Spring 4-20-2022

Designing Experiments: The Impact of Peer Review Structure on Organic Chemistry Students' Experimental Designs

Katie Patterson
University of Nebraska-Lincoln, katiepatterson3327@gmail.com

DESIGNING EXPERIMENTS: THE IMPACT OF PEER REVIEW STRUCTURE ON ORGANIC CHEMISTRY STUDENTS' EXPERIMENTAL DESIGNS

by

Katie E. Patterson

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Chemistry

Under the Supervision of Professor Alena Moon

Lincoln, Nebraska

May, 2022

DESIGNING EXPERIMENTS: THE IMPACT OF PEER REVIEW STRUCTURE ON ORGANIC CHEMISTRY STUDENTS' EXPERIMENTAL DESIGNS

Katie E. Patterson, M.S.

University of Nebraska, 2022

Advisor: Alena Moon

To better understand how peer review can be used to support students when designing experiments, the current thesis examined how the structure of the peer review template affects the kinds of feedback students give and the revisions that they make. I utilized a quasi-experimental design to investigate the effects that the peer review template had on the outcome of an experimental design task. A sample of 195 students enrolled in an Organic Chemistry I course participated in the study. The students were divided into two groups, one of which was given a scaffolded peer review template and the other was given a non-scaffolded peer review template. The students in both groups turned in an initial draft, then peer reviewed three students while also receiving feedback from three peers, and then turned in a final draft. I categorized students' feedback and scored their initial and final drafts with rubrics. Afterwards, statistical tests were run to determine if there were significant differences in the frequency of feedback students gave and the frequency of revisions students made. Based on the findings, implications for practice and research were offered.

Table of Contents

CHAPTER 1: INTRODUCTION 1

1.1 Background and Need 3

1.2 Purpose of the Study 7

1.3 Research Questions 8

CHAPTER 2: LITERATURE REVIEW 9

2.1 Experimental Design Activities 9

2.2 Peer Review 14

2.4 Social Comparison Theory 17

2.3 Summary 20

CHAPTER 3: METHODS 1

3.1 Participants and Context 1

3.2 Study Design 2

3.3 Task Development 5

3.4 Data Analysis 8

3.4.1 Experimental Design Analysis 9

3.4.2 Peer Review Analysis 14

3.4.3 Statistical Analysis 19

3.4.3.1 Research Question One: How does the structure of the peer review template impact the feedback students give for an experimental design task? 19

3.4.3.2 Research Questions Two: How does the structure of the peer review template impact students' revisions for an experimental design task? 20
CHAPTER 4: RESULTS 22
4.1 Research Question One: How does the structure of the peer review template impact the feedback students give for an experimental design task? 22
4.1.1 Chi-Square Test of Independence 22
4.1.2 Adjusted Residuals 23
4.1.3 Post-hoc Testing 24
4.2 Research Question Two: How does the structure of the peer review template impact students' revisions for an experimental design task? 26
4.2.1 Chi-Square Test of Independence 26
4.2.2 Adjusted Residuals 29
4.2.3 Post-hoc Testing 30
4.3 Score Revision 31
CHAPTER 5: DISCUSSION 35
5.1 Implications for Research and Teaching 39
References 41

