Instructional Science
https://doi.org/10.1007/s11251-023-09633-9

ORIGINAL RESEARCH

Open science in the classroom: students designing and peer reviewing studies in human brain and behavior research

Camillia Matuk[1] · Lucy Yetman-Michaelson[1] · Rebecca Martin[2] · Veena Vasudevan[3] · Kim Burgas[4] · Ido Davidesco[5] · Yury Shevchenko[6] · Kim Chaloner[7] · Suzanne Dikker[1]

Received: 11 November 2021 / Accepted: 17 April 2023

Abstract

Citizen science programs offer opportunities for K-12 students to engage in authentic science inquiry. However, these programs often fall short of including learners as agents in the entire process, and thus contrast with the growing open science movement within scientific communities. Notably, study ideation and peer review, which are central to the making of science, are typically reserved for professional scientists. This study describes the implementation of an open science curriculum that engages high school students in a full cycle of scientific inquiry. We explored the focus and quality of students' study designs and peer reviews, and their perceptions of open science based on their participation in the program. Specifically, we implemented a human brain and behavior citizen science unit in 6 classrooms across 3 high schools. After learning about open science and citizen science, students (N = 104) participated in scientist-initiated research studies, and then collaboratively proposed their own studies to investigate personally interesting questions about human behavior and the brain. Students then peer reviewed proposals of students from other schools. Based on a qualitative and quantitative analysis of students' artifacts created in-unit and on a pre and posttest, we describe their interests, abilities, and self-reported experiences with study design and peer review. Our findings suggest that participation in open science in a human brain and behavior research context can engage students with critical aspects of experiment design, as well as with issues that are unique to human subjects research, such as research ethics. Meanwhile, the quality of students' study designs and reviews changed in notable, but mixed, ways: While students improved in justifying the importance of research studies, they did not improve in their abilities to align methods to their research questions. In terms of peer review, students generally reported that their peers' feedback was helpful, but our analysis showed that student reviewers struggled to articulate concrete recommendations for improvement. In light of these findings, we discuss the need for curricula that support the development of research and review abilities by building on students' interests, while also guiding students in transferring these abilities across a range of research foci.

Keywords Citizen science · Open science · Classrooms · Scientific literacy · Experiment design · Peer review

Extended author information available on the last page of the article

Published online: 13 July 2023

Introduction

Citizen science programs can promote public scientific literacy by opening the process of inquiry to the public. In many cases, however, it is scientists who define which questions to answer and how, while public participation is limited to either contributing or analyzing data (Phillips et al., 2018). Less often is the public invited to conceptualize and review studies. Thus, many citizen science programs may miss opportunities for the public to develop scientific literacy through the pursuit of questions that interest them, and by experiencing community-driven processes that can generate scientific knowledge.

The present research explores how a classroom-based, open science approach to citizen science can engage high school students in the parts of inquiry that tend to be reserved for scientists: Study conceptualization and peer review. Our research took place during the early stages of developing a new citizen science platform and curriculum, which guided students in generating research questions and conceptualizing study designs to be later vetted by scientists, and opened for data collection via public participation. At the time of this writing, some of the student-led studies had accrued hundreds of study participants (Dikker et al., 2022). By describing the study design and review abilities observed of students during this process, this research contributes to informing how citizen science can offer students authentic inquiry learning experiences that are otherwise difficult to achieve in classroom contexts, and that also help to increase participants' scientific literacy.

Participatory science learning: developing scientific literacy through open science

Our work is grounded in participatory science learning. This sociocultural perspective views student learning as a process of identification with a community, strengthened through a focus on authentic problems, on practices for negotiating understanding, and on the roles of more knowledgeable colleagues (Barab & Hay, 2001; Gee, 2003; Koomen et al., 2018; Lave & Wenger, 1991; National Research Council et al., 2012; NGSS Lead States, 2013). Participatory science learning reflects the social nature of authentic science, which is notably emphasized in citizen science.

While expert definitions of citizen science vary depending on geography, scientific discipline, and context (Haklay et al., 2021b), the term broadly describes a spectrum of approaches for engaging the public in the process of science, from contributing data, to analyzing data, to partnering non-experts with experts to conceptualize and carry out a scientific project (Bonney et al., 2009; Haklay et al., 2021a; Haklay et al., 2021b). Citizen science is part of a broader movement called *open science*, a collective effort to make the process of scientific inquiry more transparent and accessible to the public (Fecher & Friesike, 2014). The six tenets of open science (Fecher & Friesike, 2014; Leible et al., 2019) are: Make knowledge freely available (Democratic), make the science process more efficient and goal-oriented (Pragmatic), make science accessible to everyone (Public), create and maintain tools and services (Infrastructure), measure the scientific impact of research (Measurement), and support community inclusion and commitment (Community). The importance of open science became especially apparent during the current pandemic, as scientists worldwide coordinated efforts to treat and prevent COVID-19 by prioritizing research foci, and sharing tools and early findings with one another and with the public in real time (Fry et al., 2020; Lee & Campbell, 2020; Rempel, 2020).

Here, we focus specifically on experimental design and peer review. These two keystones of the scientific process (Scott, 2007) gain further importance in an open science framework as they reflect the social and participatory nature of science, and the role of a community in generating scientific knowledge. In a typical peer review process, a journal editor invites 2–3 reviewers with relevant expertise to offer written critiques of research. Review criteria include the significance of the research questions, groundedness in existing work, soundness of the methodology, and validity of the findings (including attention to limitations, potential confounds, and alternative explanations) (Kelly et al., 2014). Ideally, reviewers also offer constructive recommendations for revisions that will ensure the rigor and relevance of the research. Various conventions ensure high quality reviews. For example, journals may follow an open, single blind, or double blind review process. Journals also tend to solicit independent reviews, whereas grant agencies tend to convene panels of reviewers, who coalesce their evaluations to produce joint reviews. While these conventions differ by field, journal, and agency, they share the goal of ensuring helpful feedback by mitigating reviewers' potential bias, and enabling them to be critical without worry of retaliation (Kelly et al., 2014).

In an open science model, peer review occurs not just on completed research, but also on proposed research. Inviting peer reviewers early into the process of science introduces multiple perspectives that can improve research, and places reviewers in a position to make more substantial contributions to the direction of research. More specifically, open science advocates encourage researchers to pre-register their research protocols with a journal, i.e., to commit to an analysis plan *before* data is collected (Ledgerwood, 2018; Nosek et al., 2018; Open Science Collaboration, 2015; Van't Veer & Giner-Sorolla, 2016). If, upon peer review, the research questions are deemed important and the methods for answering them suitable, the journal will publish the paper regardless of the outcome. At the time of this writing, more than 80 science journals offer registered reports. In contrast to typical publishing models, an open science model, in which scientists make their study plans public, encourages more rapid scientific advancements, reduces publication biases, and increases the reproducibility of research findings to cope with the replication crisis (Maxwell et al., 2015; Open Science Collaboration, 2012).

Promoting scientific literacy through authentic inquiry learning experiences

Exposure to open science, including citizen science, offers authentic inquiry learning experiences that have the potential to impact how students understand and value science. For example, participation in citizen science has been found to positively impact middle and high school students' understanding of the nature of science inquiry (Crawford, 2012), and their attitudes toward STEM (Meyer et al., 2014). Moreover, students who participate in study conceptualization and peer review gain a more accurate picture of the social nature of science, and of the roles that they could have in generating scientific knowledge (Harris et al., 2020; Robnett et al., 2015). Such roles can furthermore highlight the personal and social relevance of science: Rather than replicate classic experiments for which the outcomes are already known—a typical approach to classroom-based science (Furtak & Penuel, 2019; Linn et al., 2016)—students are given agency to ask and answer open-ended questions that are relevant to them and their communities. Such student-driven inquiry approaches can be motivating for learners, and further allow them to experience and appreciate the satisfaction of overcoming personally meaningful challenges (Buchanan, 2019).

Early science inquiry experiences can shape the positive values and attitudes toward science that are critical for developing *scientific literacy*, a major goal of science education (American Association for the Advancement of Science, 1989, 2011; National Research Council, 1996; National Science Foundation, 1996; Osborne, 2010), as well as a goal of citizen science efforts (Brossard et al., 2005; Jordan et al., 2011; Roche et al., 2020; Saunders et al., 2018). As defined by the National Research Council (American Association for the Advancement of Science, 1989, 2011; National Research Council, 1996; National Science Foundation, 1996; Osborne, 2010), a person who is scientifically literate understands and can apply scientific processes and concepts to participate effectively in both everyday and high stakes decision making. Scientific literacy includes the ability to formulate and determine the answers to questions; to describe, explain, and predict observations; and to generate and evaluate scientific information and arguments in terms of the quality of evidence, and the methods that generated it.

Thus, scientific literacy encompasses abilities in *experimental design* and *peer review*, which each require understanding and applying concepts to design and evaluate the processes by which scientific knowledge is generated.

Students' experimental design abilities

What makes an effective experimental design

Robust experimental designs have several characteristics (Walters, 2020): They are driven by testable questions or hypotheses that specify relationships between variables; they distinguish independent and dependent variables, aligned with the question or hypothesis; they use valid and reliable means to operationalize variables; they control for other variables to allow comparison and causal inference; they select and sample populations with attention to representativeness and generalizability; and they draw conclusions based in evidence from findings, with attention to limitations and possible alternative explanations.

There are various challenges to building students' experimental design skills in classroom contexts, from elementary to university levels (Roche et al., 2020). Cognitive barriers include challenges with understanding and applying concepts and skills related to the discipline, and to experimental design; and with having a sufficient grasp of the field to conceptualize useful scientific contributions. Affective barriers include students' lack of motivation to learn research methods due to anxiety over the perceived difficulty, or their failure to recognize its personal relevance (Earley, 2014). Logistic barriers include limited access to equipment, human subjects, and time (Perlman & McCann, 2005; Woolley et al., 2018), as the length of a class period and the structure of traditional school curricula are typically not conducive to student-led inquiry (Fitzgerald et al., 2019). Cultural barriers include the challenge of cultivating open science practices in classroom environments, in which standards for assessment tend to encourage independent work, and to counter the tenets of open science. Importantly, teachers, who are critical mediators of students' citizen science participation, may also lack the confidence and preparation to support their students' learning (Fitzgerald et al., 2019; Jenkins et al., 2015; Kelemen-Finan & Dedova, 2014).

Supporting students in learning experimental design

Given these barriers, it is not surprising that students struggle with experimental design. For example they have difficulty formulating answerable research questions; identifying

and operationalizing variables; designing methods aligned with research questions; and interpreting data with attention to limitations, potential confounds, and alternative explanations (Chen & Klahr, 1999; Dasgupta et al., 2014; Fuller, 2002; Kuhn & Dean, 2005; Shi et al., 2011; Woolley et al., 2018). Both middle and high school students have difficulty identifying and manipulating variables (Bullock et al., 1999; Burns et al., 1985; Dolan & Grady, 2010; Fuller, 2002), and controlling for variability in their experimental designs (Chen & Klahr, 1999; Kuhn & Dean, 2005). Elementary students tend to find it easier to identify faults in others' experimental controls than to design those experimental controls themselves (Bullock et al., 1999). These challenges are not limited to K-12 classrooms: undergraduate students also struggle to identify which variables to measure to address a hypothesis (Salangam, 2007), to align treatment and outcome variables in articulating hypotheses (Beck & Blumer, 2012; D'Costa & Schlueter, 2013; Griffith, 2007; Harker, 2009; Libarkin & Ording, 2012; Salangam, 2007), and to draw conclusions based in evidence from an experiment's findings (Dolan & Grady, 2010; Harker, 2009; Hiebert, 2007; Tobin & Capie, 1982).

Prior research has found that students in higher education contexts develop abilities in experimental design through authentic and meaningful scientific inquiry (e.g., Killpack et al., 2020; Sujarittham et al., 2019). Yet, few such experiences are offered at the pre-college level (Deemer et al., 2021), and those that do exist tend to focus on the physical and life sciences (e.g., (e.g., Etkina et al., 2003; Harris et al., 2020; Robnett et al., 2015). Thus, less is known about students' experimental design and peer review abilities in psychological sciences, in spite of the increasing relevance of this domain at the pre-college level, and to various career paths (Iandoli & Shen, 2021; Kadir & Broberg, 2021; Shneiderman, 2020).

Students' peer review abilities

The value of peer review

In contrast to experimental design, peer review tends to be more familiar to both students and teachers as a standard formative assessment practice in educational settings. Also known as peer assessment or peer feedback, classroom-based peer review involves students exchanging feedback on one another's work. Although peer review is a key part of scientific inquiry, it has mostly been studied in language arts classrooms and in higher education settings (Double et al., 2020), or else in the context of students learning physical sciences such as astronomy, and life sciences such as ecology (Anker-Hansen & Andrée, 2019; Ketonen et al., 2020; Tsivitanidou et al., 2012).

These studies on peer review have shown the multiple benefits of this assessment practice for both students and teachers across educational levels (Black & Wiliam, 1998; Dolezal et al., 2018; Double et al., 2020; Dysthe et al., 2010; Ed et al., 2001; Hattie & Timperley, 2007; Kennedy et al., 2008; Li et al., 2020; Loretto et al., 2016; Noonan & Randy Duncan, 2005; Shute, 2008; Sluijsmans, 2002; Tsai et al., 2002; Tsivitanidou et al., 2011; van Gennip et al., 2010; Wanner & Palmer, 2018; Xiao & Lucking, 2008). Specifically, teachers benefit from peer review in that the burden of assessment shifts from them to their students (Wanner & Palmer, 2018). Meanwhile, the benefits of peer review for students are bi-directional: Students' work benefits from peers' recommendations, and reviewers also benefit by giving reviews (Huisman et al., 2018). The process of reviewing moreover promotes reflection as students compare their own work to that of their classmates

(Panadero, 2016; Race, 2014). In doing so, student reviewers can learn what constitutes high quality work (Kollar & Fischer, 2010), and then apply these criteria to their own work. Peer review implemented in a classroom context can ensure that students receive timely feedback; feedback from multiple perspectives; and sometimes, more useful feedback, as students tend to find feedback from peers to be more understandable than feedback from teachers (Black & Wiliam, 1998). Importantly, peer review can heighten students' sense of autonomy by lessening students' dependence on their teacher (Shen et al., 2020). In line with the ideals of participatory learning, peer review can empower students by foregrounding their expertise (Sackstein, 2017). As evidence of these outcomes, studies show that peer assessment results in better academic performance compared to both teacher assessment and no assessment, and similar outcomes compared to self-assessment (Dolezal et al., 2018; Huisman et al., 2018; Li et al., 2020; Loretto et al., 2016; Noonan & Randy Duncan, 2005; Tsai et al., 2002; Tsivitanidou et al., 2011).

What makes an effective peer review

In science education contexts, high quality peer reviews can be defined as justified evaluations of the strengths and weaknesses of peers' work, with specific suggestions for improvement. To effectively review a research study requires a particular set of abilities, understandings, and dispositions (Gaynor, 2020), including sufficient conceptual understanding of experimental design such that issues can be identified and paired with concrete solutions; as well as an ability to communicate these critiques in ways that will be helpful to the recipient, that is, by offering clear explanations for a critique, with concrete recommendations for improvement.

A meta-analysis of the impact of peer feedback (Hattie & Timperley, 2007) showed that the most impactful feedback provided specific recommendations for better performing a task. Less impactful feedback focuses only on whether goals were achieved, and the least impactful feedback merely praises or criticizes work. Prior research found that high school students were more likely to understand peer feedback that recommended solutions for improvement; more likely to agree with feedback that softened criticism with accompanying praise; and more likely to implement feedback that was understood, with which they agreed, that offered explanation, and that expressed uncertainty (e.g., it is possible that… maybe you should…) (Wu & Schunn, 2020a).

