awareness of errors in their own experiment and in an experiment conducted by a fictitious peer. Our results can provide insight for targeted instructional materials focusing on experimentation and the CVS.

The control-of-variables strategy as a prerequisite for successful experimentation

Experiments can be defined as a method of investigating the causal relationship between dependent and independent variables (Arnold et al., 2014; Duit & Tesch, 2010). To draw valid conclusions from experiment's results, the experiment must be unconfounded (Chen & Klahr, 1999). Siler and Klahr (2012) propose three rules for designing unconfounded experiments: (1) identify the independent variable, (2) create two conditions/treatments and vary the value of the independent variable, and (3) keep all other variables constant in both conditions/treatments. These rules are integral to the CVS (Chen & Klahr, 1999) and Schwichow et al. (2022) define an experimental design error as a deviation from these rules. A full understanding of the CVS includes procedural aspects (identifying and planning unconfounded experiments) and logical aspects (interpreting controlled experiments and understanding the inherent indeterminacy of confounded experiments) (Chen & Klahr, 1999; Schwichow et al., 2022). Kranz et al. (2022) provide an overview of the challenges students face during experimentation, including those related to the application of the CVS. Common CVS errors include the indiscriminate handling of experimental materials and substances (e.g. Baur, 2018), the lack of determination of test times (e.g. Arnold et al., 2014) and the failure to control variables (e.g. Chen & Klahr, 1999; Siegler & Liebert, 1975). As previously mentioned, many studies assess students' experimentation skills with interviews (Dasgupta et al., 2016) or paper-pencil tests (García-Carmona et al., 2016) rather than hands-on activities. Consequently, these tests may not provide an accurate assessment of practical performance (Hammann et al., 2008; Kranz et al., 2022; Schreiber & Gut, 2022). Nevertheless, these studies show how difficult it is for students to understand and apply the CVS, highlighting the need for innovative teaching and learning strategies. One promising approach might be the use of students' errors as an opportunity for learning (Tulis et al., 2016). Research reveals that this is true for both self-generated and vicarious errors (Loibl & Leuders, 2019; Oser et al., 1999). To facilitate such learning through errors, three key mechanisms must be engaged: (1) prior knowledge is activated, (2) students' attention is directed to the strategy/content to be learned, and (3) students become aware of errors (Loibl et al., 2017; Sinha & Kapur, 2019, 2021; Zhang & Fiorella, 2023). Thus, to present a learning opportunity, students need to assess errors accurately.

Self-assessment, peer assessment and error awareness

Self-assessment involves descriptive and evaluative processes regarding one's own work and/or abilities (Brown & Harris, 2013; Panadero, Jönsson, et al., 2016), while peer assessment entails judging real or fictitious work of others (Reinholz, 2016; Safadi, 2022). The accuracy of self- and peer assessment is typically determined by juxtaposing students' results with external expert assessment. The higher the percentual match, the more accurate the assessment (Andrade, 2019). In sum, self- and peer assessment involve the recognition of areas for improvement and the acknowledgement of strengths. Error awareness on the other hand can be considered to be part of this self- and peer assessment as conceptualised by Panadero, Jönsson, et al. (2016). It is a student's ability to detect errors (Zhang & Fiorella, 2023), which is important, as recognising that something is erroneous is an important starting point for learning from errors.

Self-assessment involves a high degree of self-regulation, but it prompts learners to generate internal feedback to update prior knowledge (Andrade, 2019). This deep reflection can successfully promote learning (Andrade, 2019; Bond & Ellis, 2013). In a review of research on student self-assessment, Andrade (2019) notes that the accuracy of self-assessment varies. For example, there is a difference between task-specific self-assessment (the assessment of one's results from a specific task) and general self-assessment (the assessment of one's general knowledge in a specific area), with task-specific self-assessment generally being more accurate (Andrade, 2019). Age also plays a role, as younger students tend to overestimate their abilities (Andrade, 2019; Brown & Harris, 2013). In addition, proficiency in a self-assessed skill improves this accuracy (Panadero, Brown, et al., 2016). Students also benefit from clear criteria, such as scaffolds, exemplification, or checklists (Andrade, 2019; Panadero, Brown, et al., 2016).

Peer assessment harbours high potential for learning (Li et al., 2020; Panadero, Jönsson, et al., 2016; Reinholz, 2016; Stančić, 2021). Not only does it benefit the assessee but it also benefits the assessor by exposing them to shortcomings they may not have been aware of, while at the same time encouraging them to take responsibility (Topping, 2017). Like self-assessment, accuracy of peer assessment is higher for more advanced students and when it is well organised with clear criteria (Li et al., 2016; Topping, 2017). The interpersonal aspects of peer assessment can be challenging but can be mitigated by making it anonymous (Panadero & Alqassab, 2019) or by using fictitious erroneous examples. This provides students with the benefits associated with peer feedback, while mitigating some of its challenges (Reinholz, 2016; Safadi, 2022). In this study, students judge fictitious erroneous examples (i.e. teacher-crafted solutions that include vicarious errors by a fictitious student).

Self-assessment, peer assessment and error awareness in science experimentation

There is a lack of studies focusing on students' error awareness in experiments, and few studies have investigated self- and peer assessment in this specific context. Stefani (1994) examined the quality of self- and peer assessment of undergraduate students. However, the students in the study assessed laboratory reports rather than experiments. 87 students took part in the self-assessment, while another 67 students took part in the peer assessment. Stefani (1994) reported satisfying accuracies for both assessment types. Schreiber (2016) examined the accuracy of self-assessment of 28 10th graders in the context of experimentation. Overall, the results revealed high accuracy (Schreiber, 2016). Schreiber and Theyssen (2019) replicated this in a study with 180 students, once again showing that the accuracy of the self-assessment was high. The authors also investigated possible influential factors and found that experimentation skills, motivation, and aspects of cognitive load were the strongest discriminants, leading them to assume that the accuracy of the self-assessment depends on the task at hand (Schreiber & Theyssen, 2019).

In summary, existing studies demonstrate promising findings regarding the accuracy of self- and peer assessment. However, there is a research gap in term of students' ability to self- and peer assess experiments, particularly when it comes to error awareness. For students to effectively learn from errors, it is crucial that they can accurately identify and assess errors. This process of identifying errors allows students to understand where they went wrong and enables them to gain insights into the underlying concepts or procedures (Zhang & Fiorella, 2023). By engaging in thorough error assessment, students can pinpoint areas for improvement and develop a deeper understanding of the subject matter. Consequently, our study addresses this gap by investigating the differences in self- and peer assessment with a focus on students' error awareness.

Research questions

The CVS challenges students of all ages, and errors have been proven to be excellent learning opportunities in different domains (Sinha & Kapur, 2021; Zhang & Fiorella, 2023). We argue that there is also potential for learning from design errors in experimentation. However, most studies on students' challenges during experimentation use non-practical means of assessing performance, and there is a clear lack of research regarding one of the key mechanisms by which errors induce learning: students' error awareness. In the context of this study, this refers to their ability to assess design errors in their own and others' experiments.

The main focus of this study is to investigate whether there is a difference in error awareness between self-generated and vicarious design errors in experiments. To accomplish this, first, a practical hands-on approach is used to assess experimental design errors of middle school students regarding the CVS. Then, the accuracy of their self- and peer assessment are assessed before focusing on their error awareness specifically. Hence, we investigate the following research questions: (1) Which CVS-related design errors do students make when conducting an experiment? (2) How accurate is their self-assessment, and how accurately do students judge others' experiments? (3) Is there a difference in the awareness of self-generated and vicarious experimental design errors?