List of Tables

Table 1: Hypothesis Scoring Rubric 10

Table 2: Variable Scoring Rubric 11

Table 3: Outcomes And Conclusions Scoring Rubric 12

Table 4: Adapted From Patchan And Schunn (2015) 15

Table 5: Present Study's Peer Review Coding Scheme 15

Table 6: Peer Review Contingency Table 22

Table 7: Peer Review Adjusted Residuals 24

Table 8: Praise Or Summary Contingency Table. 25

Table 9: High Prose Problem Or Solution Contingency Table 26

Table 10: Revisions Contingency Table 27

Table 11: Revisions Adjusted Residuals 29

Table 12: "No Revision" Contingency Table 30

Table 13: "Revision With Score Improvement" Contingency Table 31

Table 14: Scaffolded Student Example 32

Table 15: Non-Scaffolded Student Example 33

List of Figures

Figure 1: Scaffolded Peer Review Template 4

Figure 2: Non-Scaffolded Peer Review Template 4

Figure 3: Experimental Design Task Introduction Page 7

Figure 4: Experimental Design Task Questions 8

Figure 5: Distribution Of Peer Review Given 23

Figure 6: Non-Scaffolded Group Initial And Final Scores 28

Figure 7: Scaffolded Group Initial And Final Score 28

CHAPTER 1: INTRODUCTION

In 1996, the National Academy of Sciences (NAS) published the National Science Education Standards, establishing that the goal of science education in the United States should be for all students to achieve scientific literacy (Council, Education, Education, & Assessment, 1996). This sought to recognize the importance that scientific literacy plays in people's everyday lives as well as in people's ability to engage in public discourse about science and technology issues. The importance of scientific literacy has been made even more evident over the past two years as the Covid-19 pandemic has wreaked havoc on the United States. Simultaneously, people have watched science occur in real-time as scientists have studied and learned more about the Covid-19 virus. In addition to watching real-time science, unfortunately, people have also witnessed discourse around the Nature of science (NOS) that brings into question the trustworthiness and reliability of the scientific process and questions the credibility of science as a whole (Kennedy, Tyson, & Funk, 2022). The Covid-19 pandemic uncovered the lack of understanding that people still hold regarding the NOS. Furthermore, it showcased how this lack of understanding can have real-world consequences once students leave the classroom. Therefore, it is important to continue to investigate how science can be taught in a way that supports students in their understanding of the NOS.

Students' grasp of the NOS and the teaching of the NOS has been a topic of interest for the last several decades (Mccomas, 2011). The NOS gives people the ability to make informed decisions on scientific issues related to society and enhances students' understanding of scientific topics (Mccomas, 2011). The culmination of research on the NOS, science practices, and science teaching led to the creation of the Next Generation

Science Standard (NGSS) whose goal is “to create a set of research-based, up-to-date K-12 science standards” (Next Generation Science Standards, 2013). A defining characteristic of the NGSS is their dedication to the inclusion of science and engineering practices in the K-12 classroom, a set of skills that scientists use to investigate phenomena in the world. There are eight science and engineering practices defined by the NGSS, which will be referred to as science practices: (1) Asking questions and defining problems, (2) developing and using models, (3) planning and carrying out investigations, (4) analyzing and interpreting data, (5) using mathematics and computational thinking, (6) constructing explanations and designing solutions, (7) engaging in argument from evidence, and (8) obtaining, evaluating, and communicating information (NGSS Lead States, 2013). While the NGSS is progressive in this regard, the best ways for implementing these practices in the classroom are still being investigated (B.J. Reiser, 2013).

The NGSS has pushed for the inclusion of science practices and additional curriculum that generates opportunities for students to engage with science, supporting their understanding of the NOS. In addition, the NGSS has influenced higher education and research, putting added pressure on higher education to provide more opportunities for students to engage in science practices. Now, more than ever, a large portion of students will be entering college classrooms with skills that can be built on to further enrich their understanding of science. Additionally, there have been calls for the improvement of post-secondary science education, with the goal of better preparing future STEM professionals and developing science literacy in non-STEM majors as well (Gardner, 1983; National Academies of Sciences, Engineering, 2010; National Research

Council, 2007; Olson & Riordan, 2012). Several reports in higher education have also called for reform in curriculum that includes evidence-based practices in the classroom (Cooper et al., 2015; Laverty et al., 2016). Similarly, to secondary education, the question that remains is what are the best practices for implementing and supporting student engagement in science practices at the post-secondary level?

1.1 Background and Need

There have been some strides in the inclusion of evidence-based practices in the classroom (Erduran, Simon, & Osborne, 2004; Faize, Husain, & Nisar, 2018; James & Ladue, 2021; Kallery, Psillos, & Tselfes, 2017a; Joi Phelps Walker, Sampson, & Zimmerman, 2011), but there is still much to uncover about how science practices can be best integrated into the classroom in meaningful ways. To develop the best practices two questions must be considered: how or when to implement science practices into the curriculum, and how to support students in these practices. Meaningful implementation of science practices and their concepts is necessary because it promotes a deeper understanding of concepts and cultivates independent thinking by students (Kuhn, Arvidsson, Lesperance, & Corprew, 2017). Therefore, by considering these questions, the positive learning gains and conceptual reinforcement seen by previous research in the field have a higher probability of being replicated in the classroom (Hosbein, Lower, & Walker, 2021; Murphy et al., 2018; J. Walker, Sampson, Grooms, Anderson, & O. Zimmerman, 2012). Specifically, our study focused on how to support students engaged in experimental design activities in a lecture setting.