Supporting students' in learning to peer review

As with experimental design, peer review presents a number of cognitive and affective challenges to students. Research finds that students contend with a complex set of social influences on how reviews are given and received, including trust, motivation, and comfort with one another (e.g., Kasch et al., 2021; Kaufman & Schunn, 2011; Panadero, 2016; van Gennip et al., 2010; Zou et al., 2018). For instance, student reviewers tend to be less concerned about the accuracy of their reviews, and more about social perceptions of causing embarrassment or of being unkind (Christianakis, 2010; Peterson, 2003).

Students' perceptions of peer review are also mixed. Some research finds that students tend to mistrust the expertise of their peers (Kaufman & Schunn, 2011), to view teacher feedback to be more helpful (Tsui & Ng, 2000), and to be more likely to apply teacher feedback as opposed to peer feedback in revising their work (Yang et al., 2006). At the same time, mistrust of peer feedback can make students more mindful recipients of critique

(Bangert-Drowns et al., 1991), which is related to their greater independence and autonomy as learners (Yang et al., 2006).

Other research suggests that it is the content of feedback rather than students' perceptions of it that impacts whether students apply it to their revisions (Gielen et al., 2010; Kaufman & Schunn, 2011; Strijbos et al., 2010). This indicates the importance of students learning to produce effective peer reviews, and need to intentionally support them in doing so (Gielen et al., 2010; Hovardas et al., 2014; Lu & Law, 2012; Topping, 2009; van Zundert et al., 2010). Intentional guidance in peer review can motivate the reasons for conducting peer review; increase appreciation for the value of peer review; generate a collective understanding of the characteristics of helpful reviews, and help to make peer assessment most impactful (Li, 2017; Li & Grion, 2019; Schunn et al., 2016; Tasker & Herrenkohl, 2016). Notably, incorporating conventions of professional peer review into classroom contexts appear to help. In one study among university students, for example, anonymous peer reviews helped peer reviewers to focus on the content of their peers' work rather than on guessing their identities (Yu & Sung, 2016); increased the value perception of peer assessment; and also encouraged more critical feedback (Howard et al., 2010; Ketonen et al., 2020; Panadero & Alqassab, 2019).

Research questions and rationale

Experimental design and peer review encompass many of the skills and concepts described by scientific literacy. As well, they are each key, interrelated processes in science inquiry, with potential to mutually benefit one another: As peer reviewers apply their own experimental design expertise to evaluating others' work, they may also learn new techniques and insights to apply to their own experimental designs. Student-driven citizen science inquiry offers particularly rich opportunities for developing scientific literacy because it allows students to pursue research of interest to them. Yet, this same advantage also raises questions regarding how to build on students' personal interests, while also guiding their abilities to articulate, study, and evaluate research of broader importance, both related to, and outside of their personal interests, and in which they may lack disciplinary content knowledge.

To better understand the learning potential in open science, we explore the interests, experiences, and development of students' study design and peer review abilities, through their participation in a citizen science platform and curriculum on human brain and behavior research. Specifically, we ask:

1. What kinds of studies do students design?
    a. What are the foci of students' research questions?
    b. What dimensions of quality characterize their study designs?
    c. In what ways do their study design abilities change from pre to post?
2. What kinds of peer reviews do students generate?
    a. What dimensions of quality characterize students' peer reviews?
    b. In what ways do students' peer review abilities change from pre to post?
3. What are students' experiences participating in open science? That is, how do they perceive the value and process of designing studies, and participating in peer review?

For RQ1b and RQ1c, and RQ2a and RQ2b, we hypothesize that students will improve over the course of the unit in certain aspects of study design and peer review. However, given the complexity of these skills, and the time necessary to master them, we also hypothesize that students will continue to struggle with certain other aspects of study design and peer review. By understanding the focus and quality of students' study designs and peer reviews (RQ1 and RQ2), and describing students' perceptions of open science (RQ3), we aim to inform the design of classroom-based curriculum in citizen science, and in open science contexts more broadly.

Methods

To explore students' interests, experiences, and scientific literacies in an open science context, we created and implemented a high school human brain and behavior unit, supported by a citizen science platform called *MindHive* (Dikker et al., 2022). The platform has a Discover area where students can explore and participate in studies created by scientists and other students; and a Develop area where they can create studies, collect data, and exchange peer reviews on studies created by students within their class and at other schools. Teachers can create and invite students to a class, view their students' platform activity, and create assignments using the My Classes page. To date, MindHive has been used by ~ 350 students across 9 schools, and hosts ~ 300 studies created by both students and professional scientists.

Through quantitative and qualitative analyses of student-generated artifacts, survey responses, and interviews, we describe changes in specific dimensions of students' study designs and peer reviews, and their perceptions of the value of an open science approach.

Participants and context

Three high schools in three states of the United States participated in this study during Spring 2021 (Table 1). The schools, referred to here by pseudonyms, were Gordonia, a private school in a large northeastern city; Myrtle, a private boys high school in a small southern city; and Redberry, a large public science magnet school in a small mideastern city. There was a total of 104 students across 6 class periods (1 period in each of Gordonia and Myrtle, 4 class periods in Redberry).

The three teacher participants had 8–20 + years of K-12 teaching experience, and had previously used *MindHive* in their classrooms. The teachers also had prior formal training and professional backgrounds in science, with one teacher (Gordonia) having studied environmental science education, one (Myrtle) having previously been a cancer and immunology scientist, and the other (Redberry) having a PhD in biochemistry, and a previous career in biotechnology. Each of the teachers valued, and often provided students with hands-on student-led inquiry learning experiences, and incorporated various forms of peer assessment into their instruction.

Students in this study were using MindHive for the first time. Their own prior inquiry learning experiences were in line with the standards followed by U.S. high schools. These experiences included completing lab reports that involved developing inquiries into hypotheses, identifying variables, analyzing data, and writing evidence-based conclusions. In early grades, these projects are highly scaffolded. In later grades, they become increasingly student-driven in preparation for the variety of assessments that

Table 1 Summary of participants and data

School	Subject, grade, teacher(s)	┼Students (N = 104)	Data
Gordonia • Private urban school • 32.9% nonwhite, 7.2% Hispanic • 29% on financial aid	• Environmental Science, Grades 11–12 • 1 teacher (female), 20 + yrs teaching experience	• 1 class period • 15 students • Grades 11–12	4 interviews 6 proposals 30 peer reviews
Myrtle • Private suburban school • 24% nonwhite, 1% Hispanic • 29% on financial aid	• Molecular Biology, Grades 10–11 • 1 teacher, female, 11 yrs teaching experience • Previous career as a cancer and immunology scientist	• 1 class period • 9 students • Grades 10–12	4 interviews 3 proposals 33 peer reviews
*Redberry • Public magnet school • 71% nonwhite, 23% Hispanic • 15% free/reduced lunch	• Biology, Grade 9 • 1 teacher (female), 9 yrs teaching experience, prior experience in pediatrics research • 1 teacher (male), PhD in biochemistry, previous career in biotechnology	• 4 class periods • 80 students • Grade 9	21 proposals 78 peer reviews

*Due to IRB restrictions, we were only able to collect de-identified student artifacts at this school

┼Because only half or so of the students chose to disclose their gender identities, these data are excluded from the table

students may encounter, such as college board testing, International Baccalaureate (IB) lab requirements, or advanced level independent research projects. Thus, for students in this study, their participation in MindHive was a first inquiry experience with such a level of student-directedness.

Unit description and activity flow

All classroom activities were guided by MindHive (www.mindive.science), an online citizen science platform and accompanying curriculum that supports teachers, students, and scientists in conducting real-world human brain and behavior research (Dikker etal. 2022). MindHive is designed following open science principles, and intended to support citizen science inquiry in the field of human cognition and neuroscience (hereafter referred to as brain and behavior research). Human brain and behavior research focuses on questions of immediate personal relevance to students. As such, it offers unique opportunities for learners to grapple with important research issues, such as ethics, inclusion, and participant experience; the importance of context in interpreting findings; and the relation of all these to the validity of evidence. Moreover, the cross-cutting nature of concepts in brain and behavior research allows it to easily complement many traditional STEM courses, including those with foci on environmental science, molecular biology, and research methods; to attract broader audiences to STEM; and to prepare students for careers in human–computer interaction, engineering, and user experience, which each have foundations in human brain and behavior research (Iandoli & Shen, 2021; Kadir & Broberg, 2021; Shneiderman, 2020).

MindHive invites students and scientists into an open science community, in which they can generate and review studies on human brain and behavior. Vetted studies from both students and scientists are contributed to a bank of studies available for public participation. Members can use these studies as models, and remix them by adapting instruments and experiment design patterns. In the classroom implementations that we support, scientists also join class sessions to serve as mentors during students' teamwork. Previous research found MindHive to be successful in promoting aspects of students' science learning activation, including their fascination with science, and their agency as citizen scientists (Matuk et al., 2021).

In this study, students completed the unit's 14 lessons in 1–4 class periods per week over 6–9 weeks (Table 2), during which they worked in small groups to conceptualize and propose study designs, and then to peer review one another's proposals. Below, we describe the activities designed to support students' study designs and peer reviews.

Students' study design process

Students' study design process was guided in multiple ways. First, they were introduced to the fundamentals of experimental design through interactive lectures and small and whole-group discussion activities. Second, they were exposed to models of experimental designs by participating in and reviewing scientist-designed MindHive studies. Four scientist mentors—neuroscience students enrolled in PhD programs across the United States, and whom we recruited from our network of neuroscientist colleagues—joined some of these class sessions (~ 5 class sessions/implementation, per school) to share their work, and to provide feedback on students' ideas. One mentor also designed a study featured on the MindHive platform that we paired with a lesson on mindfulness and the brain.

Table 2 Overview of lessons in the MindHive unit

Lesson number and topic	Assignments/homework
0 MindHive curriculum overview & onboarding	• Pretest survey
1A Making science—asking questions • Students learn about the ingredients of a good research question. What makes a question relevant, generalizable, and testable? • Students begin by exploring how they encounter everyday questions in science, and start to develop language to describe the nature of scientific inquiries	• Journal entry: reflect on how the changes in our everyday lives due to the pandemic have affected you. Did you learn anything about yourself or your friends? • Read a synthesized paper about human brain & behavior research during Covid-19 for next class
1B Making science-process & communication • Students learn how scientists communicate with each other and with the public at different stages of the research process. How can the scientific community strike the appropriate balance between rapid discovery and scientific rigor?	• Watch a short video on citizen science using some guiding questions for next class
2A Involving the public-citizen science • Students learn about different models of citizen science and discuss the value and possible limitations of scientist-public partnerships	• Watch Washington Post Journal video: "How China Is Using Artificial Intelligence in Classrooms" (youtube.com/watch?v = JMLsHI8aV0g&t = 13 s)
2B Involving the public—human subjects • Students learn about benefits and pitfalls related to conducting science on human research subjects through examples from the past and present • Students experience and reflect on what it is like to be a human subject, and engage in class discussions about how science and society should approach data from human subjects	• Participate in a MindHive Study and reflect on the participant experience
3A Brain & behavior research: making science • Students learn about basic human neuroscience concepts and the tools used by neuroscience and psychologists to understand how our brains support and explain our behavior	
3B Brain & behavior case studies—risk taking • Students learn about dopamine as it relates to age, risk taking, and mood	• Participate in and reflect on the MindHive Risk-taking Study
3C Brain & behavior case studies—social influence • Students learn about the social brain and how empathy and social influence can explain human behavior	• Participate in and reflect on the MindHive Climate Choices Study
3D Brain & behavior case studies—mindfulness • Students learn how different parts of the human brain map onto different brain functions through different types of mindfulness meditation	• Participate in and reflect on the MindHive Mindfulness Study
4A Developing your research: coming up with a research topic • Students revisit Lesson 1A (what makes for a good research question?) and reflect on the MindHive studies they've participated in thus far	• Create MindHive Study Workspace • Fill out Brainstorm Proposal Cards
4B Align your research question with your task • Students learn how to translate their research question into testable hypotheses and create an appropriate study design	

Table 2 (continued)

Lesson number and topic	Assignments/homework
4C Developing your proposal: Background research • Students learn how to do background research on their study topics and explore the capabilities of the tasks on the MindHive platform	• Find and share 2–3 articles related to their research topics • Participate in a new MindHive task and describe what it measures
5A Peer review •and revised proposals, peer reviewsStudents learn about the role of peer review in scientific research and explore and discuss principles and best practices of peer review. They further discuss how to revise studies based on peer feedback • Individual students first provide peer feedback to studies designed by students from another class and then discuss and synthesize their reviews as small groups	• Journal entry: how might your study proposals be impacted by the peer review process? • Individual peer review • Group synthesis
5B Peer review wrap-up • Students continue to synthesize their individual reviews in groups and reflect on the peer review experience as a whole	• Journal entry: how did you find the peer review process? • Post-review survey
5C Revise your study • Students read through the peer reviews of their studies and decide how to revise their studies	• Posttest survey

Adapted from authors (DATE)

Third, students followed a series of prompts we designed to guide them in expressing their curiosities about human brain and behavior; articulating and justifying research questions; reviewing relevant literature; articulating hypotheses and variables; describing participants and recruitment; designing methods; and discussing potential limitations, confounds, and alternative explanations, limitations (Table 3, Appendix Fig. 1). Students then used MindHive to create their proposed studies, which involved adapting and organizing surveys and tasks from a bank of research instruments (Appendix Fig. 2), and writing introductions, instructions, and debriefs for participants.

Students' peer review process

After students had created their own studies, they engaged in a discussion-based lesson on the role of peer review in science, including principles of, and best practices for peer review, and how to use feedback to revise studies. Following this lesson, students submitted their completed proposals through the peer review tool in MindHive, which made them visible to their classmates and students from their partner school in the class network's "Review" Dashboard (Appendix Fig. 3). Students were then assigned to individually review 1–7 studies from either their own, or a partner classroom at a different school.

To scaffold students' peer reviews, students used a MindHive tool that allowed them to select proposals to review. Prompts guided students' to give star ratings and written feedback on different aspects of the proposals, such as the importance of the research questions, the definition of variables, and the appropriateness of the study tasks (Table 4).

After completing their individual reviews, students reconvened in their study design groups, and served as review panels to discuss and synthesize their individual evaluations

Table 3 Prompts to guide students' proposal development

Proposal card category	Proposal card title	Proposal card description (abbreviated from the originals)
Brainstorm cards (optional)	Study topic: group synthesis	Fill out this card as a group (assign a scribe) AFTER everyone spends some time individually on the other cards in this section. Use this card to synthesize everyone's ideas and decide on a research topic
	Student study ideas 1–6	Fill out one "study ideas card" individually. When you're done, you will discuss your ideas with the group so you can collectively decide on a study topic. Use this card to synthesize your potential study ideas. Collect your class notes, journal entries, weblinks and any ideas that came to you in the shower or in conversations with your friends. Also list any MindHive tasks and surveys you are interested in and explain why
Research goals	Big research question	Describe your Big Picture research question in one sentence. Don't worry about this question being specific enough to be testable; that will happen elsewhere
	Significance	Summarize the goals and importance of your study in a few sentences. Describe the phenomenon you are trying to study and what motivated your team to study this phenomenon, based on what you have observed and read while conducting your background research. Keep your audience in mind: Who are you speaking to? Will they understand your description? Will they be convinced of your study's relevance to them and/or to society?
Background & knowledge gap	Knowledge gap	What knowledge gap is your study trying to fill? How will filling this gap help the broader public and scientific community?
	Background: brain	Which brain areas, processes, or chemicals are involved/relevant to your research topic and question? Synthesize what everyone discovered in their literature research. (see other cards in this section)
	Background: behavior	What is known about human behavior in relation to your research topic and question? Synthesize what everyone discovered in their literature research. (see other cards in this section)
	Background: discovery 1–6	Each student fills out one card individually. Explore your preliminary ideas and research question(s)! Conduct some literature research and organize what you found below: List and summarize 2 or 3 resources, and reflect on what you learned. Did you notice any patterns about your topic? Were there any facts that surprised you? Is there something you wanted to know about that is currently missing in the literature? Do you now have a burning question that is keeping you awake in the night?
Methods	Tasks | surveys	Describe the tasks and surveys you will use for this study
	Participants	Which study population are you investigating? Explain who your participants are and how many people you will recruit (in each group, if you have different participant groups)
	Procedure	What will your participants be asked to do? How often? How long will the study take to complete?
	Recruitment plan	Describe how you plan to recruit your participants and in which time frame. Will you send out an email? Recruit your classmates? or?