Methodology

Participants

Prior to the main study, the learning materials were piloted with a sample of 65 students (7th and 8th grade, $M_{age} = 12.85$; SD = 0.79) from two classes at a school that also participated in the main study. These students were not included in the main study. A total of 144 middle school students (7th and 8th grade, $age_{range} = 12–15$) across 12 classes from five schools in Austria participated in the main experimental study. Both the pilot and the main study were conducted in the *Austrian Educational Competence Centre for Biology (AECC Biology)* teaching-and-learning-laboratory at the University of Vienna.

All data were collected during the COVID-19 pandemic, in the periods between school closures due to government lockdowns. School selection was based on the ability to conduct out-of-school activities during the pandemic. Prior to participation, students received information about the research objectives, procedure, and data confidentiality. Participation was voluntary, and only students whose parents provided consent were included. Data were collected and analysed pseudonymously, using personal codes to link different parts of their worksheets.

As their experiments did not qualify for our assessment of CVS design errors (see section 'measures'), 17 of the 144 students were excluded from the analysis of the main study. Ten students did not include all necessary components in their experiments, seven students built single condition experiments. Hence, the final sample consisted of 127 students ($M_{age} = 12.75$; SD = 0.67; 42.43% female) (Table 1).

Overview of the study procedure

This study comprises two steps: the preparation and the study itself. The preparation involved the development of worksheets, training of pre-service teachers to teach the lessons, and piloting the worksheets. During the study itself, data were collected. Afterwards, the measures were quantified, and the results were analysed (see Figure 1).

Preparation

The worksheets for our lessons were developed based on Schmidt (2016), Lui (2021) and Scholtes (2021). Schmidt (2016) developed a method for the practical assessment of experimentation skills, where students compared the insulating capacity of sheep's wool and pig fat. As suggested by the author, pigs fat was exchanged for down feathers. Scholtes (2021) and Lui (2021) built upon Schmidt (2016), focusing on fostering students' experimentation skills by learning from errors. Their materials were adapted to the equipment of our *AECC Biology* teaching-and-learning laboratory to enhance the workflow (see section 'Detailed explanation of the lessons' for a detailed description). The planned lessons align with the Austrian curriculum for 7th and 8th grade, both on a methodological level (by teaching experimental skills) as well as content wise (by offering an introduction to the connections between structures, ways of live and habitats of animals; BMBWF, 2019).

Table 1. Demographic characteristics of the participants ($n = 127$).

Groups	*N*	Age *M*	Age SD	Age n.a. (%)	Gender (%) Female	Gender (%) Male	Gender (%) n.a.
Self-generated errors	58	12.8	0.70	5.17	41.38	53.45	5.17
Vicarious errors	69	12.7	0.64		43.48	52.17	4.35

Note: Gender was assessed with the following categories: female, male, and diverse; missing answers were coded as n.a.

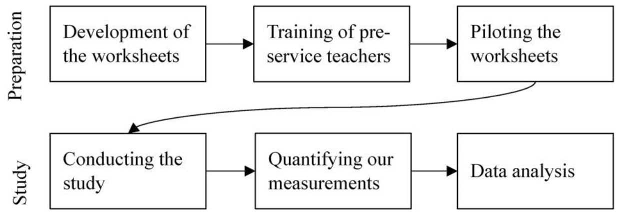

Figure 1. Overview of the study procedure.

The study, including the piloting, was conducted within the context of an internship for pre-service teachers who taught the groups in pairs. All pre-service teachers attended a seminar on teaching experimental competence prior to the study. Additionally, they attended a training session where they were informed about the exact procedure of the study and the lessons they were to teach in the form of a standardised script, including possible reactions to students' questions. Student's worksheets were designed by the authors to be self-explanatory, requiring no intervention from the pre-service teachers. Teacher educators acted as mentors and accompanied each team of pre-service teachers, ensuring adherence to the script and preventing disruptions.

The intervention worksheets were piloted with 65 students from 7th to 8th grade (M_{age} = 12.85; SD = 0.79) to ensure that the workflow was smooth. As with the data collection, piloting took place in a period between school closures. Based on this, analogue thermometers were exchanged for digital thermometers and the written explanation of the students' experiment was supplemented with a sketch.

Research setting

Per session, one school class visited our *AECC Biology* teaching-and-learning-laboratory. Students from each class were randomised into treatment groups at the entrance. All students worked individually. The groups were separated by room dividers to prevent treatment diffusion. Both groups followed the 'problem solving prior to instruction' approach (Sinha & Kapur, 2021). This approach is organised in two phases. Figure 2 outlines the

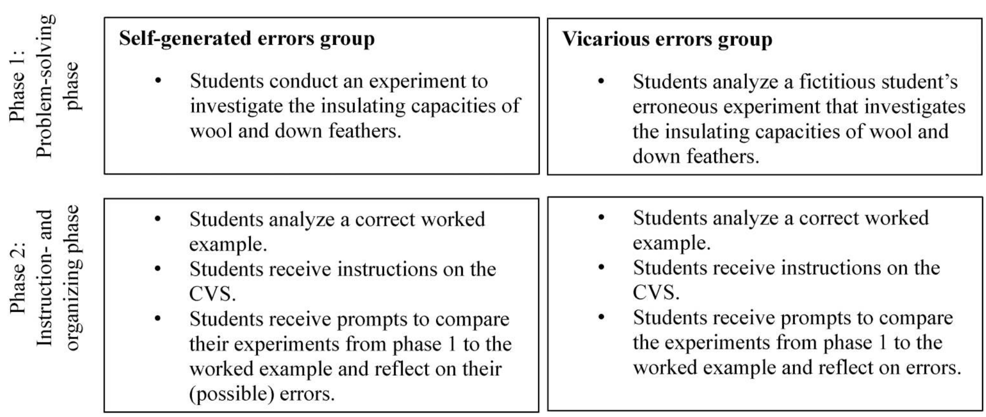

Figure 2. Simplified representation of the lessons delivered to both treatment groups.

key aspects of the worksheets for the self-generated errors and vicarious errors group that are analysed in this study. The students finished both parts in 75 min.

In phase 1, students engage in problem-solving activities regarding concepts they have yet to learn. In our study, these activities involved experiments with a focus on the CVS. This usually creates erroneous solution attempts. In phase 2, learners receive instructions on the concepts (i.e. the CVS) and reflect on phase 1, aiming for a reorganisation of their prior knowledge (Sinha & Kapur, 2021). A detailed explanation on students' worksheets and workflow can be found in section 'Detailed explanation of the lessons'. In preparation for the statistical analyses, our measures were quantified, which included occurring errors, the accuracy of students' self- and peer assessment, and their error awareness (see section 'measures'). Afterwards, statistical analyses were conducted (see section 'data analysis').

Detailed explanation of the lessons

All students worked with worksheets. A translated version can be found in the supplemental material. Both groups were provided their research question, hypothesis, and a list of available materials that could be used to conduct the experiment. Both groups were provided with written explanations on how the experiment could be conducted, but without step-by-step instructions or mentioning the CVS (see Figure 3). These worksheets were designed based on the practical performance assessment from Schmidt (2016).

The pre-service teachers provided standardised instructions on the materials at the beginning of the lesson, omitting instructions on the CVS.