How or when to implement science practices into the post-secondary curriculum is important to consider when developing activities for students to do. Most of the research in chemistry education that focuses on the implementation of evidence-based practices and science practices focuses on incorporating activities in the laboratory setting or involves a complete reconfiguration of laboratory curriculum (Collison et al., 2018; Williams & Reddish, 2018). Argument-driven inquiry laboratories are the culmination of much of this research. In these laboratories, the focus is put on having students go through steps a scientist would in the real world rather than the traditional "cookbook" style laboratories that instruct students what to do during laboratory (Carlo & Flokstra, 2017; Choi et al., 2013; J. Walker et al., 2012; Joi P. Walker & Wolf, 2017; Walker et al., 2011, 2019; Walker & Sampson, 2013). Argument-driven inquiry laboratories guide students through seven steps that mirror what a researcher would follow: (1) identification of task, (2) generation and analysis of data, (3) production of a tentative argument, (4) argumentation session, (5) explicit and reflective discussion, (6) creation of written investigation report, (7) double-blind peer review of the reports, and (8) revision of report (Sampson, Grooms, & Walker, 2011). In addition, laboratories and other supportive courses are frequently taught by teaching assistants (TAs) that also come from diverse backgrounds, often time having little to no experience with supporting students engaging in science practices. This means that in addition to supporting and training students in science practices, TAs also need new training on how to conduct these laboratories, adding another layer of complexity in implementing science practices in laboratory settings or other supportive courses taught by TAs (Wheeler, Clark, & Grisham, 2017). However, even if someone has the ability and power to make this kind of decision, only changing the laboratory component could further isolate the laboratory and lecture components (Collison et al., 2018). Therefore, researchers must investigate and develop ways in which science practices can be integrated into the lecture component as well.

In addition to knowing how or when to implement science practices into the classroom, knowing how to support students as they engage with these activities is essential for meaningful learning to take place. While the NGSS has been adopted in several U.S.A. states, college classrooms often have students from diverse backgrounds that may or may not have engaged with science practices before. Whether or not students have engaged with science practices before, extra support for these activities is still needed given the difficulty associated with learning epistemic practices. Epistemic practices are "*the socially organized and interactionally accomplished ways that members of a group propose, communicate, evaluate, and legitimize knowledge claims*" (Matthews, n.d.). Not all science practices fall under the broader construct of epistemic practices, however, designing and carrying out experiments does as it is the mechanism by which claims in science are evaluated and legitimized (Jiménez-Aleixandre & Crujeiras, 2017). Previous research in experimental design suggests that students need support in understanding the scope of what they can do in these activities and how they should carry out these activities (Arnold, Kremer, & Mayer, 2014). Providing this level of support can be challenging at the post-secondary level given how large enrollment typically is in introductory chemistry courses (Henderson, Ryan, & Phillips, 2019). A potential solution to this problem that can still provide support for students is the inclusion of peer review in science practice activities. In addition to alleviating some of the work that would otherwise be put on the instructor, peer review can help students develop evaluative judgment (McConlogue, 2015; Nicol & McCallum, 2021; Nicol, Thomson, & Breslin, 2014). This is especially true if the activity requires students to create a final product, which is typical of science activities in the laboratory and lecture setting. By including peer review, science activities have the potential to support students' development of deeper understanding related to both science practices and evaluative judgment (Berg & Moon, 2022). However, most peer review literature centers around longer writing assignments and the outcomes associated with different kinds of feedback (Carless & Boud, 2018; Finkenstaedt-Quinn, Snyder-White, Connor, Gere, & Shultz, 2019; Patchan & Schunn, 2015, 2016; Patchan, Schunn, & Correnti, 2016). Therefore, more research is needed on how peer review can be leveraged to help students develop and improve upon their experimental design abilities.

Current research in science practices points towards potential solutions to the challenges associated with implementing science practices, in our case experimental design, in the classroom. Bringing opportunities to design experiments into the classroom has the potential to alleviate problems, such as logistics and the need for additional TA training (Wheeler et al., 2017), that are encountered when integrating them into laboratories. Additionally, research suggests that including peer review in science practices can support students engaged in these practices while also having the potential to increase students' epistemic practices (Kuhn et al., 2017). However, few studies have investigated the integration of peer review into experimental design activities (Basso, 2020; J. Walker et al., 2012) and none have investigated their effects in a lecture setting. Therefore, additional research is needed that investigates different aspects of peer review and its effect on the outcome of going through the peer review process.