Table 3 (continued)

Proposal card category	Proposal card title	Proposal card description (abbreviated from the originals)
Research questions & hypotheses	Research question | hypothesis | predictor & outcome variables 1–4	Translate your Big Picture research question into a testable (sub) research question, and list your hypothesis. What is your predictor variable (the independent variable)? What is your outcome variable (the dependent variable)? Identify them here. Use multiple cards if you are testing multiple questions
Discussion	Alternative outcomes	What are any alternative outcomes of your study? go through your hypothesis/hypotheses and think critically about possible other ways that your study may pan out
	Study limitations	Discuss some possible limitations of your study below. These could be imperfections in your design or the scope of your research, etc
	Future directions	Discuss any implications of your findings in terms of possible future questions and studies

Table 4 Prompts to guide students' peer reviews

Peer reviewer prompts
Is the research question important? Why or why not?
Is the study design appropriate? How might you improve on the study design, if at all?
Do the predicted outcomes support the researchers' hypothesis?
Do the researchers consider possible alternative explanations for the study findings? Which might they be?
Does the study respect participants' privacy, health, and effort? Explain your reasoning
What further question could you address in a follow-up study?
What was it like to participate? Was it the right duration? Was the task clear? Were you motivated to put effort into your responses? Explain your answer
Does the study seem interesting? Would you choose to participate in this study? Would you recommend it? Why (not)?

for each of the 2–3 studies they reviewed. To do this, they responded to the prompts: "I like… I wish… I wonder…". This review panel activity was intended to further mimic certain professional review practices, which aim to strengthen reviewers' feedback by encouraging them to verify interpretations and come to consensus on their evaluation. It was also intended to enhance students' learning experience as it allowed peer reviewers to share their reasoning for their evaluations. In this study, we restrict our analysis to students' individual reviews, written before this group review synthesis.

Data

Our data include students' initial and revised proposals, peer reviews, and survey responses; interviews with teachers from three schools; and interviews with students from Gordonia and Myrtle (Table 1).

Study proposals. Student teams created written proposals for their research studies by responding to a series of prompts (Table 3, Appendix Fig. 1). Among other things, these prompts asked students to motivate their research focus, articulate hypotheses, define variables, and discuss potential confounds (see sample in Appendix Table 1). While our intention was for students to revise their proposals in response to their peer reviews, scheduling issues left them little time to do so. Our analysis thus focuses on students' final proposals, which were effectively identical to the proposals they submitted for peer review.

MindHive studies were the series of online participant activities, flanked with introductions and debriefings, that students built alongside their proposal development. MindHive studies included the surveys and tasks that students had chosen or adapted from the public bank, or that they had designed themselves (e.g., in Google Forms). Links to MindHive studies could be distributed to collect participant data, and were also available for peers to examine alongside the study proposals during peer review.

Peer reviews. Each student's peer review consisted of short responses (one to several sentences long) to 8 prompts (Table 4). These asked reviewers to comment on the significance of the research question, and the degree to which the proposals aligned methods with

hypotheses, fully considered alternative explanations, and addressed the ethical treatment of human subjects. Due to the current state of development of the MindHive review system, reviews were not traceable to individual students, so that it was not possible for us to link student reviews between the pre test, the unit, and the posttest.

Pre and posttests. Before and after the unit, we administered a survey designed to address various topics relevant to our broader project's research and design goals. Our analyses focus on several selected items. These include two sets of pre and post open-ended items, *Design a Study* and *Review a Study*, which together, consisted of 10 short-answer questions that prompted students to design a research study given a research question, and to review a hypothetical study (Table 5); as well as two posttest items that asked students to rate their perceived helpfulness of, and trust in peer reviews. Finally, we used an open-ended posttest item that asked students, "What did you learn from the peer review process?" in order to qualitatively capture some of the students' experiences. Further detail on these items is provided in the description of our data analysis below.

Interviews. We drew on student and teacher interviews to provide context to our interpretations of other analyses. Interviews were conducted on Zoom following implementations with each of the 4 teachers at all three of the implementing schools, and with 8 (5 males, 3 females) students from the two private schools, who were selected based on the recommendation of the teacher, and on assent and parental consent. Among other topics, teachers' interviews (~ 1 h long) addressed their experiences with implementing the unit and their impressions of students' learning. Student interviews (~ 30 min long) addressed students' experiences with study design and peer review.

Table 5 The pre/posttest items targeting students' study design and review abilities

Study design prompt	Study review prompt
Pollution in the local river has been mapped over many years. How could scientists study how the river's water quality impacts students' school performance? • Explain why the research question is important • How would you investigate the relationship between water quality and student performance? (e.g. Who would be your participants? What would you measure, when, and why?) • What outcomes do you expect to find if your prediction is true? • What might be possible alternative explanations for your findings?	Researchers want to find out if air pollution interferes with performance in outdoor sports. They decide to measure 10 runners' times in a 100-m sprint on a day where air quality is high, and again on a day where air quality is low. If most of the runners have slower race times on the day with low air quality, and faster race times on the day with higher air quality, then the researchers will conclude that air pollution has a negative impact on outdoor sports performance • Is the research question important? Why or why not? • Is the study design appropriate? How might you improve the study design, if at all? • Do the predicted outcomes support the researchers' hypothesis? • Do the researchers consider possible alternative explanations for the study findings? Which might they be? • Does the study respect participants' privacy, health, and effort? Explain your reasoning • What further question could you address in a follow-up study?

Data analysis

Quantitative analysis of the foci and quality of students' study designs (RQ1)

To describe the studies that students proposed, we analyzed the written proposals and the MindHive studies that students created during the unit. To describe the foci of these studies (RQ1a), we categorized their research questions according to the construct students proposed to investigate, and calculated the frequencies of studies that fell into these categories.

To determine students' study design abilities (RQ1b and RQ1c), two researchers iteratively refined a *study quality* rubric through initial rounds of independent coding and discussion of 3–4 studies at a time. This rubric is adapted from published rubrics for assessing students' inquiry in experimental sciences, which address similar dimensions of study design (e.g., conceptualization, problem solving, ethical reasoning) (e.g., Fine & Pryiomka, 2020; Halonen et al., 2003). Our own rubric was designed to align with the scope and focus of our curriculum, and to the developmental stage of our high school level participants. Once we had defined a working rubric, the two researchers used it to independently score approximately 10 studies, iterating on the categories through discussion and re-coding until we had achieved near perfect agreement ($K_w = 0.81$) (Fleiss et al., 2003; Landis & Koch, 1977). One researcher then coded the rest of the studies.

This *study quality* rubric rated students' studies along several categories based on characteristics that were salient to us across the proposals, and that reflected the literature on students' experiment design abilities. These categories included a justification of the importance of the research questions, a basis in relevant existing research, proper definition and operationalization of variables, alignment between the study proposed and the study tasks created in MindHive, awareness of potential confounds and limitations, adherence to ethics for research with human subjects, and attendance to the participant experience (Table 6).

To determine students' overall study design abilities (RQ1b), we calculated an *overall study quality* score for each study—a sum of the scores across each dimension of the rubric—and calculated the average of this overall score. To determine students' performance on specific aspects of study design we calculated average scores along each dimension of the *study quality* rubric (e.g., defining variables). To characterize the quality of students' study designs, we compared mean scores of their studies' individual quality dimensions using an ANOVA, with Tukey's HSD posthoc tests to follow up on any significant differences. We used this analysis for studies created at each timepoint, i.e., pretest, in-unit, and posttest.

To determine the change in students' study design abilities (RQ1c), we scored their individual responses to the pre and posttest Study Design item (Table 5) using a rubric adapted from the Study Quality rubric (Table 6). This adapted rubric (Table 7) ensured that we aligned our categories to the ways that the pre/posttests prompted students to respond. We randomized the responses so that coders were blind to which responses were pre vs. post. Once we had achieved substantial agreement ($K = 0.78$), one researcher coded the rest of the responses.

To increase statistical power, we combined data across all classes for the remainder of our analyses after determining that there were no significant differences between the overall pretest study quality scores of the two classes who completed the pretest (Gordonia and Myrtle), $F(1, 16) = 1.48$, $p = 0.242$, nor significant differences between the six classes on the posttest, $F(5, 77) = 0.787$, $p = 0.562$. With classes combined, we used a pooled t-test

Table 6 Rubric for scoring students' in-unit study designs

Code	Description	Scoring
Justifies importance	Does the proposal justify the importance of the research question by: • Giving personal reasons for their curiosity? • Explaining real-world implications of potential findings? • Avoiding overstatements of the importance, impact, and/or applicability of potential findings?	0: 0 criteria are met 1: 1–2 criteria are met 2: 3 criteria are met
Support from literature	Does the proposal build on existing literature? • Is the literature identified relevant to the research question? • Are the methods of at least one source mentioned? Mentions can be vague, e.g., "researchers compared, investigated, tested, surveyed." • Are the findings of at least one source described? • Is there an explanation provided for how at least one source informs the students' own proposal?	0: No literature identified, or is irrelevant, or is not summarized (neither methods nor findings) 1: Literature is identified and relevant, but the summary mentions only the methods, or only superficial findings (e.g., saying that personality is related to music preference without specifying how) 2: Relevant literature is summarized, its methods and superficial findings are described, or only the findings are described specifically (e.g., saying how different personalities are related to music preferences). The response MAY also explain its connection to the proposal, but this is not necessary for a 2
Variables	Are students able to articulate their IV and DV? • Are the IV and DV clearly defined and operationalized such that it is clear what and how something will be measured?	0: All variables are incorrectly defined and operationalized in relation to the hypothesis and/or were not addressed 1: Only some variables relevant to the RQs are defined and/or operationalized, and/or are not clearly stated, or are absent 2: All variables are clearly defined and operationalized
Alignment	• Do students build a MindHive study that is aligned with their proposed hypothesis? • Are study tasks/variables aligned with the hypotheses? In other words, would data from the MindHive study address the hypotheses?	0: No, the task will definitely not produce the data needed to verify the hypothesis 1: Mostly. There are more/fewer tasks than necessary, or the data generated would only allow proposed hypotheses to be partially addressed, or the study tasks proposed are good but the ones in the MindHive Study are not, or vice versa 2: Yes, all the tasks are well aligned with the hypothesis proposed. The hypotheses could definitely be addressed with the data generated by the study. There are no more/fewer key tasks than are necessary (surveys asking for demographic or other contextual info relevant to the hypothesis are fine)

Table 6 (continued)

Code	Description	Scoring
Participant recruitment	Criterion 1: • Target population of participants is described and seems reasonable for the study's goals Criterion 2: • Sample size is mentioned and seems reasonable for the study's goals; OR • There is a rationale given for the proportion of participants in each group being compared Criterion 3: • There is a recruitment plan (e.g., study links will be distributed on social media)	0: No 1: Only one criterion met 2: Two criteria are met, or only one is met but is nicely explained
Confounds, limitations	• Does the proposal rationalize a study design decision in terms of avoiding potential confounds? • Does the proposal acknowledge potential confounds? • Does the proposal consider alternative explanations for the study's anticipated findings? • Does the designer explain the identified confounds and/or alternative explanations in terms of how these might impact anticipated findings?	0: No confounds, alternative outcomes, or limitations are mentioned 1: At least two of either a confound, alternative outcome, or limitation are mentioned, but not explained OR At least one of either a confound, alternative outcome, or limitation is mentioned AND it's explained 2: At least two of either a confound, alternative outcome, or limitation are mentioned AND at least one of these is explained, OR 2 + confounds or 2 + alternative outcomes or 2 + limitations are mentioned and at least one of these is explained
Research ethics	Does the protocol adhere to ethical standards for human subjects research? Criteria: • There are no invasive questions (e.g., everything is anonymous. No personally identifying information is requested) • No tasks put participants at risk, and no questions make them feel uncomfortable (e.g. questions are not triggering. Tasks are not knowingly unsafe. If so, there are measures taken to warn participants, and to give them the choice to proceed) • Data is kept confidential, not shared with outside organizations (e.g., no need to sign into your Google account to complete a Google survey)	0: 1 criterion is met 1: 2 criteria are met 2: 3 criteria are met

Table 6 (continued)

Code	Description	Scoring
Participant experience	Do the proposal and the study attend to creating a positive participant experience? Criteria: • Tasks are of appropriate length and relevant to the research question (i.e., there are no unnecessary tasks that will generate data that was not proposed) • Includes a motivation/explanation for participating (explains the importance of the study and/or reasons for particular tasks) • Instructions for how to participate are clear • The affordances/constraints of the tools are used appropriately such that it is easy to participate (e.g., tasks are feasible, survey links are accessible)	0: 0–1 criterion met 1: 2–3 criteria met 2: 4 criteria met

Table 7 Rubric for scoring students' pre and posttest study designs

Code and applicable prompts	Description	Scoring	Examples
Justifies importance Explain why the research question is important	Does the proposal justify the importance of the research question by: • Explaining why it is important to know the answer to the research question? • Explaining real-world implications of potential findings? • Avoiding overstatements of the importance, impact, and/or applicability of potential findings?	0: None of the criteria are met 1: 1–2 criteria are met 2: 3 criteria are met	0: The research question is the basis of any scientific breakthrough and is thus key to the scientific method 1: *Because it is focusing on the well being of the affected community* 2: it is important because pollution is extremely bad for the environment especially when found in freshwater that humans drink, which is a limited resource in itself. This question looks at its effect on students and the performance which *can show the dangerous effects of pollution in drinking water and can further help the understandings of the dangers of pollution in the developing body and brain of a human*
Variables How would you investigate the relationship between water quality and student performance? (e.g. ….What would you measure, when, and why?) "What outcomes do you expect to find if your prediction is true?"	Are students able to articulate their IV and DV? • Are the IV and DV clearly defined and operationalized such that it is clear what and how something will be measured?	0: All variables are incorrectly defined and operationalized in relation to the hypothesis and/or were not addressed 1: Only some variables relevant to the RQs are defined and/or operationalized, and/or are not clearly stated, or are absent 2: All variables are clearly defined and operationalized	0: I would expect to find an environmental factor that is harming the water *[no mention of how water quality and student performance will be operationalized]* 1: I would *measure one grade only* for consistency over the course of 5 years 2: All of the students, see *where they all live, see what water quality they are exposed to*, and then look at *their grades*. You need all of this information to try and see a correlation

Table 7 (continued)

Code and applicable prompts	Description	Scoring	Examples
Alignment How would you investigate the relationship between water quality and student performance? (e.g. ….What would you measure, when, and why?) What outcomes do you expect to find if your prediction is true?	• Do students propose a study that is aligned with their proposed hypothesis? • Are study tasks/variables aligned with the hypotheses? In other words, would data from the study address the hypotheses?	0: No, the task will definitely not produce the data needed to verify the hypothesis 1: Mostly. There are more/fewer tasks than necessary, or the data generated would only allow proposed hypotheses to be partially addressed 2: Yes, all the tasks are well aligned with the hypothesis proposed. The hypotheses could definitely be addressed with the data generated by the study. There are no more/fewer key tasks than are necessary (surveys asking for demographic or other contextual info relevant to the hypothesis are fine)	0: I would use students that are close in proximity to the local river and ask them about how they feel after they drink from this water source. If they experience diarrhea, abdominal cramps, nausea, or vomiting, I will be able to conclude that the water is polluted. *[This approach will not produce data on student performance.]* 1: I would measure one grade only for consistency over the course of 5 years. *[This approach will only produce data on student performance, not on the relationship between performance and water quality.]* 2: One could *look at things like test scores, who is drinking the water. Are students with better water quality near by performing better?*

Table 7 (continued)