During phase 1, both treatment groups worked with an experiment to answer the question if wool or down is the better insulator. An experiment to answer this question could look like this: A bigger beaker filled with ice creates a cold environment. A smaller beaker with the insulator is placed inside, representing the insulating layer. A test tube with warm water is placed inside the insulator to represent an animal's body. A thermometer and a stopwatch are used to measure how quickly the water temperature decreases.

The primary difference in the way the two groups engaged with the experiment is that the self-generated errors group conducted the experiment. Students were provided with all the necessary laboratory materials to correctly conduct the experiment, plus additional materials to enlarge the error potential (e.g. four different sizes of beakers instead of only two). The vicarious errors group was physically presented with an erroneous experiment from a fictitious peer and asked to examine the provided experiment with a checklist (see Figure 4), ensuring that they were actively engaging with it. The experiment included five errors that were based on findings from Schmidt (2016) and Scholtes (2021): (1) varying ice quantity between treatments, (2) varying insulator quantity, (3) varying water level, (4) varying test tube position, and (5) varying beaker position. Each respective error was presented well visible (e.g. there was a big difference in the water level). Students in the vicarious errors group were told that the experiments in phase 1 were erroneous, but not told how many errors the experiment included.

In phase 2, both groups received a written explanation of the CVS and a correct worked example to contrast with the experiment from phase 1. As suggested in the

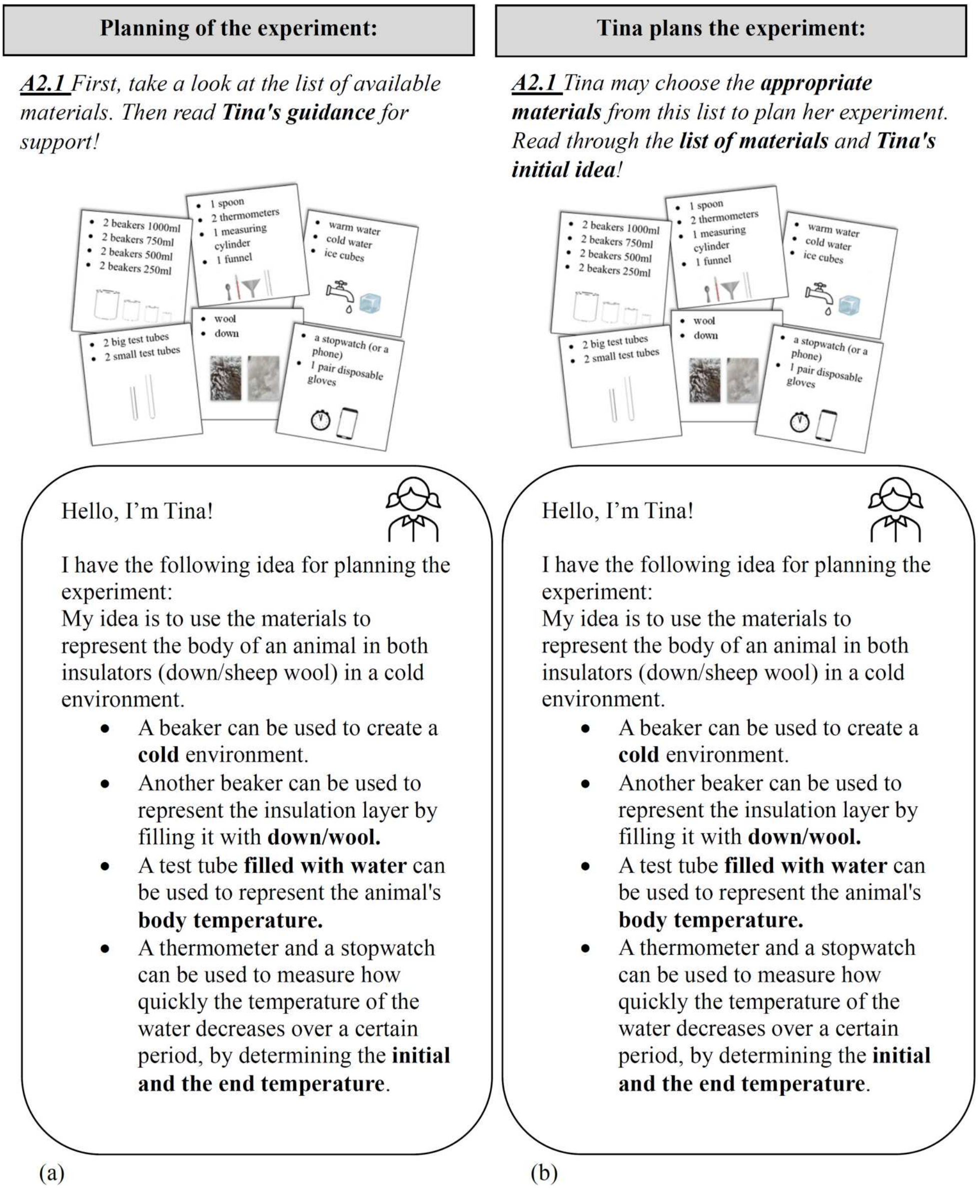

Figure 3. Examples of the worksheets from phase 1. Both groups were provided with the same information about how the experiment could be conducted, with the prompts slightly varying. (a) Excerpt of the prompts given to the students from the self-generated errors group. (b) Excerpt of the prompts given to the students from the vicarious errors group.

literature, students were prompted to reflect on potential errors in the experiments from phase 1 (Loibl & Leuders, 2019), with both groups given identical prompts by a fictional character Tom. These included a checklist of the experiments' variables, with students asked to identify if they were held constant across both conditions (see Figure 5).

(a)

Tina conducts the experiment:

<u>A3.</u> Tina conducted the experiment. You can see **the experimental setup** in front of you. **Examine** Tina's experiment and use the table below to indicate which materials she used for the down and the wool condition.

Material	Down condition		Wool condition	
a: Size of the outer beakers	☐ Different sizes		☐ Equal sizes	
b: Size of the inner beakers	☐ Different sizes		☐ Equal sizes	
c: Ice quantity	☐ A lot	☐ Little	☐ A lot	☐ Little
d: Insulator quantity	☐ A lot	☐ Little	☐ A lot	☐ Little

(b)

Figure 4. Examples of the worksheets given to the vicarious errors group in phase 1. (a) Photo of the erroneous experiment the vicarious errors group was presented with physically. (b) Example of the prompts given to the students for them to engage with the experiment, translated from German.

<u>A5.</u> Compare **your** experiment from phase 1 with the correct experiment and Tom's tips.

Tick **all** the aspects that you did **right** or **wrong**.

	I did this **right**	I did this **wrong**
Equal-sized beakers for the **cold** environment:	☐	☐
Equal amount of ice cubes:	☐	☐
Equal-sized beakers for **thermal insulation**:	☐	☐
Same amount of insulating material:	☐	☐

(a)

<u>A5.</u> Compare **Tina's** experiment from phase 1 with the correct experiment and Tom's tips.

Tick **all** the aspects that she did **right** or **wrong**.

	Tina did this **right**	Tina did this **wrong**
Equal-sized beakers for the **cold** environment:	☐	☐
Equal amount of ice cubes:	☐	☐
Equal-sized beakers for **thermal insulation**:	☐	☐
Same amount of insulating material:	☐	☐

(b)

Figure 5. Sample of the worksheets from phase 2, translated from German. (a) Self-generated errors group. (b) Vicarious errors group.