1.2 Purpose of the Study

The purpose of this study was to investigate if the structure of a peer review template affects the nature of student feedback, and the type of revision done. To accomplish this, an experimental design task was created for organic chemistry students that prompted them to design an experiment to investigate the properties of a solvent that can be switched between hydrophobic and hydrophilic. To determine the ability of peer review to support students when engaging in science practices, an investigation into how to best elicit feedback and revisions from students is needed. Additionally, providing feedback and making revisions are essential to the scientific process and peer review offers one way in which these skills can be developed. To explore the impacts of peer review, the researcher created a task that prompted students to write out an experiment that could provide additional evidence about the "switchable" nature of a solvent. After submitting their initial draft online, students were randomly assigned three anonymous students to review. After giving and receiving feedback from their peers, students then uploaded a final draft of their design. Students in one section of Organic Chemistry I (n=76) were given a peer review template that included basic instructions about how long their feedback should be (the non-scaffolded peer review template). While students in another section of the same course (n=119) were given a peer review template that included the same basic instructions and a list of criteria that are used to evaluate an experimental design (the scaffolded peer review template). Such criteria included asking students if a discussion of acid-base theory and how it frames their hypothesis was included in the design. The goal of this study was to measure the effects that the structure of peer review had on the feedback students gave and the amount of revision made.

1.3 Research Questions

1. How does the structure of the peer review template impact the feedback students give for an experimental design task?
2. How does the structure of the peer review template impact students' revisions for an experimental design task?

CHAPTER 2: LITERATURE REVIEW

This chapter explores how experimental design activities can be utilized in a classroom setting, how to design such an activity, and how peer review can be used to support students in these practices. In section one, a review of studies investigating student competencies in experimental design will be presented. The last section will summarize current peer review research and make an argument for the use of peer review as a tool for supporting students engaging in experimental design. Overall, this chapter will summarize the research that informed the present study.

2.1 Experimental Design Activities

The purpose of this section is to further evaluate the best practices for implementing experimental design activities. To eventually be able to support students in developing their experimental design skills, understanding is needed about different competency levels that students currently possess. To gain insight into this, tasks must be created that (1) prompt students to engage in experimental design, and (2) differentiate between different experimental design competencies. The following section will synthesize three studies that provide insight into the current understanding of student competencies in experimental design and how a task can uncover difficulties students have to differentiate between competencies.

When designing experiments in an inquiry laboratory, students must coordinate three pieces; (1) the theoretical ideas related to the problem, (2) the evidence needed to elucidate the problem (representations of data or processed data), and (3) the materials from the experiment (i.e., raw data, laboratory equipment, laboratory techniques, etc.). Assisting students in making these connections is essential for them to fully engage in experimental design (Psillos, Tselfes, & Kariotoglou, 2004). To better understand how to assist students, Kallery et al. (2017) designed an experiment to investigate students' ability to make these connections when engaging in an experimental design activity. In this study, 25 secondary students were tasked with designing an experiment to investigate the claim that mugs made of two different materials heat up water at different rates when placed on a burner (i.e. investigate the relationship between heat and temperature). To evaluate the students' experimental design, researchers used a framework of analysis that focused on capturing connections students make between theory, evidence, and materials across seven dimensions related to experimental design (Lefkos, Psillos, & Hatzikraniotis, 2011). The seven dimensions included experimental procedure description, separation of variables, handling of variables, initial conditions, devices and instruments, device settings, and forming a hypothesis. Researchers classified student responses across these dimensions into three levels: missing (level 1), partially stated (level 2), and completely stated (level 3). Researchers also defined expected connections that are needed in each of the seven dimensions. For example, in the devices and instruments dimension, researchers determined that students need to first connect the theory to the evidence and then connect the evidence to the material world. Researchers tabulated the percentage of students at each level that made the expected connections between dimensions. Their results showed that students in level 3 (top-performing students) had difficulty connecting theory or concepts to the raw data or equipment needed to produce their desired results. In addition, students across all levels struggled to connect evidence and theory when forming hypotheses and when determining independent or dependent variables. This aligns with other research that has shown students often struggle to design experiments that align with the hypothesis that they are trying to test and struggle with manipulating variables in an experiment (De Jong & Van Joolingen, 1998; Lawson, 2002). This suggests that focusing on hypotheses and variables in a task could help differentiate between different competency levels in students, meaning that students who are able to connect evidence with theory when forming a hypothesis could indicate high levels of competency in experimental design. However, the authors used the difficulty students had when designing experiments as evidence that "*involving students in experimental design activities does not necessarily promote scientific ways of thinking*" (Kallery et al., 2017b). This is a contradiction to research findings from other experimental design literature and the wider science practices community (Cooper et al., 2015; Kuhn et al., 2017; National Research Council, 2012). Additionally, this claim is based on student participation in one activity, when evidence suggests that students need multiple opportunities to learn epistemic practices (Barzilai & Chinn, 2018). What these difficulties do suggest, however, is that students need additional supports when first participating in designing an experiment in order to begin to develop competencies in this practice (Beishuizen, Wilhelm, & Schimmel, 2004).