Code and applicable prompts	Description	Scoring	Examples
Participants How would you investigate the relationship between water quality and student performance? (e.g. Who would be your participants?…)	Criterion 1: • Target population of participants is described and seems reasonable for the study's goals Criterion 2: • Sample size is mentioned and seems reasonable for the study's goals; OR • There is a rationale given for the proportion of participants in each group being compared Criterion 3: • There is a recruitment plan (e.g., study links will be distributed on social media)	0: No 1: Only one criterion met 2: Two criteria are met, or only one is met but is nicely explained	0: Preform better with water worse without *[No mention of participants.]* 1: tests different groups with different water qualities and see grades and compare *[Does not indicate who comprises the groups.]* 2: I would investigate how well some *students are doing in school who actually drink from the contaminated water source* and investigate how well *students who don't drink from the water source* are doing. We would also look at those students and how they were performing in school pre-pollution within the water

Table 7 (continued)

Code and applicable prompts	Description	Scoring	Examples
Alternative explanation What might be possible alternative explanations for your findings?	• Does the response identify and explain alternative explanations for the predicted finings?	0: No alternative explanations are mentioned 1: An alternative explanation is mentioned, but not explained 2: An alternative explanation is mentioned AND explained; OR, more than one alternative explanation is mentioned, though not necessarily explained	0: Some variables may not be controlled or well regulated. *[No alternative explanations are mentioned.]* 1: Alternative explanations would be factors that vary between students, like previous academic performance or family influence. *[Alternatives are mentioned but not explained.]* 2: We might also be seeing that the *students who are drinking contaminated water are being forced to stay home because they are sick from the water and that is what is affecting their performance in school rather than the water's effect on them mentally*. There might also be the explanation that the *students who are drinking contaminated water come from lower-income families and therefore were already not focusing in school because they had outside pressures already and the water played no actual role*

Table 7 (continued)

Code and applicable prompts	Description	Scoring	Examples
Ethical considerations How would you investigate the relationship between water quality and student performance? (e.g. What would you measure, when, and why?) What outcomes do you expect to find if your prediction is true?	Does the response adhere to ethical standards for human subjects research? Criteria: • There are no invasive questions (e.g., everything is anonymous. No personally identifying information is requested) • No tasks put participants at risk, and no questions make them feel uncomfortable (e.g. questions are not triggering. Tasks are not knowingly unsafe. If so, there are measures taken to warn participants, and to give them the choice to proceed) • Data is kept confidential, not shared with outside organizations	0: There are 2 or more ethically questionable aspects of the study, or one severe 1: There is one ethically questionable aspect of the study 2: There are no ethically questionable aspects of the study. The study is carefully designed to avoid putting participants at risk	0: give clean water and not great water to different groups and see the difference in student performance *[Knowingly giving participants unclean drinking water violates participants' health and safety.]* 1: I would gather a group of twenty children, randomly selected from different grades. I would measure the pH level of drinking water, the student's daily intake of the local water (for about three times a day for a week), and notice their test grades. Every week, I would raise the drinking pH level until the pH is 7, and every week I would observe their test scores. From this experiment I would be able to notice a pattern. *[The EPA recommends that drinking water is < pH 8.5. Nevertheless, a correlational study design would arguably be a safer approach than one in which the quality of drinking water is manipulated.]* 2: I would test all the students, including those both near and far from polluted water sources. I would then contrast that data between the grades of the students with the highest pollutants and the lowest pollutants and see if there is a trend. *[A correlational study is proposed that avoids knowingly exposing participants to harmful contaminants.]*

to compare students' pre and posttest scores on the Study Design item to see whether and how students' study design abilities may have changed.

Next, to more deeply explore pre to post change in study design abilities, we looked just at those students who had matched pre and posttest responses (N = 17). Using a Shapiro Wilk Goodness-of-Fit test, we found that the pre-post difference scores on the design item were normally distributed (W = 0.921, p = 0.135). We then conducted a paired t-test to find any significant differences between the overall study design scores on these students' pre and posttests.

Quantitative analysis of the quality of students' peer reviews (RQ2)

To analyze students' peer reviews, the same two researchers iteratively refined a *review quality* rubric through independent coding and discussion of disagreements, similar to the process described for RQ1b and RQ1c. Each peer review consisted of statements of one to several sentences long in response to 8 prompts (Table 4). Two researchers read 3–4 review statements at a time, and discussed the elements that were both apparent across students' statements, and that reflected definitions from the literature on what constitutes a helpful peer review. Based on emergent themes, and on our comparison to the literature, we defined effective peer reviews as ones that were *specific* to the study being reviewed; that offered an actionable *recommendation* for improvement, and that provided an *explanation* for the evaluation. To determine the degree to which reviewers were critical of the studies, we also coded whether or not reviews explicitly *identified faults* with the study design (Table 8).

Once we had defined a working rubric, two researchers used it to independently score a set of ~ 15 review statements (~ i.e., the set of responses to all review prompts by three individual students, Table 4) at a time, refining category definitions through discussion of disagreements, then repeating this process until we had achieved near perfect agreement ($K_w = 0.881$) (Fleiss et al., 2003; Landis & Koch, 1977). One researcher then coded the rest of the peer reviews.

To determine the overall quality, as well as the particular strengths and weaknesses of students' reviews (RQ2a), we created an *overall review quality* score for each student's review, which was the sum of the *specificity*, *recommendation*, and *explanation* scores of each of their review statements. We did not include *identified faults* in this overall score because this category was intended to categorize reviews that indicated explicit weaknesses, not to describe the quality of a review. As well, whether or not students gave a *recommendation* already assumed that they had implicitly or explicitly identified an area for improvement (i.e., a fault).

To understand how review dimensions differed from one another, we created 4 *dimensional* scores for each review, which were the sum of scores on particular rubric categories (*recommendation, specificity, explanation, identified fault*) across each statement of a review. We explored how mean scores on individual quality dimensions (recommendation, specificity, explanation) differed within students' in-unit peer reviews by using an ANOVA with Tukey's HSD posthoc tests to follow up on any significant differences. We also compared overall review scores between students' pre and posttest responses using pooled t-tests. To explore whether students were becoming more or less critical in their reviews over the course of the unit, we used an ANOVA to compare mean *identified fault* scores of students' reviews generated at three timepoints: pretest, in-unit, and posttest. Additionally,

Table 8 Rubric for scoring students' study reviews

Code (1 or 0)	Description	In-unit examples	Pre/posttest examples
Identifies fault	• The reviewer identifies, with any level of specificity, a need for improvement • A recommendation implies that the reviewer found fault	[The authors] *did not provide* alternative explanations	It is appropriate, *however* multiple trials should be run, since the runners could have a bad day and not perform well, regardless of the air quality
Gives recommendation	• The reviewer provides concrete suggestions for improvement that give clear indications to the authors on how to improve	*I'd suggest telling the participant to pay attention to the emotions*	It is appropriate, however *multiple trials should be run*, since the runners could have a bad day and not perform well, regardless of the air quality
Is specific	• The review is specific to the proposal, using keywords and terminology related to the study's content or design, such as by describing specific constructs identified, or methods/approaches used. There is no doubt that the review applies to a particular study	I think [the research question] is important, *masks* have become a big part of our life, and the *ability to gauge emotion* affects social interactions	It is appropriate, however multiple trials should be run, since the *runners* could have a bad day and not *perform* well, regardless of the *air quality*
Gives explanation	• The review provides reasons for their evaluation (There is an explicit or implied "because," and not simply a Yes/No.) • The reasons do not simply restate the prompt but use different terminology to suggest that reviewers are thoughtful of the underlying meaning	I'd suggest telling the participant to pay attention to the emotions *so they know what to look for, especially since they can't go back to the video link in the form after they submit*	It is appropriate, however multiple trials should be run, *since the runners could have a bad day and not perform well, regardless of the air quality*

we used a pooled t-test to compare if and how pre and posttest means of the category *identified fault* differed.

To see how students' review abilities may have changed over time (RQ2b), we used the *Review Quality* rubric (Table 8) to score their individual responses to the pre and posttest *Study Review* item. We randomized the responses to ensure that the coder was blind to which responses were pre vs. post. As with the pretest study quality scores (see above) we increased statistical power by combining data from all classes after determining that there were no significant differences in overall review quality between classes at the pretest, $F(1, 16) = 1.48$, $p = 0.242$, $N = 18$, nor significant differences in overall review quality between classes at the posttest, $F(5, 69) = 1.48$, $p = 0.207$, $N = 75$. With classes combined, we did a pooled t-test to compare students' pre and posttest scores on the *Study Review* item.

To further investigate pre to post change in study review abilities, we next considered just those students who had matched pre and posttest responses ($N = 17$). We used a Shapiro Wilk Goodness-of-Fit test, and found that the difference scores on the review item were normally distributed ($W = 0.968$, $p = 0.764$). We then conducted a paired t-test to find any significant differences between the overall study review scores on these students' pre and posttests.

Qualitative analysis of students' study designs and peer reviews (RQ1 and RQ2)

To complement the quantitative analyses used to answer RQ1 and RQ2, we also performed a qualitative analysis of students' work. Specifically, guided by the dimensions of the *study quality* and *review quality* rubrics described above, two researchers examined and compared students' study designs and peer reviews. Our goal was to describe the range of students' abilities visible across their work, and to illuminate features not captured by our coding rubrics. We therefore focused on identifying examples from across students' study designs and peer reviews that would illustrate their varying abilities to reason about, and critique: implications for research; how a proposed study builds on prior research; variable definition and operationalization; potential confounding variables; adherence to research ethics; and the optimization of research participants' experiences. Findings from this analysis were organized according to these rubric themes, and presented as a qualitative description of illustrative and contrasting examples of students' work.

Determining students' experiences with open science (RQ3)

We triangulated different data sources to describe students' experiences participating in open science (RQ3). First, we analyzed students' responses to two 7-point Likert-type posttest items, which asked students to rate their agreement with the following statements: "Our peers' reviews gave us helpful feedback that we used to revise our work," and "I mostly disagreed with our peers' reviews." We used these responses to determine the perceived helpfulness of peers' reviews, and the trust that students had in their peers' reviews.

Second, we used a consensus approach (Cascio et al., 2019) to categorize students' responses to an open-ended item on the post-survey, which asked "What did you learn from the peer review process?" One researcher developed categories based on an initial reading and coding of all responses. A second coder then used these categories to independently code all the responses. We resolved disagreements through discussion, and modified and re-coded the categories accordingly. We then calculated the frequencies of responses in each category.

Third, we read transcripts of students' and teachers' interviews, and identified instances of students' self-reports, and of teachers' observations, of students' experiences participating in open science. These instances included descriptions of the challenges and/or perceived value of engaging in the study design and peer review processes. We used these instances to qualify students' experiences, and to complement our interpretation of our other data sources. In our findings, we use quotes from interviews and the short-answer survey item described here to illustrate the themes that emerged.

Findings

For ease of readability, this section first presents quantitative findings regarding students' study designs (RQ1) and peer reviews (RQ2). Following this, we present qualitative findings that address dimensions of quality across both students' study designs and peer reviews.

RQ1 what kinds of research studies do students design?

RQ1a students' research foci

Documents of individual students' research ideation demonstrate their curiosity about, and intuitions for the questions that can be answered by human brain and behavior science methods. Most of the students' questions appeared to stem from, and to have direct implications for students' everyday experiences and concerns. As one student wrote of their interest in superstitions:

> As a person who participates in sports, I see a lot of teammates do things or not do certain things for luck. I just really wonder how we as humans develop superstition, and why we develop it.

Another student wrote of their curiosity about birth order effects:

> My question was "Does being the youngest, oldest, middle, or only child affect your independence in daily life at the ages (14–18)? ". One thing that draws me to my questions is my wonders in life and things that I've noticed around me. I've always taken curiosity in the way people are in their family like middle or only or oldest or youngest since I feel like there's something that changes or alters your behavior based on that.

The 26 proposals that students eventually coalesced upon with their teams covered diverse foci, from impacts of the pandemic on mental health, to the effects of music on memory (Table 9). For example, a number of teams proposed studies to investigate ideal conditions and strategies for studying, such as "how the amount of time after [a person] wake[s] up affects their performance on critical thinking skills and memory" to show "what time they are the best at test taking." Meanwhile, others sought to describe the impacts of living in the pandemic on people's mental and physical health. For instance, one team's proposal asked "What impact does wearing a mask have on high schoolers' ability to gauge emotion?".

Without exception, students proposed participants who were, or at least included, adolescents like themselves. Besides being a personally relevant age group, students seemed to

Table 9 Student teams' proposal topics

Number	Topic	Sample research question
10	Education	Is it easier to focus online or in person?
8	Memory	How do different speeds of music affect high school students' ability to retain new information?
8	Mental health	How does lock down, being stuck indoors and social isolation affect how likely teens are to show emotional disorders symptoms (like depression)?
7	Music	How do different types of music affect emotion?
6	COVID-19	How optimistic are teenagers about a return to normalcy following the initial COVID-19 vaccine rollout?
5	Emotion	How does the emotional salience of the words impact your short-term memory of them?
4	Development	How has the pandemic affected different age groups' mental health?
3	Social influence	Is a teenager more likely to alter their opinion of a tweet or piece of information on social media after seeing a well-known and verified account had posted the tweet?
3	Environment	Do sources of news impact behaviors regarding environmental warnings?
2	Personality	Does extroversion or introversion impact mental health during the pandemic?
2	Color perception	Do people of different ethnicities make the same emotion-color associations?
1	Sleep	What is the effect of a stressful day on the amount of sleep someone gets?
1	Food	Has the Pandemic Changed the Environmental Impact of People's Eating Habits?
1	Language	How do word-color associations impact organization and perception? How do category-color associations affect the speed of classification? (Study 12)
1	Social perception	Does political affiliation affect how we interact with and view others?
1	Art	Does art inspire people to create social change?
1	Culture	Do people of different ethnicities make the same emotion-color associations?

Categories are not mutually exclusive

recognize the more general need for a scientific understanding of this population. As one student wrote:

> I… think we should stick to the ages 14–18 or highschool age as that's an age a lot of scientists want to focus on but don't have a lot of information about as we are a very interesting age.

RQ1b quality of students' in-unit study designs

The team team studies created during the unit (N = 104) had a mean overall quality score of 12.74 out of a possible 16 points (SD = 2.30, range = 4–16). Scores on individual dimensions (each with scores ranging from 0–2) were significantly different from one another, F (7,824) = 19.34, p < 0.0001. (Table 10, Appendix Fig. 5). Study designs scored highest on the dimensions *Participant recruitment* (M = 1.88, SD = 0.46), *Attention to participant experience,* and *Builds on literature.* They scored lowest on *Alignment between hypothesis and methods* (M = 1.22, SD = 0.65), followed by *Attention to confounds and limitations*.

RQ1c pre to post changes in study quality

A pooled t-test of students' pre and posttest responses to the *Study Design* item showed no significant differences in overall study quality scores between the pre (M = 7.70, SD = 2.64, N = 20) and posttest (M = 7.33, SD = 2.72, N = 83), t(29.53) = 0.57, p = 0.58, d = 0.14 (Table 10). We found similar results when we excluded incomplete responses, and used a paired t-test to compare only those students who had completed both the pre (M = 7.78, SD = 0.59, N = 18) and posttest (M = 7.44, SD = 0.59, N = 18), t (16) = 1.04, p = 0.31, d = 0.58.

Certain individual study quality dimensions differed significantly from pre to post (Appendix Fig. 5). Specifically, students were better at justifying the importance of a research study by the posttest (M = 1.34, SD = 0.85, N = 83) than they were at the pretest (M = 0.85, SD = 0.88, N = 20), t (28.17) = -2.25, p = 0.03, d = 0.57. However, students appeared to decline in their abilities to discuss potential confounds, limitations, and alternative explanations of their study designs from the pretest (M = 1.35, SD = 0.88, N = 20) to the posttest (M = 0.82, SD = 0.78, N = 83), t (26.81) = 2.48, p = 0.02, d = 0.64. The effect sizes of each of these results suggests medium practical significance. Students' abilities on other study quality dimensions remained unchanged from pre to post.