To keep the groups comparable despite being taught by different pre-service teachers, the students did not receive individual feedback on either their own experiments and their self-assessment or the peer assessment of the fictitious students' experiments at the end of phase 2 (see also section 'limitations').

Measures

Occurring errors

To assess the errors that occurred during the experiment, the experiments from the self-generated errors group were photographed and analysed. The students did not need to write a detailed protocol about their procedure, however, they were asked to note the measurement durations (i.e. wait times between temperature measurements in both conditions) and the starting temperature, as these details would not be visible in a photograph.

We adapted the category system from Schmidt (2016) to assess all variables of the experiments, with a focus on the CVS. This coding frame was mostly concept-driven, with categories and subcategories emerging from the logic of the experiment (Schreier, 2012). This means that the main categories are based on the variables that occur in the experiment, e.g. the size of test tubes or the amount of insulating material used in each condition. Each category had two subcategories: Each variable (main category) was assessed to see if it was held constant across treatments (subcategory 1) or if it was erroneously varied (subcategory 2). The coding guideline includes definitions for each category, rules for the subcategories, and student examples, as recommended in the literature (Schreier, 2012). At the beginning of the coding process, additional exclusion criteria were inductively established (e.g. students building only one condition instead of two, meaning that the experiment could not be assessed in the context of the CVS). The first author and a trained research assistant applied the coding guideline. This coding will be referred to as 'expert assessment'. Cohen's kappa coefficient ($\kappa = .99$) was calculated on a randomly selected 25% of the sample to determine reliability (Cohen, 1960). Coding was done using MAXQDA 2022 (VERBI Software, 2021). The codes were extracted from the software and used as the basis for a subsequent frequency analysis, as detailed in the section 'data analysis'.

Accuracy of the students' self- and peer assessment

To answer the questions of whether their self- and peer assessment is accurate, data was gathered within part 2 of the worksheets, with the checklist including the experiment variables. The self-generated errors group assessed their experiments, while the vicarious errors group assessed the fictitious peer's experiment. Both groups were asked to determine if each respective variable was correctly held constant across both conditions (see Figure 5). Students' self-assessment was compared with the expert assessment to determine its accuracy, as is common procedure (e.g. Andrade, 2019). Similarly, the vicarious errors groups' peer assessment was compared with the errors that were included in the fictitious student's experiment. This was done by creating a dummy variable for each of the assessed variables, where one of four codes was assigned: (1) error recognised by the student, (2) error not recognised by the student, (3) correct application of the CVS recognised by the student, and (4) correct application of the CVS not recognised by the student. Then, the percentages of the cases where the students' assessment matched the expert assessment were calculated. The higher the percentual match, the higher the accuracy of the self- or peer assessment.

Error awareness in the context of the CVS: self-generated vs. vicarious errors

Error awareness can be described as students' ability to detect errors (Zhang & Fiorella, 2023) and thus quantified similarly to the self- and peer assessment. We drew from the same material as for the self- and peer assessment, but focused on the errors (i.e. the variables that were varied according to the expert assessment). For both treatment groups, the students' assessment of those variables was juxtaposed to the expert assessment: The higher the percentual match, the higher the error awareness.

Data analysis

Statistical analysis was conducted using the software RStudio (RStudio Team, 2020). In order to investigate which CVS-related design errors occurred during experimentation, the codes from the expert's assessment of the photographed experiments and students' notes on the duration of the measurement and the starting temperature (see section 'measurement') were converted into dummy variables. This allowed for a frequency analysis of each error. To investigate the accuracy of the self- and peer assessment, a frequency analysis of the cases where the self-/peer assessment and expert assessment did or did not match was conducted. To compare students' awareness of self-generated and vicarious errors, the Wilcoxon rank sum test was used to compare the two treatment groups. The effect size Pearson's r was calculated, with $r = 0.10$ indicating a small effect, $r = 0.30$ a medium effect, and $r = 0.50$ a large effect (Cohen, 1992). Due to the nature of the tasks undertaken by the two treatment groups, it was not possible to control for the number of errors in the self-generated errors group, while each student in the vicarious errors group received an experiment with five predefined errors. To mitigate potential bias, additional comparisons of only students that were exposed to the same errors were done using Pearson's Chi-square test with Yates' continuity correction. The effect size Cohen's w was calculated, with $w = 0.10$ representing a small effect, $w = 0.30$ a medium effect, and $w = 0.50$ a large effect (Cohen, 1992).

Results

Occurring errors

A variety of CVS-related design errors were observed in the self-generated errors group. Figure 6 showcases the experiments from two students, including typical design errors the students made when they conducted the experiment individually. In this example, the students varied the beaker or test tube sizes, the amount of materials used or the water level between the treatments. All of these constitute deviations from the CVS, since the independent variable in the students' experiment was the quality of insulator (wool or down) and all other variables were to be held constant across treatments.

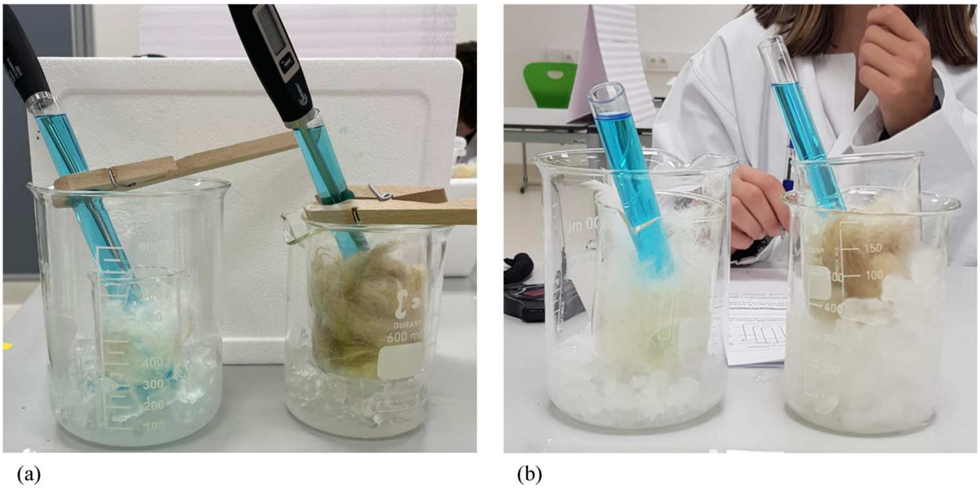

Figure 6. Examples of the students' experimental design errors. (a) Student 1: Varying outer beaker size, varying inner beaker size, varying amount of insulating material. (b) Student 2: Varying outer beaker size, varying test tube size, varying ice quantity, varying water level.

While these errors were specific to the experiment used in the study, they can be categorised into three generic categories that are applicable to most experimental settings in the natural sciences: (1) categorical variables, such as beaker and test tube sizes, were subsumed under the term 'choice of material'; (2) continuous variables, such as the water level or the insulator quantity, were subsumed under the category 'handling of material'; and (3) variables that related to the measuring process were categorised as 'measurement'. Figure 7 shows that most of the errors could be categorised under 'handling of material'.

Aggregated at a student level, nearly all students (98.28%) faced problems at least once with handling the given materials or substances. 29.31% of the students chose material indiscriminately at least once and did not consider that the treatment conditions must be kept comparable unless they are to be specifically changed. 10.34% of the students encountered difficulties in the measuring process; they did not determine the test times.