Van Riesen et al (2018) sought to provide some insights on how to support students when designing experiments by developing and employing an Experimental Design Tool (EDT) with 120 secondary students. The EDT provides students with a step-by-step structure to design an experiment and has built-in heuristics to guide students through the activity. This tool is meant to act as scaffolding for designing an experiment, which helps students perform a task that is difficult to accomplish on their own (Brian J. Reiser, 2018; Simons & Klein, 2007; van Riesen et al., 2018). To test the effectiveness of the EDT to support students in experimental design, students were divided into three conditions: the experimental condition, the control specific (CS) condition, and the control main (CM) condition. Students in all the conditions worked through a virtual laboratory about buoyancy and Archimedes' principle that included thirteen research questions that could be organized under five broad research questions. In the experimental condition, students worked through the laboratory using the EDT. In the CS condition, students did not have the EDT assisting them but had the thirteen research questions organized under the five broad research questions. In the final condition, the CM condition, the students were only given the five broad research questions. To test for differences between the conditions, researchers administered a pre-and post-test that measured students' conceptual knowledge of the principles covered in the experiment (buoyancy and Archimedes' principle). The test included questions about the concepts in the experiment and had students apply those concepts as well. Results from this assessment found that there was no significant difference in gains when the three conditions were compared. However, researchers found that students in the EDT condition that had low prior knowledge did have significantly higher gains pre to post-test than students with low prior knowledge in the CS, but not the CM condition. Similar results were seen by Alexander and Judy (1988), who found that students with lower prior knowledge benefit more from additional scaffolding and guidance (Alexander & Judy, 1988; Hmelo, Holton, & Kolodner, 2014; Tuovinen & Sweller, 1999). The researchers in the study suggested that the significant gains in the EDT condition only being present when compared to one control condition signals that scaffolding is not a "one-size-fits-all principle". Similar conclusions were drawn by Perez et al. (2017) who found that when students with lower prior knowledge completed simpler experiments, they experience higher learning gains than when given more complex experiments.

While it is possible to see conceptual learning gains in students when they complete an experimental design activity with scaffolding, these studies still do not provide insight into improvements made in designing actual experiments. Dasgupta et al. (2014) provided one solution to this problem through the development and validation of a rubric to evaluate students' experimental designs referred to as the rubric for experimental design (RED). RED offers one way to capture improvements students make in their experimental designs. To develop a rubric able to do this, researchers deployed three experimental design activities from the literature into an undergraduate biology course. Researchers coordinated difficulties had by students in the literature with their responses to the activities to come up with five areas of difficulties that students face. The following areas were identified: the variable properties of an experimental subject; the manipulated variables; measurement of outcomes; accounting for variability; and the scope of inference appropriate for experimental finding (Dasgupta et al., 2014). For each area, the authors defined completely correct ideas and the evidence that tended to signal difficulty in the area. To validate the utility of the RED, additional testing was done by collecting student data pre- and post-instruction online via five additional experimental activities over the course of a semester. Then, the student products were analyzed using RED to determine if differences between students and pre/post instruction could be detected. Researchers found that the RED was able to detect changes or improvements in student answers from the pre/post activities. Additionally, it was found that students have similar difficulties when designing experiments as defined in the literature. This study provides evidence of areas that students may need extra support on when completing. They attributed this to novice students having difficulty applying concepts when the context was changed, a finding seen in another study (Barnett & Ceci, 2002; Goodman, Wood, & Chen, 2011). However, this could also be attributed to students not receiving feedback after each activity. Feedback has been found to be an effective tool for helping students learn (Hattie & Timperley, 2007; Mory, 1996; Shute, 2008).