Individual quality dimensions of students' pretest study designs differed significantly from one another, F (5, 114) = 5.02, p = 0.0003, N = 20. Pretest scores on *Justification of Importance* and *Research Ethics* were especially lower than pretest scores on other dimensions.

Individual study quality dimensions also differed significantly from one another at the posttest, F (5, 492) = 21.46, p < 0.0001, N = 83. Like in the pretest, students struggled with considering ethics in their research design decisions (M = 0.73, SD = 0.44). Unlike the pretest, students appeared to struggle with attending to confounds and limitations (M = 0.82 SD = 0.78) while having no trouble justifying the importance of the research (M = 1.34, SD = 0.85).

Examples of student responses that demonstrate these pre-post trends can be viewed in Table 11, and possible reasons for them are explored in the Discussion section.

Table 10 Means and standard deviations of individual quality dimensions of student teams' pretest, in-unit, and posttest studies

Dimension of study quality	Mean (SD)			Pre-post t-test
	In-unit (N = 104)	Pretest (N = 20)	Posttest (N = 83)	
*Justifies importance	1.62 (.60)	.85 (.88)	1.34 (.85)	t (28.17) = − 2.25, p = .03, d = .57
Support from literature	1.85 (.34)	N/A	N/A	N/A
Definition & operationalization of variables	1.63 (.55)	1.5 (.51)	1.47 (.61)	t (33.34) = .23, p = .82, d = .05
Alignment	1.19 (.64)	1.55 (.78)	1.52 (.65)	t (27.82) = .19, p = .85, d = .04
Participant recruitment	1.87 (.46)	1.60 (.60)	1.45 (.72)	t (33.63) = 1.00, p = .33, d = .23
*Discussion of confounds, limitations, alternative explanations	1.53 (.70)	1.35 (.88)	.82 (.78)	t (26.81) = 2.48, p = .02, d = .64
Attention to research ethics	1.78 (.54)	.85 (.49)	.73 (.44)	t (27.04) = .96, p = .35, d = .26
Attention to the participant experience	1.3 (.61)	N/A	N/A	N/A

*indicates significant pre-post differences

Table 11 Selected student responses showing pre-post improvement and decline in quality dimensions of study design

	Pre response (score)	Post response (score)
	Prompt: Pollution in the local river has been mapped over many years. How could scientists study how the river's water quality impacts students' school performance?	*Prompt: Researchers want to find out if air pollution interferes with performance in outdoor sports. They decide to measure 10 runners' times in a 100-m sprint on a day where air quality is high, and again on a day where air quality is low. If most of the runners have slower race times on the day with low air quality, and faster race times on the day with higher air quality, then the researchers will conclude that air pollution has a negative impact on outdoor sports performance*
Justifying importance *Explain why the research question is important*	A student body's school performance and the quality of a local river are both important things to measure and compare for relationships. (0)	The question is important because if the water quality negatively impacted student performance, a solution could be identified and presented. (2)
	The research question is important because it sets up what the scientist is going to investigate. (0)	It's important because it affects the lives of children who will grow up to create more offspring and if there is a negative cycle regarding how both the parents and offspring learn, we will not be able to advance. (2)
	Pollution in the local river will soon enter the local town's freshwater system, if it is not the freshwater source already. The students drinking the water will have a worse school performance, as they are consuming toxins. (1)	The research question is important because it is about a social issue (climate change and water pollution) and it could be having a negative effect on the community, which would require immediate action. (2)
Alternative explanations *What might be possible alternative explanations for your findings?*	There are other variables affecting the children's education (0)	External factors, like what the learning environment is like, what the teachers and home situations are like. (1)
	Some variables may not be controlled or well regulated. (0)	The pollutant is actually beneficial to performance. (1)
	There are alternate factors that are damaging the child's performance, such as family or health issues. (2)	It might be that the range of the scores is not too different and therefore not deemed an important issue. (1)
	Other external causes that could affect performance such as a pandemic, trends, etc. (2)	Perhaps something else is affecting student behavior and not the river. (1)
	The students could have contracted a virus or could have eaten something that made them feel sick. However, it's unlikely that all of them would feel sick from another source as they all share a water supply. (2)	A possible alternative explanation for my findings could be the school system in the area of the polluted water; maybe the teachers aren't productive? (1)

RQ2 what kinds of reviews did students generate?

RQ2a quality of peer reviews generated during the unit

Peer reviews ranged greatly along the dimensions we analyzed (see Table 12 for examples), from brief responses to each prompt ("I think it is apporpriate [sic], I don't see any problems."), to elaborate evaluations of multiple aspects of the study designs (e.g., Example 2 in the *Recommendation* row of Table 12). Students' reviews were generally positive, mostly rating their peers' studies 3–4.5/5 stars (1–2 stars were only given to proposals that lacked required sections), and identifying few faults ($M = 2.28/6$, $SD = 1.15$, $N = 240$). Reviews were equally positive about peer-designed studies during the unit (e.g., "Honestly really good job on the [study] design- I really like how it is planning on being set up.") as they were about studies reviewed on the pretest ($M = 2.33$, $SD = 0.89$, $N = 18$) and the posttest ($M = 2.08$, $SD = 0.97$), $F(2, 237) = 1.66$, $p = 0.19$. As well, this positivity did not change from pre to post, $t(27.12) = 1.05$, $p = 0.30$, $d = 0.27$.

Scores on individual review quality dimensions differed significantly, $F(2, 438) = 212.64$, $p < 0.0001$ (Table 13, Appendix Fig. 4). Students' reviews were more likely to be specific ($M = 3.91$, $SD = 1.37$) than they were to offer a recommendation ($M = 1.15$, $SD = 1.07$), $p < 0.0001$, $d = 2.25$). Reviews were also more likely to give an explanation ($M = 3.78$, $SD = 1.42$) than they were to offer a recommendation, $p < 0.0001$, $d = 2.09$). The large effect sizes of these results suggest that they had high practical significance. To illustrate, consider this peer reviewer's comments on the appropriateness of one student group's proposal to use a self-report survey to investigate changing dietary habits before and during the pandemic:

> I think the study design is appropriate, however, it will be difficult for participants to determine what they were craving before the pandemic and to get them to be honest about it.

Reflective of the trends found across our sample, this excerpt shows how the student reviewer identified a specific threat to the validity of the data, but stopped short of offering recommendations for addressing it.

RQ2b pre to post changes in review quality

There were no significant differences between the overall review quality scores on the pre ($M = 7.06$, $SD = 3.00$, $N = 18$) and posttest ($M = 6.83$, $SD = 3.03$, $N = 75$), $t(26.11) = 0.29$, $p = 0.77$, $d = 0.08$. We also found no significant differences in overall review quality scores when we did a paired t-test among only those students who had completed both the pre ($M = 7.12$, $SD = 3.06$, $N = 17$) and the posttest ($M = 6.06$, $SD = 3.47$, $N = 17$), $t(16) = 1.526$, $p = 0.147$, $d = 0.32$. Individual review quality dimensions also did not differ significantly between pre and post (Table 13).

Among pretest responses, however, scores on individual review quality dimensions differed significantly, $F(2, 51) = 4.68$, $p = 0.01$, $N = 18$. In particular—and similar to what we observed of peer reviews generated during the unit—students' reviews were more likely to be specific ($M = 2.94$, $SD = 1.30$) than they were to give a recommendation for improvement ($M = 1.67$, $SD = 1.03$), $p = 0.003$, $d = 1.08$). The effect size of this result indicates high practical significance. As an example, consider the following student's pre-test review

Table 12 Contrasting examples of peer reviews for each dimension of review quality

Dimension of review quality	Examples of students' peer review statements
Specificity	**Study reviewed:** "The Effect of School on Sleep" **Reviewer prompt***:* Q1: Is the research question important? Why or why not? **Example 1** Reviewer prompt: I think this research question is important, as it is very applicable to a large part of the population **Example 2** It is important to understand how stress plays a role in sleep. It might be a good idea to in the future extend the study to find ways to reduce stress **Example 3** I think a confounding variable would be that it would take people some time to get used to the program in the beginning, so that may be a reason why people may be slower in the beginning and faster at the end. Also, it was a little annoying that all the words had to be correctly capitalized
Recommendation	**Study reviewed:** "Musical Memory" **Reviewer prompt:** What was it like to participate? [Comment on duration, clarity of instructions, quality of engagement] **Example 1** I think it felt a bit long and boring because I had to complete the memory recall task multiple times, but overall I was motivated because I think it's a really interesting research question. I would recommend people to participate in this study because I think this question is one that requires a lot of data points, so it would be interesting to see any general trends that come out of this **Example 2** The first task I did, I found it pretty fun to participate however I began losing motivation when the same images kept on repeating for each task. I feel like that made it harder for me to answer correctly because I was more confused about whether I had seen that image from this task or the previous task than focused on the music. Also, I think it would be nice if the tasks were put in a more clear order since they had numbers next to them but weren't put in order of those numbers but the duration was right, somewhere around 15 min. This study definitely does seem very interesting and I would recommend it since I'm interested to know the results! **Study reviewed:** "Let's Argue! Hearing Hot Topics" **Reviewer prompt:** Is the study design appropriate? How might you improve on the study design, if at all? **Example 1** "I think this study design was appropriate however, some of the questions were worded a bit confusing. Also, for the Initial Affiliation Survey, the buttons weren't working." **Example 2** "In the Initial Affiliation Survey, I couldn't really tell if I was clicking my answer because it didn't show anything. I think that a slider with strongly agree at one end and strongly disagree at the other end would be better and easier for the participants to use. I think that in the first set of scenarios, there could be an other space for the "how would you talk to her". On one of the questions, I didn't really agree with any of the possible answers, so having an other button would be helpful and more effective." **Study reviewed:** What impact does wearing a mask have on high schoolers' ability to gauge emotion?" **Reviewer prompt:** Is the study design appropriate? How might you improve on the study design, if at all? **Example 1** It is a very well-thought-out process and was quite enjoyable to participate in. The researchers may want to bear in mind that everyone interprets emotions differently, though. Also, the acting was fantastic! **Example 2** The goal of the study is extremely clear, so good job with that! Coming up with clear, direct research questions and plans is hard. The recording was a bit echo-ey, but my main issue with the study is the fact that the video is long and we read the script in English class. There's too much nuance. Since we analyzed this text in English class, you know that it's complicated and people could have different interpretations. I think the ability to pick up on emotions would be better tested if the actors expressed a series of emotions, then the participants had to identify the emotions from a large word bank. If you want to keep the current format, I think there should be a direction somewhere that says "please ONLY write 1–3 sentences" or "please name an emotion here:" depending on how much detail you're looking for—something specific to get more consistent data. I don't remember if this question was on the form, but I'd suggest telling the participant to pay attention to the emotions so they know what to look for, especially since they can't go back to the video link in the form after they submit. Also, changing the 1 to 100 scale to something like 1–20 might make data analysis easier. I like the "how confident are you in your assessment?" question though. And of course, the acting skills were on point.:D

Table 12 (continued)

Dimension of review quality	Examples of students' peer review statements
Explanation	**Study reviewed:** "How Characteristics of Music Affect Emotion" **Reviewer prompt***:* Do the predicted outcomes support the researchers' hypothesis? **Example 1** "Yes, the predicted outcomes support the researchers' hypothesis" **Example 2** "Mostly, but some of the research questions are a little broad, and could be narrowed down a little so the hypotheses make more sense." **Example 3** "There seems to be a bit of a disconnect in the group, because one person put that the outcome variable is the memories triggered by listening to the music. The hypotheses all have differences as well; one says that faster tempos will induce anxiety, another says they will make the listeners happier. All of the individual hypotheses are supported by their respective predicted outcome, but there isn't an overall consensus."

of a given study design, written in response to the prompt "Do the researchers consider possible alternative explanations for the study findings? Which might they be?":

> The researcher does not consider alternate possible outcomes. They think that the air quality and the data received on these sprints will prove their question about air quality and its effects right but in actuality, there are many other explanations that should be considered.

This student's review offers a specific critique (i.e., that the researchers only consider air quality as impacting runners' performance), but not a concrete recommendation for improvement (i.e., suggestions for other variables that could impact runners' performance).

We also found significant differences between quality dimensions of reviews at the post-test, $F(2, 222) = 33.32$, $p < 0.0001$ (Appendix Fig. 4). In particular, students were more likely to make their comments specific to the study proposal ($M = 2.95$, $SD = 1.34$) than they were to give recommendations ($M = 1.40$, $SD = 0.94$), $p < 0.0001$, $d = 1.34$). They were also more likely to give explanations for their evaluations of the studies ($M = 2.48$, $SD = 1.25$) than they were to give recommendations ($p < 0.0001$, $d = 0.98$). Both of these differences had high practical significance. Finally, students' reviews were more likely to be specific than they were to give an explanation, although this finding was marginally significant, and had low to medium practical significance ($p = 0.045$, $d = 0.363$).

An example of a post-test review that demonstrates these two findings is the following comment, made in response to the prompt, "Does the study respect participants' privacy, health, and effort? Explain your reasoning":

> I believe so; however, they are testing the participant's health in extreme pollution levels so it may harm them in vain of the experiment.

This review offers a critique specific to the study (i.e., in mentioning the study's variables and context), and an explanation for why the study design does not quite address the criterion of respecting participants' health (i.e., because it involves exposure to extreme pollution). However, this review does not offer a recommendation for improvement (e.g., by suggesting that the study designers conduct a natural experiment rather than an intervention-based one).

Qualitative description of students' study designs and peer reviews

To complement the quantitative findings described above for RQ1 and RQ2, this section offers a qualitative description of the quality dimensions of students' study designs and peer reviews. Specifically, we describe students' reasoning about and critique of: the research implications of a study; the proposed study's contributions to existing research; the definition and operationalization of variables; potential confounding variables; the navigation of research ethics; and the optimization of research participants' experiences.

Identifying the personal and broader implications of research

As many of students' research foci stemmed from their personal experiences (e.g., how to succeed in school), they tended to justify their studies in terms of implications for other youth like them (e.g., improving test performance). For example, one team designed a study to determine the time at which people perform optimally on memory, typing, and reaction time tests. They justified the importance of this study as follows:

> By seeing what time someone does the best, one can organize their tasks and activities to be based off the time they do the best. People can do their homework at say 3PM (if they do the best then) and see if that helps with them not being distracted.

Students also demonstrated abilities to consider applications of their research questions beyond themselves. One team, for instance, proposed studying the relationship between personality type, stress, and memory, a theme that could easily have translated to explaining test performance. Instead, this team described implications of their research for testimony in criminal cases. As they explained: "witnesses may be stressed because of anxiety when giving testimonies, so officials will be able to know to what degree they should believe the witness and arrive at a truer conclusion." These examples show students' abilities to articulate both the personal and broader implications of anticipated findings, which is key to ensuring the relevance and applicability of research efforts.