Accuracy of the students' self- and peer assessment

The accuracy of self-assessment was relatively high in all categories (choice of material: 89.66% agreement, handling of material: 58.91% agreement, measurement: 69% agreement). 4.66% of the overall variables from the students' experiments could not be analysed, because the students either did not conduct a valid self-assessment for the respective variable (e.g. they did not use the provided checklist correctly), or the variable could not be analysed regarding potential errors using the photos of their experiments (due to bad photo quality, e.g. the water level was hidden by the insulator) or worksheets (e.g. the students did not note the starting temperatures). Table 2 provides an overview of the self-assessment, split into the variables that the experts judged to be held constant across the treatments (CVS applied) and those that were judged to be varied across the treatments (CVS not applied). It shows a clear difference between students' ability to assess what they did correctly vs. what they did not.

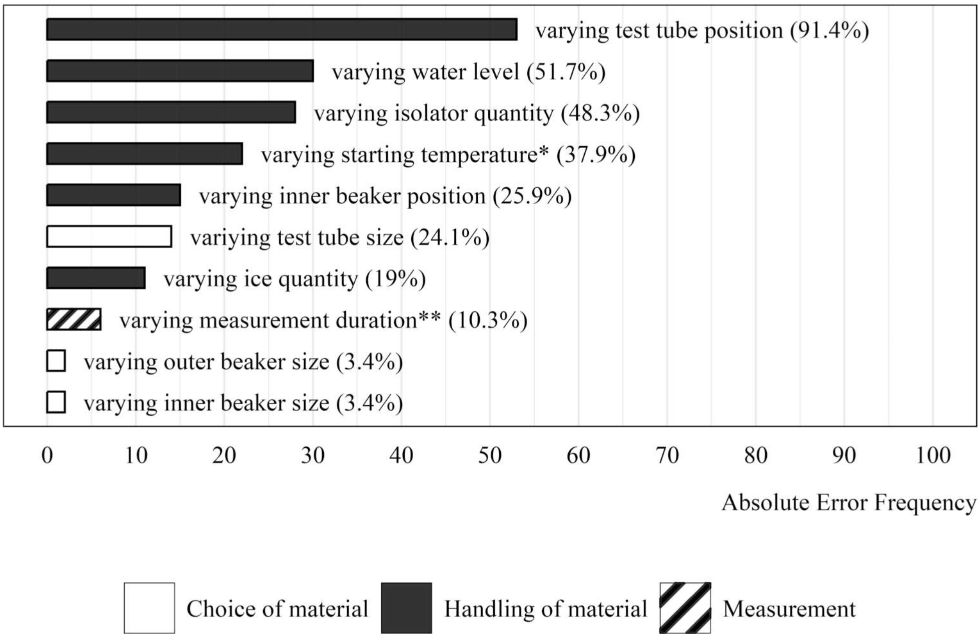

Figure 7. Overview of the errors observed in the self-generated errors group ($n_{students} = 58$), including the percentage of students that made the corresponding design error. Data with an * were derived from students' worksheets.

In cases where the variables were held correctly held constant across the treatments according to the expert judgement, the accuracy of students' self-assessment was high in all three categories. Choice of material was the easiest for students to assess (98.04% match with the expert assessment), followed by measurement (86.05%) and handling of material (78.73%).

Assessing their self-generated design errors was more difficult for students, meaning their error awareness was rather low. In the category 'measurement', only half the design errors were recognised. Looking at the categories 'choice' and 'handling of material', it was less than half of them (choice of material: 33.33%, handling of material: 42.77%).

The peer assessment conducted by the vicarious errors group exhibited high overall accuracy across all categories (choice of material 94.69% agreement; handling of material 85.54% agreement; measurement 88.41% agreement). A small portion of the data (4.38%) could not be assessed due to invalid student answers. Unlike with self-assessment, there was no stark difference between the variables that the experts judged to be held constant across treatments and those that were judged to be varied. The agreement between the peer and expert assessment when the CVS was applied was 95.15% for choice of material, 88.24% for handling of material and 92.42% for measurement. The students' error awareness could only be judged in one category as only variables that fall into the category 'handling of material' were varied in the erroneous example, with students recognising 85% of the vicarious errors (see Table 3).

Table 2. Overview of the accuracy of students' self-assessment in the 'choice of material', 'handling of material' and 'measurement' categories.

	Accuracy of self-assessment when CVS was applied				Accuracy of self-assessment when CVS was *not* applied				
Categories	Correct application of the CVS recognised	%	Correct application of the CVS not recognised	%	Error recognised	%	Error not recognised	%	Invalid cases
Choice of material	150	98.04	3	1.96	6	33.33	12	66.67	3
Handling of material	137	78.73	37	21.26	68	42.77	91	57.23	15
Measurement	37	86.05	6	13.95	3	50.00	3	50.00	9

Note: The numbers are calculated by juxtaposing the students' self-assessment with the expert assessment. The percentages are based on the number of variables that were rated by the experts to be correctly held constant or erroneously varied over the treatments, respectively. 'Invalid cases' refers to students that either had invalid self-assessment (e.g. not using the provided table to self-assess) or to photos of experiments that could not be accurately analysed for the corresponding category of design error.

Table 3. Overview of the accuracy of students' peer assessment in the 'choice of material', 'handling of material', and 'measurement' categories.

	Accuracy of peer assessment when CVS was applied				Accuracy of peer assessment when CVS was *not* applied				
Category	Correct application of the CVS recognised	%	Correct application of the CVS not recognised	%	Error recognised	%	Error not recognised	%	Invalid cases
Choice of material	196	95.15	10	4.85	NA	NA	NA	NA	1
Handling of material	60	88.24	8	11.76	289	85	51	15	6
Measurement	61	92.42	5	7.58	NA	NA	NA	NA	3

Note: The numbers are calculated by juxtaposing the students' peer-assessment with the expert assessment. The percentages are based on the number of variables that were rated by the experts to be correctly held constant or erroneously varied over the treatments, respectively. 'Invalid cases' refer to students that had invalid peer assessments (e.g. not using the provided table to peer-assess).

Error awareness in the context of the CVS: self-generated vs. vicarious errors

When comparing the percentages of recognised design errors, the self-generated errors group recognised fewer errors than the vicarious errors group ($Mdn_{self\text{-}generated\ errors\ group} = 50$; $Mdn_{vicarious\ errors\ group} = 100$; see Figure 8). This difference is statistically significant with a large effect size, with $U = 633.5$; $Z = 6.77$; $p < .001$; $r = .60$ ($N_{self\text{-}generated\ errors\ group} = 57$; $N_{vicarious\ errors\ group} = 69$) and moderate.