The studies in this section offered insight into areas that students often struggle with when designing an experiment and how scaffolding could offer support in these areas. However, scaffolding is not a "one-size-fits-all" mechanism and Van Riesen et al. (2018) found that significant improvements were only seen in low prior knowledge students. In addition to scaffolding, providing feedback to students could be an effective tool for supporting students. However, for large introductory courses, this may not be an option for the instructor given the high enrollment and necessity for students to have multiple opportunities to practice. Therefore, I posit that peer review could be an effective tool for supporting students when they are designing experiments. The following section will review research on feedback and how peer review plays a role in student learning.

2.2 Peer Review

Peer review is an essential part of the scientific process that all scientists must go through. In the real world, peer review acts as a mechanism to evaluate the integrity, credibility, and quality of the research to ensure valid conclusions are drawn. Within education, peer review provides an opportunity for students to receive feedback on their ideas and provide feedback for their peers' ideas. There is a large body of literature investigating the potential positive effects that receiving feedback can have on student learning (Azevedo & Bernard, 1995; Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Chi & Wylie, 2014; Koedinger, Corbett, & Perfetti, 2012), but the effectiveness of feedback is still debated in the community (Molloy & Boud, 2014; Mory, 1996, 2004; Shute, 2008). However, the research by Nicol et al. (2014) suggests that engaging in the process of peer review could help students develop evaluative judgment (Nicol et al., 2014). Evaluative judgment is "*the capability to make decisions about the quality of work of oneself and others*" (Tai, Ajjawi, Boud, Dawson, & Panadero, 2018). As the previous studies show, students often struggle with making decisions about experimental design, particularly when manipulating variables and forming a hypothesis (De Jong & Van Joolingen, 1998; Lawson, 2002). Therefore, developing students' evaluative judgment through the peer review process could offer much needed support to students when designing experiments. Additionally, a goal of engaging students in science practices is to instill students with a deeper understanding of the epistemic criteria of science (Kuhn et al., 2017), which includes competencies in evaluative judgment.

The focus of research in the peer review space has been mostly on the process of receiving feedback. While receiving feedback from peers has been shown to improve student's quality of work (Cho & MacArthur, 2011), just the act of receiving feedback does not always prompt students to make revisions to their work (Finkenstaedt-Quinn et al., 2019). To better understand the phenomenon of receiving feedback, Patchan et al. (2016) investigated the nature of feedback and the influence it has on the rate that students make revisions. Similar to the advice given about providing helpful feedback, they found that when students received praise for their work (Bienstock et al., 2007;

Hesketh & Laidlaw, 2002) or localized or specific feedback (Goodman & Wood, 2004; Goodman et al., 2011; Goodman, Wood, & Hendrickx, 2004; Patchan & Schunn, 2016) were more motivated to revise their work. However, the inclusion of praise in feedback was found to sometimes impede student revisions. Carless and Boud (2018) also found that affect plays a role in receiving feedback, along with the perceived value of the feedback process. In addition, the student must also judge what feedback warrants revisions (Carless & Boud, 2018).

The complexity of feedback uptake has shifted focus to the process of giving feedback. Current work in this area has found that revisions for students are higher when they give feedback versus when they receive feedback (Anker-Hansen & Andrée, 2015; Cho & MacArthur, 2011; Ion, Sánchez Martí, & Agud Morell, 2019; Lundstrom & Baker, 2009; Nicol & McCallum, 2021). When students are giving feedback, students compare their work with the student they review (McConlogue, 2015; Nicol et al., 2014; van Popta, Kral, Camp, Martens, & Simons, 2017). Comparing their work prompts students to reevaluate their own work and its alignment to task criteria (McConlogue, 2015; Nicol & McCallum, 2021; Nicol et al., 2014). This process of evaluating their own work against another can prompt the generation of internal feedback. Internal feedback is what drives students to revise their understanding and make improvements (Butler & Winne, 1995; Nicol et al., 2014). Students also report that formulating internal feedback reduces the need for additional feedback, as they were already able to identify and make the changes suggested (Anker-Hansen & Andrée, 2015; Nicol & McCallum, 2021; Nicol et al., 2014). Recent peer review literature suggests that peer review can provide a mechanism for students to generate internal feedback (Nicol, 2021; Nicol & McCallum, 2021). During the peer review process, students are using their work as the standard for which they evaluate other's work. The act of comparing one's work to another promotes reflection, helping students generate internal feedback about their own work.