Table 13 Mean scores on individual quality dimensions of the reviews students generated in-unit, and on the pre and posttests

Review quality dimension	Mean (SD), range			Pre to post t-test
	In-unit (N = 147)	Pre (N = 18)	Post (N = 75)	
Overall	8.84 (3.19) 0–15	7.06 (2.98) 1–13	6.83 (3.03) 0–13	t(26.11) = .29, p = .77, d = .08
Gives recommendation	1.15 (1.07) 0–4	1.67 (1.01) 0–4	1.40 (.94) 0–3	t(24.34) = 1.00, p = .33, d = .28
Specificity	3.91 (1.37) 0–6	2.94 (1.30) 0–5	2.95 (1.34) 0–5	t(26.38) = -.01, p = .99, d = -.01
Explanation	3.78 (1.42) 0–6	2.44 (1.42) 0–5	2.48 (1.25) 0–5	t(23.63) = -.10, p = .92, d = -.03
Identifies fault	2.37 (1.25) 0–6	2.33 (.89) 1–4	2.08 (.97) 0–5	t(27.12) = 1.05, p = .30, d = .27

Building on existing research

Students' showed varying abilities to build their studies upon existing literature. Given broad criteria for what background research would count, some students identified published, peer reviewed articles, while others relied on news articles and blogs relevant to their topics. Students also showed varying attempts to draw upon prior research to inform their own study designs. These attempts varied from brief mentions of other studies' research questions (e.g., "This study investigated how a traumatic event in one's life could affect their risk-taking behavior."), to providing cursory summaries of other studies' findings (e.g., "It's best to listen to music without words (preferably classical music). However, it does depend on people's tastes.") to reflecting upon those studies' findings. For example:

> The greatest reduction in stress is achieved when listening to music in the presence of others, or alone and for the purpose of relaxing. I wonder what music they used in the study? Based on the wording, I'd assume that the participants got to choose, but it really could be either way [...]. I didn't expect that listening [to] music in the presence of others would reduce stress! (Appendix Table 1)

While many responses showed room for growth, their abilities to connect their own questions to existing research demonstrated their potential to recognize their role in advancing scientific knowledge, which was a central lesson of our curriculum.

Defining variables and aligning methods

Aligning methods with hypotheses is central to creating robust experiments, but was also among the most challenging aspects for students. Student teams that used or adapted the existing studies, surveys, and tasks available in MindHive seemed to experience less issues. Meanwhile, students who sought to answer questions that could not be answered by building on existing resources appeared to struggle.

As an example, one team asked "Does art inspire people to create social change?" These students proposed using tasks and surveys to measure participants' stress, emotion, and mindfulness before and after viewing a series of artworks created by environmentally conscious artists. Their goal was to understand how these measures change before and after participants "learn about how certain artists are impacting the environment: either through the visual aspects of their work itself or if the money received from the work goes to a bigger cause, thus helping the environment."

One peer reviewer commented on the challenge of defining the variables identified ("How exactly will you rate the change in people's inspiration?"). Another pointed out that "participants […] interpret art in their own way, which could differ dramatically from the intended message of the artist." This same reviewer remarked that "some of the tasks in the study also don't seem well connected to the question asked. The research question also doesn't stay consistent, it goes from social change to environmentalism."

This example demonstrates both students' abilities to identify and critique misalignments between a study's methods and research questions; as well as their struggle to implement these alignments in their own designs—a feeling that even experts can appreciate. This example also indicates the value of the availability of MindHive's examples and assets for supporting students' inquiry. Currently, these are largely suited for research that involves questions about decision making, perception, and reaction time. However, future

development might explore ways to support students' various research interests through the provision of more diverse examples and instruments.

Handling confounds in experimental design

Students' abilities to reason about the limitations and potential confounds of study designs varied widely (see examples in Table 11). For instance, one student team asked "how do different types of music affect one's stress level?" This team proposed administering a survey to measure participants' perceived stress before and after listening to pop music played from a linked YouTube video. Whereas some peer reviewers found no issue with this study design, others pointed out that musical preference may be a potential hidden variable. One reviewer suggested addressing this by collecting additional information about participants' musical preferences:

> I might suggest considering adding in a survey about preferred genres of music to help account for the possibility that music taste changes which type of music relieves stress so that you can still make a reasonable conclusion even if there isn't one genre that applies to the majority of the participants.

Another reviewer commented:

> The study design is well done. However, it would be great if a personal choice option was added into the study as one of the independent variables. I feel there could be some misleading evidence from the study otherwise as people will be performing better on their preferred genres rather than simply due to the genre.

Other examples show the advantage of the personal relevance of human brain and behavior research for students' reasoning about the validity of study designs. For instance, on the pre and posttest item regarding how to design a study to investigate the impacts of contaminated drinking water on students' academic performance, students drew on contemporary events (i.e., the COVID-19 pandemic) to illustrate possible hidden variables ("There could have been a virus or other sickness that happened to emerge at the same time as the study was conducted." and "Other external causes that could affect performance such as a pandemic."). Students also drew on other familiar experiences with school. As this student wrote:

> We might also be seeing that the students who are drinking contaminated water are being forced to stay home because they are sick from the water and that is what is affecting their performance in school rather than the water's effect on them mentally. There might also be the explanation that the students who are drinking contaminated water come from lower-income families and therefore were already not focusing in school because they had outside pressures already and the water played no actual role.

Together, these examples illustrate how students used their everyday understanding to reason about confounds. In particular, they show how a focus on human brain and behavior research allowed students to draw on familiar experiences to reason about the impacts of contextual factors on the validity of potential findings, a task that may have been more difficult in a domain about which they had less personal knowledge.

Reasoning about research ethics

Students approached research ethics and participant experience in different ways. In one review of a peer's study, "Has the Pandemic Changed the Environmental Impact of People's Eating Habits?", a student reviewer commented on the need for researchers to be aware of the impacts of their tasks on different participants, and on the need to anticipate the variety of participants' experiences to ensure the validity of data. In response to the prompt: "Does the study respect participants' privacy, health, and effort? Explain your reasoning." this student replied:

> Yes based on the proposal, however I rate this 4/5 stars because while they did nothing wrong, they did not account for eating disorders (EDs) or how they may have changed/developed over quarantine. EDs are a serious and personal issue that affects many people and this topic of food patterns may be incredibly triggering. Or, if someone with an ED participates without being triggered, their data may create an outlier or not support the common trends of the data found.

Research ethics was also an issue in the studies that students designed at the pre and posttest. Students were asked to propose a study based on this prompt: "Pollution in the local river has been mapped over many years. How could scientists study how the river's water quality impacts students' school performance?" Without commenting on the ethics of their choice, several students proposed observing the impacts of manipulating the quality of participants' drinking water (e.g., "give clean water and not great water to different groups and see the difference in student performance."). As well, few students proposed using historical/existing data to investigate this question, which demonstrates students' limited understanding of the variety of possible research methods available to them.

These students' challenges in reasoning about research ethics may be due to the nature of the study context given on the pre and posttests, which contrasted with the contexts for which they designed during the unit. We further discuss this possibility in the Discussion section.

Optimizing the participant experience

Important considerations for the validity and reliability of the data collected from a study include ensuring clear instructions and feasible tasks that do not make unreasonable demands of participants. Because peer reviewers could examine their peers' MindHive studies alongside their written proposals, we noticed that students were able to provide detailed and constructive comments on usability issues. For example, in reviewing the study "What impact does wearing a mask have on high schoolers' ability to gauge emotion?", one reviewer wrote: "I'd suggest telling the participant to pay attention to the emotions so they know what to look for, especially since they can't go back to the video link in the form after they submit."

We also observed students' designing studies that balanced their research goals with a sensitivity to the participant experience. One student team closed their MindHive study with a question that asked participants to comment on their experience completing the study. Meanwhile, another team described:

> While we want to gather data from participants over an extended period of time to see the long-term effects of music on students, we are limited and cannot do so without making the study needlessly tedious and stretched out

Together, these examples illustrate the advantage of engaging with MindHive studies alongside written study proposals for making usability issues more salient to both designers and reviewers.

RQ3 what are students' experiences participating in open science?

In terms of their experiences participating in study design and peer review, students generally rated the helpfulness of their peers' reviews highly (M = 5.24, SD, 1.38, median = 6), with 78.6% of the 84 students who responded indicating that they *Somewhat Agreed* to *Strongly Agreed* that their peers provided helpful feedback that improved their study, and 9.5% saying that they *Somewhat Disagreed* to *Strongly Disagreed* (Fig. 1). Regarding trust in the reviewer, 10.1% of students said that they disagreed with the peer reviews they received (M = 2.98, SD = 1.41, median = 3), while less than half (48.6% of 84 students) indicated that they did not disagree with their peers' reviews (Fig. 2). This finding suggests that almost half of students remained somewhat skeptical of their peers' reviews.

In their open-ended responses, students reported various benefits gained from their participation in the peer review process (Table 14). In particular, students described how reviewing their peers' work allowed them to identify strategies they could use to improve their own proposals. Students additionally commented on learning strategies for effectively engaging in the process of peer review. Importantly, students also noted the value of peer review as a form of collaboration between study designers and reviewers, which is especially characteristic of an open science model.

In their interviews, teachers reported that their students found the experience of participating in MindHive to be novel and compelling. They appreciated the opportunity for students to engage in each step of the research design process. As one teacher noted: "I think that [students] did benefit from the creative challenge of creating the experiments that they came up with." One student interviewed remarked that engaging in the unit helped to convey "the amount of work and time [that] gets put into creating a study." Another remarked that the experience was valuable for helping them "to think critically, which is really important throughout science and life as a whole… just being able to again delve beneath the surface of a certain question…. and then also just seeing how asking a question can develop into this huge research study."

Students also reported valuing the open science process more specifically. Teachers appreciated the chance for students, through their interaction with mentors, and by participating in real studies, to connect with scientists and other professionals engaged in similar work. Meanwhile, students commented on how "it was really nice to have such easy access to, like other people's work, and also the ability to have, like so many people respond to your project to get your views like almost instantly."

Discussion

Most citizen science projects do not involve the general public nor K-12 students in the study ideation and peer review stages of inquiry. We worked with 6 classrooms across 3 high schools to implement and test an open science, citizen science unit themed around human brain and behavior research. Students designed their own studies and then peer-reviewed proposals of their classmates, and of students at different schools. Our analysis

explored students' research interests, abilities, and experiences with study design and peer review. Specifically, we examined (RQ1) the focal topics of studies students designed and how their study design abilities changed from pre- to post-unit; (RQ2) the kinds of peer reviews students generated, and how their review abilities changed from pre- to post-unit; and (RQ3) the value that students perceived in open science by participating in one example of an open science program, MindHive.

Our complementary quantitative and qualitative analyses showed that while students' overall study design abilities did not change from pre to post (RQ1b, RQ1c), there were notable pre-post changes in their abilities within specific dimensions of study quality. For instance, students improved at justifying the importance of a research study, which indicates an increased ability to perceive and articulate the real-world value of research. However, students *decreased* in their abilities to identify confounds and limitations from the pre to the posttest.

Students also struggled to align their methods to their hypotheses in their in-unit study designs, which confirms existing research on students' challenges with experimental design (Woolley et al., 2018). Interestingly, students did not struggle to align methods with hypotheses in the pre and posttests. Students' abilities to navigate research ethics also differed between the in-unit vs. pre/posttests, but in the opposite direction: Among the study quality dimensions we analyzed, students scored the lowest on the *Research Ethics* dimension based on their pre and posttest responses, but the highest on this dimension based on their team-generated studies during the unit.

There are two potential, non-mutually exclusive reasons for these findings. First, our assessment may have been inadequate in capturing changes in students' study design abilities; and second, our curriculum could have better supported students in learning to design studies. These issues are especially important to consider given that these students were simultaneously developing knowledge and abilities in research, as well as in the science of human behavior. Notable in this regard, and as highlighted by our qualitative analysis of students' work, is the disjunct between the study topics generated by students during the unit (RQ1a)—which mostly involved online tasks and variables familiar to students (e.g., genre of background music, study strategies)—and the topic that was the focus of the pre/post study design item (the impacts of water contamination on student performance). It may have been challenging for students, who, during the unit, learned and practiced applying research methods in the context of personally relevant research foci, to then be asked to apply these methods to research contexts outside of their everyday experiences. Had the research foci of our pre/post assessment been more similar to students' research foci during the unit, this assessment may have better detected changes in students' abilities. This observation reflects broader challenges in the design of assessments that are both adequately aligned to learning tasks, and that adequately capture the transfer of learning (Harris et al., 2019; Tiruneh et al., 2018). Moreover, we might expect students to have more refined research and peer review abilities with their greater conceptual grasp of the subject domain. Such a hypothesis is in line with research on the relationship between domain knowledge and skills (Huang et al., 2017).

These findings suggest the importance of understanding students' curiosities in classroom-based citizen science contexts (RQ1a), and particularly in contexts that place students as agents in conceiving of and carrying out research. Knowing what makes students curious can allow us to support, validate, and build on the range of their interests. For instance, we might incorporate further examples of scientist-created studies that illustrate diverse research questions, methodological approaches, and instruments, which students can adapt for their own studies. We might also design supports for students to better

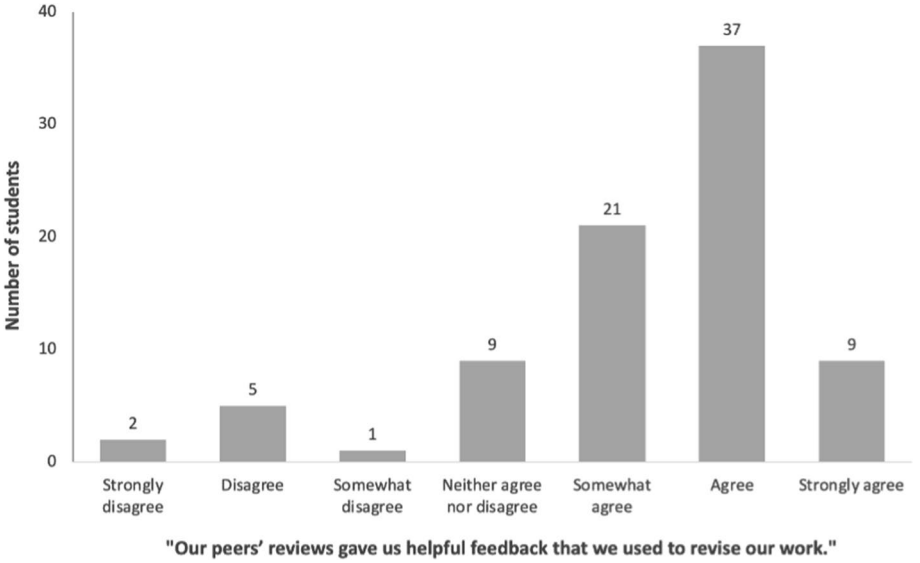

Fig. 1 Distribution of student ratings of "Our peers' reviews gave us helpful feedback that we used to revise our work"

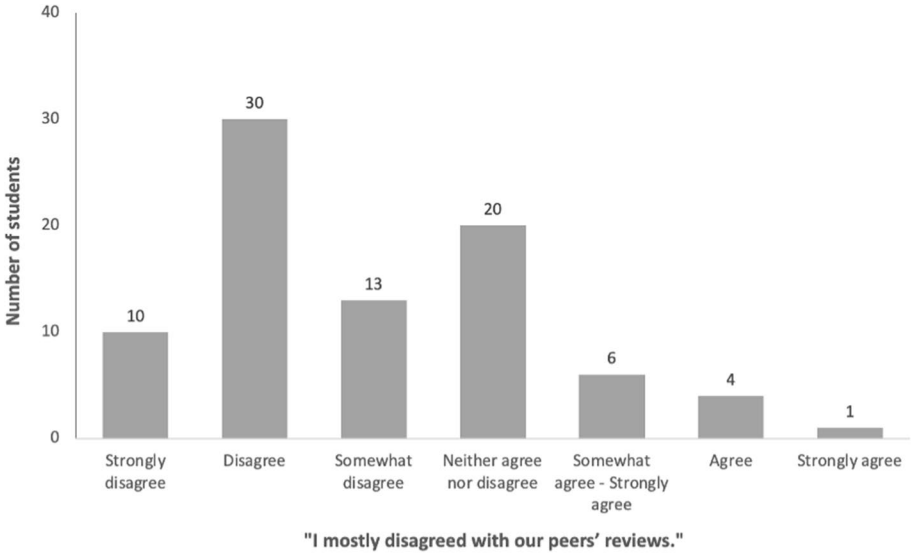

Fig. 2 Distribution of student ratings of "I mostly disagreed with our peers' reviews"

articulate their interests into research questions that are of broader importance. Finally, we might guide their abilities to abstract and transfer their study design skills between more or less similar research contexts, abilities that are important to scientific literacy.