To ensure that this difference in error awareness between the treatment groups was not skewed by the difference in the number of occurring design errors, students of the vicarious errors group were additionally compared with students from the self-generated

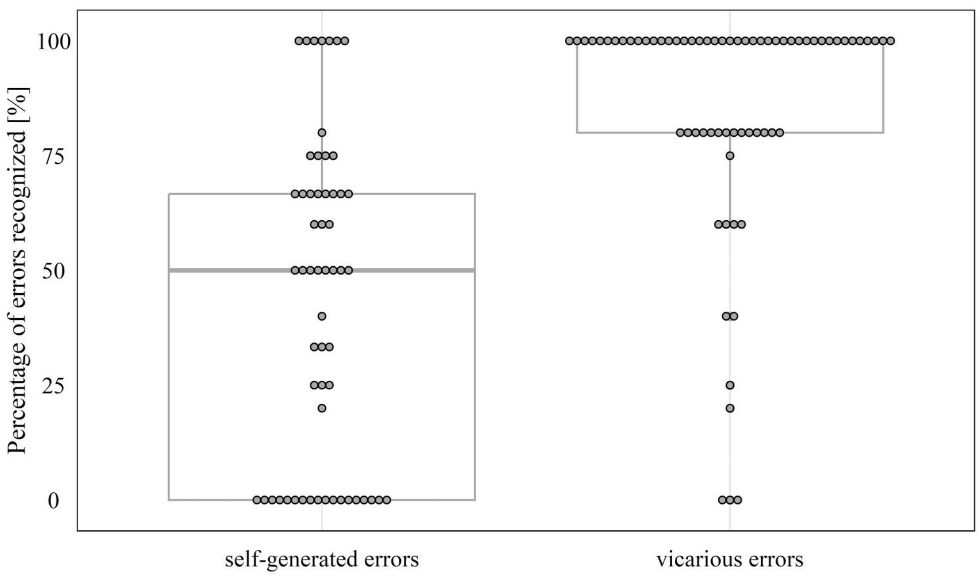

Figure 8. Errors recognised (%) in the self-generated errors group ($n = 57$) and the vicarious errors group ($n = 69$).

Table 4. Results of the Pearson's Chi-square test with Yates' continuity correction for the errors that were present in both the self-generated and the vicarious errors group.

Error	$N_{self\text{-}generated\ errors\text{-}group}$	$N_{vicarious\text{-}errors\text{-}group}$	X^2	df	*P*	Cohen's *W*
Ice quantity	11	68	23.52	1	<.001	.59
Insulator quantity	28	68	19.43	1	<.001	.48
Water level	30	68	28.61	1	<.001	.56
Inner beaker Position	15	68	13.83	1	<.001	.44
Test tube position	53	68	13.25	1	<.001	.35

Note: The number of students in the self-generated errors group varies between errors as only students that made the corresponding error were included. The number of students in the vicarious errors group remains the same for all categories as all students in this group were presented with the same deliberate errors.

errors group that made the respective errors for each of the five errors in the fictitious example. For each design error, the difference in error awareness remained large, with a moderate to large effect size, as shown in table 4.

Discussion

Occurring errors

In line with prior research (see review from Kranz et al., 2022), we identified various errors in the students' experiments. As previously stated, 17 of 144 participants were excluded from our detailed analysis. 10 students omitted crucial components like the inner beakers or water in the test tubes. Baur (2018) found similar errors. Another seven students built single-condition experiments; this was also consistent with previous studies (Baur, 2018; Chen & Klahr, 1999; Hammann et al., 2008; Schwichow et al., 2022; Siler & Klahr, 2012). As the students' worksheets included instructions on preparing two conditions, this suggests that students either did not read the instructions carefully enough or actively chose to ignore them, e.g. because they thought the outcome of a single condition experiment suffices for drawing conclusions (Siler & Klahr, 2012). This suggests considering qualitative approaches in the future, such as think-aloud approaches which could help to identify students' thought processes by talking about their decisions, thus providing immediate clarification.

Nearly all students (98.3%) who conducted the experiment encountered challenges with handling the materials, consistent with the findings of previous studies (Baur, 2018). This may be attributed to the nature of the category. 'Choice of material' includes categorical variables; thus, they are confined to a few distinct expressions. Variables associated with the 'handling of the material', such as the quantity of insulating material used by the students, are continuous and therefore not subject to such constraints. These deviations from the CVS, coupled with the occurring lack of determination of test times (Arnold et al., 2014; García-Carmona et al., 2016), suggests that students may overlook the comparability of substance amounts or laboratory material between the treatments, implying a non-scientific understanding of experimentation (Hammann et al., 2008). Alternative explanations include beliefs that additional varied variables lack causal effects, tendencies to ignore these variables, preferences for testing multiple variables simultaneously, and/or intentions to produce extreme differences between treatments (Schwichow et al., 2022). Lacking manual skills may also contribute (Baur, 2018; Schmidt, 2016). Procedural and manual skills are intertwined, and most assessments do not allow a distinction between the two (Schmidt, 2016). However, according to the Austrian curriculum for Biology (BMBWF, 2019), students should conduct simple experiments as early as grade 5, which fosters their manual skills among other objectives. We therefore assume that the occurring errors cannot only be attributed to a lack of manual skills; nevertheless, more research needs to be done.

Accuracy of the students' self- and peer assessment

Our data shows high accuracy for both self- and peer assessment, consistent with prior research in the physics education context by Schreiber (2016), Schreiber and Theyssen (2019) and a meta-analysis on peer assessment by Li et al. (2016). The overall accuracy of self- and peer assessment was similar, consistent with findings of Stefani (1994), who focused on the assessment of laboratory reports. The relatively high accuracies (58.91–94.69%) may be attributed to the specificity associated with assessing concrete examples of an experiment instead of more general skills such as experimental competence (Andrade, 2019; Brown & Harris, 2013; Li et al., 2016). Another beneficial factor might be the provision of criteria for the assessment in both treatment groups (Andrade, 2019; Panadero, Brown, et al., 2016).

Despite the relatively high overall accuracy, students struggled the most with assessing continuous variables. In a review of studies on self-assessment, Ross (2006) reported that some students simply do not understand what high performance in a task looks like and are therefore unable to assess the task accurately. This leads to the conclusion that students could benefit from even clearer criteria, such as the inclusion of decision rules that help them determine if the CVS has been applied. Involving the students in criterion development and clarification (Li et al., 2016; Panadero, Jönsson, et al., 2016) and training them in criteria application could further enhance assessment accuracy (Andrade, 2018; Panadero, Jönsson, et al., 2016; Stančić, 2021).

Error awareness in the context of the CVS: self-generated vs. vicarious errors

The difference in error awareness between self-generated and vicarious errors was found to be statistically significant, with students showing higher error awareness when assessing vicarious errors. This could be attributed to several reasons, which should be further examined in future studies. Unlike self-assessment, the assessment of vicarious errors may not require the same level of self-reflection (Tulis et al., 2016). Another explanation might be that students may be less biased when assessing vicarious errors, as they lack personal involvement in the task and are less influenced by subjective factors such as exerted effort (Ross, 2006). Additionally, the difference in error awareness could have been enhanced through the peer assessment's anonymity, as the students had to assess a fictitious student's work. Two reviews on self- and peer assessment suggest that anonymity in peer assessment eliminates possible inhibiting factors on interpersonal relationships, such as not wanting to rate peers badly or lack of self-confidence in a student's ability to rate peers (Panadero, 2016; Panadero, Jönsson, et al., 2016). However, in a meta-analysis, Li et al. (2016) suggest that students tend to take non-anonymous assessments more seriously, resulting in a more accurate assessment of their peers. Answering the question of whether anonymity helps or hinders peer assessment is beyond the scope of this study, further research is needed.

Limitations and future direction

Due to the study design, students from the vicarious errors group knew they were assessing an erroneous experiment, while students in the self-generated errors group were not aware of this. Consequently, this may have primed the vicarious errors group, meaning the information they received might have influenced their responses (Ramscar, 2016), which might have played into the difference in error awareness. Further research including a feedback loop for the self-generated errors group is needed. Providing students with individual feedback for either their own experiments or their self- and peer-assessment would have exceeded the limits of our study, in which the intervention took place on a single day and with multiple pre-service teachers teaching the lessons.