Berg and Moon (2022) further investigated the ability of peer review to help students generate internal feedback on a data analysis and interpretation activity. Data analysis and interpretation are one of the science practices outlined by the NGSS. Students generated a response to the data analysis and interpretation task and then were given other responses to evaluate. Researchers asked students what they thought of the other responses and if they would make any changes to their response, simulating the process of peer review. Through this, they found that students generated internal feedback when looking at a response similar to their own and when looking at responses different from their own. Once internal feedback was generated, it either validated their response (leading to no revision) or incentivized students to improve their response (leading to revision). In addition, some students completely revised and improved their whole argument upon generating internal feedback. This study showcases the potential of peer review in helping students to regulate and develop competencies in science practices. However, follow-up investigations are needed to determine if these outcomes are replicable with peer review in a classroom setting.

2.4 Social Comparison Theory

Peer review is a social process that typically involves students comparing their own work with their peer's work. During this comparison, the way a student perceives their own work relative to others influences the product of the comparison. This can then impact the feedback students give and the evaluation of their own work to make revisions. I wanted to leverage these social comparisons to investigate how the structuring of peer review templates affects the kind of feedback students give and the revisions they make.

Social comparison theory was developed in 1954 by social psychologist, Leon Festinger, who theorized what happens when an individual is placed in an environment in which they are uncertain about how to behave or think. He theorized that individuals will compare themselves to others to reduce uncertainty (Festinger, 1954). People will often engage in this comparison when the environment has specific standards and criteria that must be met (Levine, 1983; Martin, 2000; Smith and Arnkelsson, 2000; Alicke, 2007; Pomery et al., 2012; Miller et al., 2015; Greenwood, 2017). By comparing themselves to others, an individual can appraise their relative ability and performance.

The subject to whom they are comparing themselves is referred to as the “target” (Martin, 2000; Smith and Arnkelsson, 2000; Alicke, 2007; Pomery et al., 2012; Miller et al., 2015; Greenwood, 2017). The target is generally a subject, real or imaginary (i.e., a simulated response or product), that exists in a similar environment to the individual making the comparison. How the individual perceives the target’s performance determines the direction of social comparison that is being made. There are three types of social comparisons that an individual can make, an upwards comparison, a downwards comparison, or a lateral comparison. During an upwards comparison, an individual views the targets as superior or of higher quality than themselves Whereas during a downwards comparison, an individual views the target as inferior or of lower quality. Lastly, if an individual views the target as being like themselves, then this is considered a lateral comparison. The kind of social comparison an individual makes is motivated by the reason for comparing as well as the need to reduce uncertainty (Pomery, Gibbons, & Stock, 2012).

In addition to reducing uncertainty, researchers have recognized three primary types of motivation: self-evaluation, self-improvement, and self-enhancement (Dijkstra, Kuyper, Van Der Werf, Buunk, & Van Der Zee, 2008; Pomery et al., 2012). Self-evaluation is associated with an individual's motivation to evaluate their own work or standing by comparing to someone they perceive to be like their own (i.e., lateral comparison). Festinger theorized that the more similar a target is to an individual, the more precise their evaluation will be (Pomery et al., 2012). The next two motivations are associated with individuals comparing themselves to targets that they perceive as different from themselves. Self-improvement is the desire to improve oneself by comparing to others. This type of motivation is generally associated with an upward comparison, as an individual will seek out a person they view as doing better than oneself to learn new skills (Pomery et al., 2012). On the other hand, when an individual is making a downward comparison, they are motivated by self-enhancement. In this scenario, an individual is motivated to improve their feelings about their own work, such as ease anxiety, by comparing to a target they view as worse off than their own (Pomery et al., 2012).