Table 14 Student-reported benefits of their participation in the peer review process

Benefit	Description	Example	Frequency
About study design			
Strategies to apply to their own project	Identifies ways to improve their own study	We learned that one of the aspects of our study was confusing and hard to follow, as it required the user to start a timer and never identified when to start it in our survey. We went back and changed this to improve our study	19
Process of study design	Reflects on the experience of designing a study	I learned that as the study developer it is easy to get caught up in the research and overlook important details	4
About peer review			
Peer Review Process	Reflects on the experience of peer review and/or on particular approaches to conducting a review	I learned about how to incorporate other people's ideas into my own studies and how to analyze the structure of other studies while providing constructive feedback	18
Challenge of giving a good review	Recognizes the difficulty of giving a helpful review	I learned how hard it is to give good critique that is constructive and helpful	3
General			
Value of collaboration	Recognizes the contribution of peer reviewers to improving a study, that reviewers will interpret the same thing differently from one another, and from the study designer, and the value of different perspectives	I learned that it helpful to get an outside opinion on your work because they can pick up things that you never noticed	13
Importance of communication	Recognizes the importance of clear communication	I learned that we should elaborate on instructions so everyone can understand them	4
About another topic	Learned about the topic of the study they reviewed	I learned about the topic of the study I reviewed, which I found pretty interesting	3
Other			
Vague	The statement is not specific enough to be categorized	I learnt the true value of the peer review process while in the peer review process	6
Nothing	States that nothing was learned from the review process	To be completely honest, we had already gone through this process before, and thus I did not learn much from this process	4

Despite these mixed results in researcher ratings of student study quality, teachers valued the opportunity for students to grapple with issues of ethics in human subjects research. As one teacher noted, students "were exposed to the considerations that go into human testing, which is something they've never been exposed to…" This finding suggests the value of offering students' meaningful contexts in which to learn about topics in ethics, which are becoming increasingly prevalent across nearly all STEM fields, including human–computer interaction, engineering, user experience design, and data science (Gasparich & Wimmers, 2014).

Regarding RQ2a (students' peer review quality), the fact that students were generally positive in their peer reviews resonates with literature on the quality of peer assessment, which finds that students tend to be superficial and positive in their peer assessments (Hovardas et al., 2014). As with our analysis of students' study design abilities, we found that students' review abilities were consistent from pre to post (RQ2b). Similar to our findings on students' experimental design abilities, this observation may be explained by students' need for better in-unit support for learning to effectively review studies, and/or to shortcomings of our assessments to capture changes in students' review abilities. More specifically, students may have been challenged to review studies in a research context (i.e., the impact of air pollution on outdoor sports performance) that differed greatly from the contexts for which they designed their own studies during the unit.

In terms of specific dimensions of review quality, it is notable that both in students' in-unit reviews, and in their pre and posttest reviews, students were better able to be specific in their comments and to provide explanations for their evaluations, than they were able to produce concrete recommendations for improvement. This finding may reflect the relative difficulty of making a recommendation, which requires generating new ideas, compared to specificity and explanation, which only require interpreting and communicating one's interpretations.

Together, these trends in students' peer reviews likely reflect their still developing abilities. They suggest a need to support students in abstracting and applying their review abilities to research contexts outside of the ones that are close to their personal experiences. At the same time, these findings resonate with research on the role of expertise in professional scientific peer review, in which individual reviewers often represent different degrees of familiarity with either the methods or the topic of the study under review. While a reviewer may be less able to provide specific recommendations for studies outside of their domain area, close alignment in domain expertise between reviewer and study designer may also lead to harsher—albeit more specific—reviews (Gallo et al., 2016). Ultimately, there is an opportunity for an open science curriculum to demonstrate to students the advantages of including multiple perspectives in peer review (Lee et al., 2013; Resnik & Elmore, 2016).

Regarding RQ3 (how students value peer review), students generally found it rewarding to evaluate their peers' proposals, and identified various personal benefits gained from the experience. However, the review process was not without challenges. In line with our *review quality* rubric, students whom we interviewed expressed wanting peer reviews to recommend specific actions that would improve their work, to comment on the substance rather than to simply re-word their proposals, and to explain the reasons for their evaluations. These findings resonate with Hattie and Timperley (2007), who found that the most helpful reviews are those that provide specific recommendations for improvement and that explain why they offer these recommendations. Additionally, the fact that almost half

of students reported disagreeing to some extent with their peers' reviews resonates with other research on students' mistrust of peer assessment compared to teacher assessment (Anker-Hansen & Andrée, 2019). It is likely that there were important differences between classrooms with respect to both communication style and expectations of the amount and quality of feedback. These differences in expectations may explain some students' disappointment with the cursory and generic peer reviews they received. While anecdotal, this observation highlights the importance of adequate training to ensure that all students understand the role and expectations of peer review.

Limitations

One limitation of this study is the small number of participants (n = 20) who completed the pretest. Three of the 6 classes were unable to participate in the pretest, in part due to COVID-19 related restrictions on classroom research. As a result, our pre and post samples differed greatly in size. Future work replicating these findings will be needed that include larger, matched samples.

A second limitation of our study is that we did not consider the various affective qualities of peer reviews. Given the existing literature on the role of affect management and communication strategies such as hedging (e.g., including words like "probably" and "maybe," examples of which can be informally observed in Table 10) in impactful feedback (Wu & Schunn, 2020a), future research on the affective dimensions of peer reviews would provide a fuller picture of students' review abilities.

A third limitation is that, due to scheduling issues, most students did not have the chance to revise their studies. As such, we are unable to see how peers' reviews might have impacted students' study designs. Other research suggests that even when given the time, students tend not to use their peers' feedback to improve their work, either because they do not perceive the feedback to be useful, because they lack the strategies necessary to make use of feedback (Jonsson, 2013), or because of the multiple social factors that influence the uptake of peer feedback (Wu & Schunn, 2020a, 2020b). Dialogue between reviewers and review recipients is critical for encouraging the uptake of feedback; but, as we have experienced, it is challenging to create such opportunities in a classroom context (Tsivitanidou et al., 2012).

A fourth limitation is that, due to the status of our platform's data logging capabilities at the time of this study, we were unable to link in-unit and pre-post responses, and as such, we were unable to compare students' experiences based on individual differences. For example, it is possible that students with different pretest scores followed different trajectories during, and after the unit. Indeed, other research suggests that students' with different ability levels benefit differently from curriculum interventions, and even from being grouped with peers of the same or different ability levels (Kyza et al., 2011; White & Frederiksen, 1998). Continued development of our platform will allow such research to be conducted in future curriculum implementations.

A fifth limitation is that—due to COVID-related restrictions on in-person classroom research—we were unable to systematically observe how teachers and science mentors supported students' study designs and peer reviews. Future research might attend to the roles that teachers' and mentors' expertise contributes to similar classroom-based open science communities.

Conclusion

Participating in open science such as citizen science can improve students' scientific literacy. Yet, student citizen scientists are rarely agents in the full spectrum of scientific inquiry. This study described a citizen science program in which students designed and peer reviewed their own human brain and behavior research. It responds to calls to involve students as co-creators of citizen science efforts, which can ensure that such projects address both their educational and scientific objectives (Gray et al., 2012). By examining the impacts of our program on students' study design and peer review abilities, this study illustrates how citizen science can meaningfully engage students in the dialogue among authors and reviewers to produce and validate scientific knowledge.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11251-023-09633-9.

Funding National Science Foundation, 1908482, Camillia Matuk.

References

American Association for the Advancement of Science. (1989). *Science for All Americans. A Project 2061 Report on Literacy Goals in Science, Mathematics, and Technology*. Washington, D.C.: AAAS. https://doi.org/10.1177/027046769001000206

American Association for the Advancement of Science. (2011). *Vision and Change in Undergraduate Biology Education: A Call to Action. A Summary of Recommendations Made at a National Conference Organized by the American Association for the Advancement of Science*. Washington, DC: AAAS. https://www.aaas.org/sites/default/files/content_files/VC_report.pdf. Accessed on 9 July 2023

Anker-Hansen, J., & Andrée, M. (2019). Using and rejecting peer feedback in the science classroom: a study of students' negotiations on how to use peer feedback when designing experiments. *In Research in Science & Technological Education, 37*(3), 346–365. https://doi.org/10.1080/02635143.2018.1557628

Bangert-Drowns, R. L., Kulik, C.-L.C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 6 61*(2), 213–238.

Barab, S. A., & Hay, K. E. (2001). Doing science at the elbows of experts: Issues related to the science apprenticeship camp. *Journal of Research in Science Teaching, 38*(1), 70–102.

Beck, C. W., & Blumer, L. S. (2012). Inquiry-based ecology laboratory courses improve student confidence and scientific reasoning skills. *Ecosphere, 3*(12), 112.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., & Wilderman, C. C. (2009). Public Participation in Scientific Research: Defining the field and assessing its potential for Informal Science Education. A CAISE inquiry group report. *Online Submission*. http://files.eric.ed.gov/fulltext/ED519688.pdf

Brossard, D., Lewenstein, B., & Bonney, R. (2005). Scientific knowledge and attitude change: The impact of a citizen science project. *International Journal of Science Education, 27*(9), 1099–1121.

Buchanan, S. C. (2019). Using the hermeneutic phenomenological method to explore the middle school student lived experience of student driven inquiry. *Qualitative and Quantitative Methods in Libraries, Special Issue: School Library Research and Educational Resources, 6*, 61–74.

Bullock, M., Ziegler, A., Weinert, F. E., & Schneider, W. (1999). Scientific reasoning: Developmental and individual differences. In F. E. Weinert & W. Schneider (Eds.), *Individual development from 3 to 12: Findings from the Munich longitudinal study* (pp. 38–54). Cambridge University Press.

Burns, J. C., Okey, J. R., & Wise, K. C. (1985). Development of an integrated process skill test: TIPS II. *Journal of Research in Science Teaching, 22*, 169–177.

Cascio, M. A., Lee, E., Vaudrin, N., & Freedman, D. A. (2019). A team-based approach to open coding: Considerations for creating intercoder consensus. *Field Methods, 31*(2), 116–130.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098–1120.
Christianakis, M. (2010). "I Don't Need Your Help!" peer status, race, and gender during peer writing interactions. *Journal of Literacy Research: JLR, 42*(4), 418–458.
Crawford, B. A. (2012). Moving the essence of inquiry into the classroom: engaging teachers and students in authentic science. In K. C. D. Tan & M. Kim (Eds.), *Issues and challenges in science education research: moving forward* (pp. 25–42). Springer.
D'Costa, A. R., & Schlueter, M. A. (2013). Scaffolded instruction improves student understanding of the scientific method & experimental design. *The American Biology Teacher, 75*, 18–28.
Dasgupta, A. P., Anderson, T. R., & Pelaez, N. (2014). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. *CBE Life Sciences Education, 13*(2), 265–284.
Deemer, E. D., Ogas, J. P., Barr, A. C., Bowdon, R. D., Hall, M. C., Paula, S., Capobianco, B. M., & Lim, S. (2021). Scientific research identity development need not wait until college: examining the motivational impact of a pre-college authentic research experience. *Research in Science Education*. https://doi.org/10.1007/s11165-021-09994-6
Dikker, S., Shevchenko, Y., Burgas, K., Chaloner, K., Sole, M., Yetman-Michaelson, L., Davidesco, I., Martin, R., & Matuk, C. (2022). An online citizen science tool to support students and communities in authentic human brain and behavior science inquiry. *Connected Science Learning*, *4*(2).
Dolan, E., & Grady, J. (2010). Recognizing students' scientific reasoning: A tool for categorizing complexity of reasoning during teaching by inquiry. *Journal of Science Teacher Education, 21*(1), 31–55.
Dolezal, D., Motschnig, R., & Pucher, R. (2018). Peer review as a tool for person-centered learning: computer science education at secondary school level. *Teaching and learning in a digital world.* Springer.
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*. Springer.
Dysthe, O., Lillejord, S., & Wasson, B. (2010). Productive e-feedback in higher education: Two models and some critical issues. *Learning across sites* (pp. 255–270). Routledge.
Earley, M. A. (2014). A synthesis of the literature on research methods education. *Teaching in Higher Education, 19*(3), 242–253.
Ed, P. J., Ed, C. N., & Ed, G. R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. National Academy Press, 2102 Constitutions Avenue, N.W., Lockbox 285, Washington, DC 20055 ($39.95). Tel: 800–624–6242 (Toll Free); 202–334–3313; Web site: http://www.nap.edu.
Etkina, E., Matilsky, T., & Lawrence, M. (2003). Pushing to the edge: Rutgers astrophysics institute motivates talented high school students. *Journal of Research in Science Teaching, 40*(10), 958–985.
Fecher, B., & Friesike, S. (2014). Open science: one term, five schools of thought. *Opening science* (pp. 17–47). Springer.
Fine, & Pryiomka. (2020). Assessing College Readiness through Authentic Student Work: How the City University of New York and the New York Performance Standards Consortium Are …. *Learning Policy Institute*. https://eric.ed.gov/?id=ED606677
Fitzgerald, M., Danaia, L., & McKinnon, D. H. (2019). Barriers inhibiting inquiry-based science teaching and potential solutions: perceptions of positively inclined early adopters. *Research in Science Education, 49*(2), 543–566.
Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions. *In Wiley Series in Probability and Statistics*. https://doi.org/10.1002/0471445428
Fry, C. V., Cai, X., Zhang, Y., & Wagner, C. S. (2020). Consolidation in a crisis: Patterns of international collaboration in early COVID-19 research. *PLoS ONE, 15*(7), e0236307.
Fuller, R. G. (2002). The second career—science education. *A love of discovery* (pp. 303–304). Springer. https://doi.org/10.1007/978-94-007-0876-1_12
Furtak, E. M., & Penuel, W. R. (2019). Coming to terms: Addressing the persistence of "hands-on" and other reform terminology in the era of science as practice: FURTAKandPENUEL. *Science Education, 103*(1), 167–186.
Gallo, S. A., Sullivan, J. H., & Glisson, S. R. (2016). The influence of peer reviewer expertise on the evaluation of research funding applications. *PLoS ONE, 11*(10), e0165147.
Gasparich, G. E., & Wimmers, L. (2014). Integration of ethics across the curriculum: from first year through senior seminar. *Journal of Microbiology & Biology Education: JMBE, 15*(2), 218–223.
Gaynor, J. W. (2020). Peer review in the classroom: Student perceptions, peer feedback quality and the role of assessment. *Assessment & Evaluation in Higher Education, 45*(5), 758–775.

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment, 1*(1), 20.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315.

Gray, S. A., Nicosia, K., & Jordan, R. C. (2012). Lessons learned from citizen science in the classroom. A response to the future of citizen science. *Democracy and Education, 20*(2), 14.

Griffith, A. B. (2007). Semester-long engagement in science inquiry improves students' understanding of experimental design. *Teaching Issues and Experiments in Ecology*, *5*(25), 1–27.

Haklay, M. M., Dörler, D., Heigl, F., Manzoni, M., Hecker, S., & Vohland, K. (2021a). What is citizen science? In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The science of citizen science* (pp. 13–33). Springer. Springer. https://doi.org/10.1007/978-3-030-58278-4

Haklay, M. M., Fraisl, D., Greshake Tzovaras, B., Hecker, S., Gold, M., Hager, G., Ceccaroni, L., Kieslinger, B., Wehn, U., Woods, S., Nold, C., Balázs, B., Mazzonetto, M., Ruefenacht, S., Shanley, L. A., Wagenknecht, K., Motion, A., Sforzi, A., Riemenschneider, D., & Vohland, K. (2021b). Contours of citizen science: A vignette study. *Royal Society Open Science, 8*(8), 202108.

Halonen, J. S., Bosack, T., Clay, S., McCarthy, M., Dunn, D. S., Hill, G. W., IV., McEntarffer, R., Mehrotra, C., Nesmith, R., Weaver, K. A., & Whitlock, K. (2003). A rubric for learning, teaching, and assessing scientific inquiry in psychology. *Teaching of Psychology, 30*(3), 196–208.