Moreover, the experiments presented to the vicarious errors group deliberately included clear, visible errors in the continuous variables, potentially enhancing error awareness. In contrast, students in the self-generated errors group may have been less strict in identifying smaller deviations from the CVS as errors. Nevertheless, the self-generated errors group exhibited low error awareness even for the categorical variables.

Notably, the students' prior experience of working with design errors, the learning outcomes regarding students' CVS-skills and subjective process variables such as cognitive load or motivation were not considered in this study.

Conclusion

Understanding scientific practices is crucial for learners, yet teaching about scientific endeavours remains challenging. Our study contributes by examining experimental design errors, revealing challenges associated with controlling variables, especially continuous ones. Students were found to be better at identifying their correct procedures than their errors. Additionally, while with some prompts, students could correctly identify some of their self-generated errors, they were more successful at identifying design errors in experiments conducted by a fictitious student. This simulation of peer assessment – through a fictitious peer – could be beneficial in science classrooms, allowing for plannable and scalable opportunities for teaching the CVS. Our results further indicate that there is potential for approaches using experimental design errors as learning opportunities, especially since the identification of errors is a prerequisite for this approach to learning, but this has yet to be explored further. As indicated in the section 'limitations and further directions', future research should explore whether incorporating a feedback loop for self-generated errors could enhance learning outcomes. Additionally, students' prior experience with design errors, their learning outcomes in terms of CVS skills and their subjective learning process variables such as cognitive load or motivation were not examined. Addressing these variables in future research could provide a more comprehensive understanding of how the diversity of CVS-related design errors can be used when teaching experimentation and if self-generated and vicarious design errors can be successfully implemented when teaching the CVS. Further research in this direction is currently being carried out by the authors.

Disclosure statement

No potential conflict of interest was reported by the authors.

Ethics statement

As per Austrian regulations, ethics committee approval was not required as the study did not involve patients, was non-invasive, and participation was entirely voluntary. Permission to conduct this study was obtained from the principals and the students' parents, as is common procedure for under-age students in Austria. The research team was guided by the 'Guidelines for Good Scientific Practice' (Austrian Agency for Research Integrity, 2016).

Data availability statement

The data supporting the findings of this study and the R-code used to analyse them are available in the Open Science Framework (OSF) at https://osf.io/pzw8u/?view_only=b05cb388087d406e9ac62b2196ca9362.

ORCID

Linda Hämmerle http://orcid.org/0000-0002-3773-4388
Andrea Möller http://orcid.org/0000-0002-2345-1299
Alexander Bergmann-Gering http://orcid.org/0000-0002-5492-8512
Theresa Krause-Wichmann http://orcid.org/0000-0003-1345-0556
Judith Lederman http://orcid.org/0000-0002-0515-9913

References

Allchin, D. (2012). Teaching the nature of science through scientific errors. *Science Education*, *95*(5), 904–926. https://doi.org/10.1002/sce.21019

Andrade, H. L. (2018). Feedback in the context of self-assessment. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge Handbook of Instructional Feedback* (pp. 376–408). Cambridge University Press.

Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education*, *4*(87), 1–13. https://doi.org/10.3389/feduc.2019.00087

Arnold, J., Kremer, K., & Mayer, J. (2014). Understanding students' experiments—What kind of support do they need in inquiry tasks? *International Journal of Science Education*, *36*(16), 2719–2749. https://doi.org/10.1080/09500693.2014.930209

Austrian Agency for Research Integrity. (2016). *Guidelines for good scientific practice*. Retrieved February 22, 2024, from https://oeawi.at/en/guidelines/

Baur, A. (2018). Fehler, Fehlkonzepte und spezifische Vorgehensweisen von Schülerinnen und Schülern beim Experimentieren [Mistakes, misconceptions, and pupils' specific approaches to experimentation]. *Zeitschrift für Didaktik der Naturwissenschaften*, *24*(1), 115–129. https://doi.org/10.1007/s40573-018-0078-7

BMBWF. (2019). *Gesamte Rechtsvorschrift für Lehrpläne - allgemeinbildende höhere Schulen, Fassung vom 10.08.2022. BGB1. II Nr. 267/2022*. [Entire legal regulation for curricula - Secondary schools, version of 10.08.2022. BGB1. II Nr. 267/2022].

BMBWF. (2024). *Gesamte Rechtsvorschrift für Lehrpläne - allgemeinbildende höhere Schulen, Fassung vom 01.08.2024. BGB1. II Nr. 239/2023*. [Entire legal regulation for curricula - Secondary schools, version of 01.08.2024. BGB1. II Nr. 239/2023].

Bond, J. B., & Ellis, A. K. (2013). The effects of metacognitive reflective assessment on fifth and sixth graders' mathematics achievement. *School Science & Mathematics*, *113*(5), 227–234. https://doi.org/10.1111/ssm.12021

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *Research on classroom assessment* (pp. 367–393). Sage.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of control of variables strategy. *Child Development*, *70*(5), 1098–1120. https://doi.org/10.1111/1467-8624.00081

Chinn, C. A., & Colan Duncan, R. (2018). What is the value of general knowledge of scientific reasoning? In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation: The roles of domain-specific and domain-general knowledge* (pp. 77–101). Routledge.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Dasgupta, A. P., Anderson, T. R., & Pelaez, N. J. (2016). Development of the neuron assessment for measuring biology students' use of experimental design concepts and representations. *CBE Life Science Education*, *15*(2), 1–21. https://doi.org/10.1187/cbe.15-03-0077

Duit, R., & Tesch, M. (2010). *On the role of the experiment in science teaching and learning - Visions and the reality of instructional practice*. 7th International Conference on Hands-on Science, Rethymno, Greece, Juy 25th–31th 2010, University of Crete.

García-Carmona, A., Criado, A. M., & Cruz-Guzmán, M. (2016). Primary pre-service teachers' skills in planning a guided scientific inquiry. *Research in Science Education*, *47*(5), 989–1010. https://doi.org/10.1007/s11165-016-9536-8

Gartmeier, M., Bauer, J., Gruber, H., & Heid, H. (2008). Negative knowledge: Understanding professional learning and expertise. *Vocations and Learning*, *1*(2), 87–103. https://doi.org/10.1007/s12186-008-9006-1

Hammann, M., Phan, T. T. H., Ehmer, M., & Grimm, T. (2008). Assessing pupils' skills in experimentation. *Journal of Biological Education*, *42*(2), 66–72. https://doi.org/10.1080/00219266.2008.9656113

Kapur, M. (2014). Comparing learning from productive failure and vicarious failure. *Journal of the Learning Sciences*, *23*(4), 651–677. https://doi.org/10.1080/10508406.2013.819000

KMK. (2005). *Beschlüsse der Kultusministerkonferenz Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss*. [Decisions of the assembly of German ministers of education—Educational standards in biology for lower secondary school]. Luchterhand.