Harker, A. R. (2009). Full application of the scientific method in an undergraduate teaching laboratory. *Journal of College Science Teaching, 29*, 97–100.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement Issues and Practice*. https://doi.org/10.1111/emip.12253

Harris, E. M., Dixon, C. G. H., Bird, E. B., & Ballard, H. L. (2020). For science and self: youth interactions with data in community and citizen science. *Journal of the Learning Sciences, 29*(2), 224–263.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Hiebert, S. M. (2007). Teaching simple experimental design to undergraduates: Do your students understand the basics? *Advances in Physiology Education, 31*(1), 82–92.

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 & Education, 71*, 133–152.

Howard, C. D., Barrett, A. F., & Frick, T. W. (2010). Anonymity to promote peer feedback: pre-service teachers' comments in asynchronous computer-mediated communication. *Journal of Educational Computing Research, 43*(1), 89–112.

Huang, P.-S., Peng, S.-L., Chen, H.-C., Tseng, L.-C., & Hsu, L.-C. (2017). The relative influences of domain knowledge and domain-general divergent thinking on scientific creativity and mathematical creativity. *Thinking Skills and Creativity, 25*, 1–9.

Huisman, B., Saab, N., van Driel, J., & van den Broek, P. (2018). Peer feedback on academic writing: undergraduate students' peer feedback role, peer feedback perceptions and essay performance. *In Assessment & Evaluation in Higher Education, 43*(6), 955–968. https://doi.org/10.1080/02602938.2018.1424318

Iandoli, L., & Shen, J. (2021). Towards a Design Framework for Humanized AI. *Proceedings of the 5th HUMANIZE Workshop*. http://ceur-ws.org/Vol-2903/IUI21WS-HUMANIZE-2.pdf

Jenkins, L. L., Walker, R. M., Tenenbaum, Z., Sadler, K. C., & Wissehr, C. (2015). Why the secret of the great smoky mountains institute at tremont should influence science education—connecting people and nature. In M. P. Mueller & D. J. Tippins (Eds.), *EcoJustice, citizen science and youth activism: situated tensions for science education* (pp. 265–279). Springer International Publishing.

Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education, 14*(1), 63–76.

Jordan, R. C., Gray, S. A., Howe, D. V., Brooks, W. R., & Ehrenfeld, J. G. (2011). Knowledge gain and behavioral change in citizen-science programs. *Conservation Biology: THe Journal of the Society for Conservation Biology, 25*(6), 1148–1154.

Kadir, B. A., & Broberg, O. (2021). Human-centered design of work systems in the transition to industry 4.0. *Applied Ergonomics, 92*, 103334.

Kasch, J., Van Rosmalen, P., Henderikx, M., & Kalz, M. (2021). The factor structure of the peer-feedback orientation scale (PFOS): toward a measure for assessing students' peer-feedback dispositions. *In Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2021.1893650

Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science, 39*(3), 387–406.

Kelemen-Finan, J., & Dedova, I. (2014). Vermittlung von Artenkenntnis im Schulunterricht. Ergebnisse einer Befragung von Lehrpersonal in Österreich und bildungspolitische Relevanz. *Naturschutz Und Landschaftsplanung., 46*(7), 219–225.

Kelly, J., Sadeghieh, T., & Adeli, K. (2014). Peer review in scientific publications: Benefits, critiques, & a survival guide. *eJIFCC, 25*(3), 227–243.

Kennedy, K. J., Chan, J. K. S., Fok, P. K., & Yu, W. M. (2008). Forms of assessment and their potential for enhancing learning: Conceptual and cultural issues. *Educational Research for Policy and Practice, 7*(3), 197.

Ketonen, L., Hähkiöniemi, M., Nieminen, P., & Viiri, J. (2020). Pathways through peer assessment: implementing peer assessment in a lower secondary physics classroom. *International Journal of Science and Mathematics Education, 18*(8), 1465–1484.

Killpack, T. L., Fulmer, S. M., Roden, J. A., Dolce, J. L., & Skow, C. D. (2020). Increased scaffolding and inquiry in an introductory biology lab enhance experimental design skills and sense of scientific ability. *Journal of Microbiology & Biology Education: JMBE*. https://doi.org/10.1128/jmbe.v21i2.2143

Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction, 20*(4), 344–348.

Koomen, M. H., Rodriguez, E., Hoffman, A., Petersen, C., & Oberhauser, K. (2018). Authentic science with citizen science and student-driven science fair projects. *Science Education, 102*(3), 593–644.

Kuhn, D., & Dean, D., Jr. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science, 16*(11), 866–870.

Kyza, E. A., Constantinou, C. P., & Spanoudis, G. (2011). Sixth graders' co-construction of explanations of a disturbance in an ecosystem: exploring relationships between grouping, reflective scaffolding, and evidence-based explanations. *In International Journal of Science Education, 33*(18), 2489–2525. https://doi.org/10.1080/09500693.2010.550951

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lave, J., & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge University Press.

Ledgerwood, A. (2018). The preregistration revolution needs to distinguish between predictions and analyses [Review of The preregistration revolution needs to distinguish between predictions and analyses]. *Proceedings of the National Academy of Sciences of the United States of America, 115*(45), E10516–E10517.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *In Journal of the American Society for Information Science and Technology, 64*(1), 2–17. https://doi.org/10.1002/asi.22784

Lee, O., & Campbell, T. (2020). What science and STEM teachers can learn from COVID-19: harnessing data science and computer science through the convergence of multiple STEM subjects. *Journal of Science Teacher Education, 31*(8), 932–944.

Leible, S., Schlager, S., Schubotz, M., & Gipp, B. (2019). A review on blockchain technology and blockchain projects fostering open science. *Frontiers in Blockchain, 2*, 16.

Li, L. (2017). The role of anonymity in peer assessment. *Assessment & Evaluation in Higher Education, 42*(4), 645–656.

Li, L., & Grion, V. (2019). The Power of Giving Feedback and Receiving Feedback in Peer Assessment. *All Ireland Journal of Higher Education*, *11*(2). http://ojs.aishe.org/index.php/aishe-j/article/view/413

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.

Libarkin, J., & Ording, G. (2012). The utility of writing assignments in undergraduate bioscience. *CBE Life Sciences Education, 11*(1), 39–46.

Linn, M. C., Gerard, L., Matuk, C., & McElhaney, K. W. (2016). Science education: From separation to integration. *Review of Research in Education, 40*(1), 529–587.

Loretto, A., DeMartino, S., & Godley, A. (2016). Secondary students' perceptions of peer review of writing. *Research in the Teaching of English, 51*(2), 134–161.

Lu, J., & Law, N. (2012). Online peer assessment: Effects of cognitive and affective feedback. *Instructional Science, 40*(2), 257–275.

Matuk, C., Martin, R., Vasudevan, V., Burgas, K., Chaloner, K., Davidesco, I., Sadhukha, S., Shevchenko, Y., Bumbacher, E., & Dikker, S. (2021). Students learning about science by investigating an unfolding pandemic. *AERA Open, 7*, 23328584211054850.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *The American Psychologist, 70*(6), 487–498.

Meyer, N. J., Scott, S., Strauss, A. L., Nippolt, P. L., Oberhauser, K. S., & Blair, R. B. (2014). Citizen Science as a REAL Environment for Authentic Scientific Inquiry. *Journal of Extension*, *52*(4). http://www.joe.org/joe/2014august/pdf/JOE_v52_4iw3.pdf

National Research Council. (1996). *National science education standards*. National Academy Press.

National Research Council Division of Behavioral and Social Sciences and Education, Board on Science Education, & Committee on a Conceptual Framework for New K-12 Science Education Standards. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. National Academies Press.

National Science Foundation. (1996). *Shaping the future: new expectations for undergraduate education in science, mathematics, engineering, and technology*. Directorate for Education and Human Resources.

NGSS Lead States. (2013). *Next generation science standards: for states*. National Academies Press.

Noonan, B., & Randy Duncan, C. (2005). Peer and self-assessment in high schools. *Practical Assessment, Research, and Evaluation, 10*(1), 17.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America, 115*(11), 2600–2606.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 7*(6), 657–660.

Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science, 349*(6251), 4716.

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science, 328*, 463–466.

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 human and social conditions in assessment* (pp. 247–266). New York: 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, 44*(8), 1253–1278.

Perlman, B., & McCann, L. I. (2005). Undergraduate research experiences in psychology: A national study of courses and curricula. *Teaching of Psychology, 32*(1), 5–14.

Peterson, S. (2003). Peer response and students' revisions of their narrative writing. *L1 Educational Studies in Language and Literature, 3*(3), 239–272.

Phillips, T., Porticella, N., Constas, M., & Bonney, R. (2018). A framework for articulating and measuring individual learning outcomes from participation in citizen science. *Citizen Science Theory and Practice, 3*(2), 3.

Race, P. (2014). *Making learning happen: A guide for post-compulsory education*. SAGE.

Rempel, D. (2020). Scientific collaboration during the COVID-19 pandemic: N95DECON.org. *Annals of Work Exposures and Health, 64*(8), 775–777.

Resnik, D. B., & Elmore, S. A. (2016). Ensuring the quality, fairness, and integrity of journal peer review: A possible role of editors. *Science and Engineering Ethics, 22*(1), 169–188.

Robnett, R. D., Chemers, M. M., & Zurbriggen, E. L. (2015). Longitudinal associations among undergraduates' research experience, self-efficacy, and identity. *In Journal of Research in Science Teaching, 52*(6), 847–867. https://doi.org/10.1002/tea.21221

Roche, J., Bell, L., Galvão, C., Golumbic, Y. N., Kloetzer, L., Knoben, N., Laakso, M., Lorke, J., Mannion, G., Massetti, L., Mauchline, A., Pata, K., Ruck, A., Taraba, P., & Winter, S. (2020). Citizen science, education, and learning: challenges and opportunities. *Frontiers in Sociology, 5*, 613814.

Sackstein, S. (2017). *Peer feedback in the classroom: Empowering students to be the experts*. ASCD.

Salangam, J. (2007). *The impact of a prelaboratory discussion on non-biology majors' abilities to plan scientific inquiry [Masters]*. California State University.

Saunders, M. E., Roger, E., Geary, W. L., Meredith, F., Welbourne, D. J., Bako, A., Canavan, E., Herro, F., Herron, C., Hung, O., Kunstler, M., Lin, J., Ludlow, N., Paton, M., Salt, S., Simpson, T., Wang, A., Zimmerman, N., Drews, K. B., … Moles, A. T. (2018). Citizen science in schools: Engaging students in research on urban habitat for pollinators. *Austral Ecology, 43*(6), 635–642.

Schunn, C., Godley, A., & DeMartino, S. (2016). The reliability and validity of peer review of writing in high school AP English classes. *Journal of Adolescent & Adult Literacy: A Journal from the International Reading Association, 60*(1), 13–23.

Scott, A. (2007). Peer review and the relevance of science. *Futures, 39*(7), 827–845.

Shen, B., Bai, B., & Xue, W. (2020). The effects of peer assessment on learner autonomy: An empirical study in a Chinese college English writing class. *In Studies in Educational Evaluation, 64*, 100821. https://doi.org/10.1016/j.stueduc.2019.100821

Shi, J., Power, J., & Klymkowsky, M. (2011). Revealing student thinking about experimental design and the roles of control experiments. *International Journal for the Scholarship of Teaching and Learning*. https://doi.org/10.20429/ijsotl.2011.050208

Shneiderman, B. (2020). Human-centered artificial intelligence: reliable, safe & trustworthy. *International Journal of Human-Computer Interaction, 36*(6), 495–504.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Sluijsmans, D. (2002). Establishing learning effects with integrated peer assessment tasks. *The Higher Education Academy*. https://www.researchgate.net/profile/Dominique-Sluijsmans/publication/237794992_Establishing_learning_effects_with_integrated_peer_assessment_tasks/links/54bf5d9d0cf2f6bf4e04e68d/Establishing-learning-effects-with-integrated-peer-assessment-tasks.pdf

Strijbos, J.-W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*(4), 291–303.

Sujarittham, T., Tanamatayarat, J., & Kittiravechote, A. (2019). Investigating the students' experimental design ability toward guided inquiry based learning in the physics laboratory course. *The Turkish Online Journal of Educational Technology, 18*(1), 63–69.

Tasker, T. Q., & Herrenkohl, L. R. (2016). Using peer feedback to improve students' scientific inquiry. *Journal of Science Teacher Education, 27*(1), 35–59.

Tiruneh, D. T., Gu, X., De Cock, M., & Elen, J. (2018). Systematic design of domain-specific instruction on near and far transfer of critical thinking skills. *International Journal of Educational Research*. https://doi.org/10.1016/j.ijer.2017.10.005

Tobin, K. G., & Capie, W. (1982). Relationships between classroom process variables and middle-school science achievement. *Journal of Educational Psychology, 74*, 441.

Topping, K. J. (2009). Peer assessment. *Theory into Practice, 48*(1), 20–27.

Tsai, C.-C., Lin, S. S. J., & Yuan, S.-M. (2002). Developing science activities through a networked peer assessment system. *Computers & Education, 38*(1), 241–252.

Tsivitanidou, O. E., Zacharia, Z. C., & Hovardas, T. (2011). Investigating secondary school students' unmediated peer assessment skills. *Learning and Instruction, 21*(4), 506–519.

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.

Tsui, A. B. M., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing, 9*(2), 147–170.

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: The role of interpersonal variables and conceptions. *Learning and Instruction, 20*(4), 280–290.

van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*(4), 270–279.

Van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—a discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2–12.

Walters. (2020). 2.2 Research designs in psychology. In S. Walters, C. Stangor, & J. Walinga (Eds.), *Psychology-1st Canadian edition.* Thompson Rivers University.

Wanner, T., & Palmer, E. (2018). Formative self-and peer assessment for improved student learning: The crucial factors of design, teacher participation and feedback. *Assessment & Evaluation in Higher Education, 43*(7), 1032–1047.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: making science accessible to all students. *Cognition and Instruction, 16*(1), 3–118.

Woolley, J. S., Deal, A. M., Green, J., Hathenbruck, F., Kurtz, S. A., Park, T. K. H., Pollock, S. V., Bryant Transtrum, M., & Jensen, J. L. (2018). Undergraduate students demonstrate common false scientific reasoning strategies. *In Thinking Skills and Creativity, 27*, 101–113. https://doi.org/10.1016/j.tsc.2017.12.004

Wu, Y., & Schunn, C. D. (2020a). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology, 60*, 101826.

Wu, Y., & Schunn, C. D. (2020b). When peers agree, do students listen? The central role of feedback quality and feedback frequency in determining uptake of feedback. *Contemporary Educational Psychology, 62*, 101897.

Xiao, Y., & Lucking, R. (2008). The impact of two types of peer assessment on students' performance and satisfaction within a Wiki environment. *The Internet and Higher Education, 11*(3), 186–193.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179–200.
Yu, F.-Y., & Sung, S. (2016). A mixed methods approach to the assessor's targeting behavior during online peer assessment: Effects of anonymity and underlying reasons. *Interactive Learning Environments, 24*(7), 1674–1691.
Zou, Y., Schunn, C. D., Wang, Y., & Zhang, F. (2018). Student attitudes that predict participation in peer assessment. *Assessment & Evaluation in Higher Education, 43*(5), 800–811.

Authors and Affiliations

Camillia Matuk[1] · Lucy Yetman-Michaelson[1] · Rebecca Martin[2] · Veena Vasudevan[3] · Kim Burgas[4] · Ido Davidesco[5] · Yury Shevchenko[6] · Kim Chaloner[7] · Suzanne Dikker[1]

✉ Camillia Matuk
cmatuk@nyu.edu

1 New York University, New York, USA
2 University of Pennsylvania, Philadelphia, USA
3 University of Pittsburgh, Pittsburgh, USA
4 Independent Researcher, New York, USA
5 University of Connecticut, Mansfield, USA
6 University of Konstanz, Konstanz, Germany
7 Grace Church School, New York, USA