Kranz, J., Baur, A., & Möller, A. (2022). Learners' challenges in understanding and performing experiments: A systematic review of the literature. *Studies in Science Education*, *59*(2), 321–367. https://doi.org/10.1080/03057267.2022.2138151

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education*, *45*(2), 193–211. https://doi.org/10.1080/02602938.2019.1620679

Li, H., Xiong, Y., Zhang, X., Kornhaber, M. L., Lyu, Y., Chung, K. S., & Sunen, H. K. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education*, *41*(2), 245–264. https://doi.org/10.1080/02602938.2014.999746

Loibl, K., & Leuders, T. (2019). How to make failure productive: Fostering learning from errors through elaboration prompts. *Learning and Instruction*, *62*, 1–24. https://doi.org/10.1016/j.learninstruc.2019.03.002

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review*, *29*(4), 693–715. https://doi.org/10.1007/s10648-016-9379-x

Lui, H. (2021). *Lernen aus Fehlern beim Experimentieren im naturwissenschaftlichen Unterricht mit Fokus auf der Variablenkontrollstrategie* [Learning experimentation in science education from errors, with a focus on the control-of-variables strategy]. [Unpublished master's thesis]. Saarland University.

Morris, B. J., Cason, J., Asaro, K., Zhang, Y., Rivers, M., Owens, W., & Dunlosky, J. (2024). Learning the control-of-variables strateyg during an informal science lesson on popping popcorn. *International Journal of Science Education*, *47*(1), 45–64. https://doi.org/10.1080/09500693.2024.2310526

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press. https://doi.org/10.17226/13165

National Research Council. (2013). *Next generation science standards: For states, by states*. Retrieved February 20, 2024, from https://www.nextgenscience.org/get-to-know

OECD. (2007). *PISA 2006. Science competencies for tomorrow's world. Volume 1: Analysis*. OECD Publishing. https://doi.org/10.1787/19963777

OECD. (2018). *The science of teaching science: An exploration of science teaching practices in PISA 2015*.

Oser, F., Hascher, T., & Spychiger, M. (1999). Lernen aus Fehlern. Zur Psychologie des "negativen" Wissens [Learning from errors. On the psychology of "negative" knowledge.]. In W. Althof (Ed.), *Fehlerwelten. Vom Fehlermachen und Lernen aus Fehlern. Beiträge und Nachträge zu einem interdisziplinären Symposium aus Anlass des 60. Geburtstags von Fritz Oser* (pp. 11–41). Leske + Budrich.

Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: A review and future directions. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of social and human conditions in assessment* (pp. 247–266). Routledge.

Panadero, E., & Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assessment & Evaluation in Higher Education*, *44*(8), 1253–1278. https://doi.org/10.1080/02602938.2019.1600186

Panadero, E., Brown, G. T. L., & Strijbos, J.-W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review*, *28*(4), 803–830. https://doi.org/10.1007/s10648-015-9350-2

Panadero, E., Jönsson, A., & Strijbos, J.-W. (2016). Scaffolding self-regulated learning through self-assessment and peer assessment: Guidelines for classroom implementation. In D. Laveault & L. Allal (Eds.), *Assessment for learning: Meeting the challenge of implementation* (Vol. 4, pp. 311–326). Springer.

Ramscar, M. (2016). Learning and the replicability of priming effects. *Current Opinion in Psychology*, *12*, 80–84. https://doi.org/10.1016/j.copsyc.2016.07.001

Reinholz, D. (2016). The assessment cycle: A model for learning through peer assessment. *Assessment & Evaluation in Higher Education*, *41*(2), 301–315. https://doi.org/10.1080/02602938.2015.1008982

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research & Evaluation*, *11*(10), 1–13.

RStudio Team. (2020). *RStudio: Integrated development for R*. RStudio. http://www.rstudio.com/

Safadi, R. (2022). Supporting student learning from diagnosing erroneous examples when contrasting them with worked examples in the physics classroom. *International Journal of Science Education*, *44*(2), 245–270. https://doi.org/10.1080/09500693.2021.2023834

Schmidt, D. (2016). *Modellierung experimenteller Kompetenzen sowie ihre Diagnostik und Förderung im Biologieunterricht* [Modeling of experimental competences and their diagnostics and promotion in biology classes]. Logos.

Scholtes, K. (2021). *Lernen aus Fehlern beim Forschenden Experimentieren im Chemieunterricht* [Learning from errors in inquiry learning in chemistry education] [Unpublished master's thesis]. Saarland University.

Schreiber, N. (2016). Wie genau beurteilen Schülerinnen und Schüler ihre eigenen experimentellen Fähigkeiten? [How accurate can students assess their own experimental skills?]. *Physik und Didaktik in Schule und Hochschule*, *1*(15), 49–63.

Schreiber, N., & Gut, C. (2022). Die Validierungspraxis bei hands-on Experimentiertests in der naturwissenschaftsdidaktischen Forschung [The validation practice of performance assessments using practical tasks in science education research]. *Zeitschrift für Didaktik der Naturwissenschaften*, *28*(8), 1–22. https://doi.org/10.1007/s40573-022-00145-6

Schreiber, N., & Theyssen, H. (2019). Selbstbeurteilungen zur formativen Diagnostik experimenteller Performanzen - Was zeichnet genau urteilende Schülerinnen und Schüler aus? [Self-assessments for the formative diagnostics of experimental performances—What characterizes pupils with a high self-assessment accuracy?]. *Zeitschrift für Didaktik der Naturwissenschaften*, *25*(1), 43–61. https://doi.org/10.1007/s40573-019-00091-w

Schreier, M. (2012). *Qualitative content analysis in practice*. Sage Publications.

Schwichow, M., Brandenburger, M., & Wilbers, J. (2022). Analysis of experimental design errors in elementary school: How do students identify, interpret, and justify controlled and confounded experiments? *International Journal of Science Education*, *44*(1), 91–114. https://doi.org/10.1080/09500693.2021.2015544

Siegler, R. S., & Liebert, R. M. (1975). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Designing a factorial experiment. *Developmental Psychology*, *11*(3), 401–402. https://doi.org/10.1037/h0076579

Siler, S. A., & Klahr, D. (2012). Detecting, classifying, and remediating. Children's explicit and implicit misconceptions about experimental design. In R. W. Proctor & E. J. Capaldi (Eds.), *Psychology of science: Implicit and explicit processes* (pp. 137–180). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199753628.003.0007

Sinha, T., & Kapur, M. (2019). *When productive failure fails*. 41st Annual Conference of the Cognitive Science, July 24th–27th 2019, Montreal, Canada.

Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research*, *91*(5), 761–798. https://doi.org/10.3102/00346543211019105

Stančić, M. (2021). Peer assessment as a learning and self-assessment tool: A look inside the black box. *Assessment & Evaluation in Higher Education*, *46*(6), 852–864. https://doi.org/10.1080/02602938.2020.1828267

Stefani, L. A. J. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education*, *19*(1), 69–75. https://doi.org/10.1080/03075079412331382153

Topping, K. J. (2017). Peer assessment: Learning by judging and discussing the work of other learners. *Interdisciplinary Education and Psychology*, *1*(1), 1–17. https://doi.org/10.31532/InterdiscipEducPsychol.1.1.007

Tulis, M., Steuer, G., & Dresel, M. (2016). Learning from errors: A model of individual processes. *Frontline Learning Research*, *4*(2), 12–26. https://doi.org/10.14786/flr.v4i2.168

VERBI Software. (2021). *MAXQDA 2022* [Computer software]. https://www.maxqda.com/

Zhang, Q., & Fiorella, L. (2023). An integrated model of learning from errors. *Educational Psychologist*, *58*(1), 18–34. https://doi.org/10.1080/00461520.2022.2149525