The classroom provides an evaluative atmosphere that is ideal for engaging students in social comparisons (Pomery et al., 2012; Pepitone, 1972). Students are motivated to learn new material and learning new material or engaging in unfamiliar practices often generates cognitive uncertainty in students. Social comparisons offer a way for students to alleviate uncertainty by providing an avenue for students to evaluate and obtain internal feedback (Levine, 1983). Peer review provides students the opportunity to engage in social comparison to evaluate and obtain internal feedback. Doing peer review exposes students to responses of different sophistication, which could lead to social comparisons based on different motivations. Therefore, using social comparison theory to frame our investigation allows us to focus on the reviewer and the feedback they give when evaluating the effect that peer review structure has.

2.3 Summary

To help students develop competencies in experimental design, students must be given the opportunity to practice and develop their skills outside the laboratory as well. The current literature on experimental design focuses on improving conceptual learning gains and the difficulties students face when designing experiments. However, there is little understanding on how students can overcome difficulties seen repeatedly in the literature such as manipulating variables and connecting theory to the hypothesis (De Jong & Van Joolingen, 1998; Kallery et al., 2017b; Lawson, 2002). While scaffolding does appear to provide some support for students while designing experiments, it tends to only support students with low prior knowledge (Perez et al., 2017; van Riesen et al., 2018). Feedback from an instructor or peers could provide additional support to students, but the effectiveness of feedback is still debated (Molloy & Boud, 2014; Mory, 1996, 2004; Shute, 2008). Feedback is meant to help students improve their drafts and understanding, but often it does not (Molloy & Boud, 2014; Mory, 1996, 2004; Shute, 2008). However, the act of generating feedback has been shown to drive revisions and improve student responses (Lundstrom & Baker, 2009; Patchan, 2011; Sadler & Good, 2010; Wooley, Was, Schunn, & Dalton, 2011). When giving feedback, students generate internal feedback which allows them to reflect on their own response. Peer review can act as the vehicle for generating internal feedback and could be useful in supporting students engaged in science practices (Berg & Moon, 2022). However, more investigation is needed to determine if and how these results could be replicated in a classroom setting with peer review. Therefore, this study contributes to the current literature by investigating the effects that peer review structure has on the outcomes of an experimental design activity. More specifically, our study will focus on capturing the effects seen on the peer review students give and the revisions students make on their experimental designs. Our focus on the peer review students give is informed by social comparison theory.

CHAPTER 3: METHODS

To investigate the effect the peer review prompt structure has on the peer review process, a quasi-experimental design was used. In a quasi-experimental design, individuals are not randomly assigned to a treatment or control group (Maydeu-Olivares, 2009). This type of design can be chosen for several reasons, the most common being ethical or practical restraints of randomized experiments. For our study, it was not practical to collect randomized data given the limitations associated with administering a task through a course learning management system. Instead, data was collected from two sections of a course where one acted as a control group and the other as the treatment group. In our study, I defined the control group as students who were given a non-scaffolded peer review prompt and the treatment group as students who were given a scaffolded peer review prompt. The nature of these two groups will be discussed throughout the rest of the methods section. For the remainder of the paper, the two groups will be referred to as the “scaffolded group” and “non-scaffolded group”.

3.1 Participants and Context

This study was conducted with students enrolled in Organic Chemistry I at the University of Nebraska-Lincoln during the spring 2021 semester. According to the University of Nebraska Institutional Review Board (IRB) guidelines, this study is exempt from needing official approval as it takes place in an established educational setting and likely does not have any adverse effects on students or instructors. Data was collected in two sections of organic chemistry I to have a large enough sample for quantitative analysis. Both sections were taught by the same professor and covered the same content.

These sections were designated for non-chemistry STEM majors and pre-professional students whose majors included biological sciences, agriculture, pre-medical, and pre-dentistry. The course is divided into three components: lecture, recitation, and laboratory. For our study, all participation occurred through the lecture portion of the course. The lecture was taught synchronously, where a student could attend in-person or via Zoom, three days a week. During the lecture, the professor mainly utilized traditional lecture-based instruction and would occasionally use a 'flipped' classroom approach for Friday lectures. All course announcements, assignments, and exams were given online through the course learning management system. For this reason, the task was administered online via Canvas.

3.2 Study Design

I utilized a quasi-experimental design approach to help answer our research questions. In our study, students were divided into the scaffolded group and the non-scaffolded group based on what section of the course they were enrolled in. Students in both groups submitted an initial draft for the task, then were randomly assigned three students to peer review. After the peer review process was completed, students revised their initial draft and submitted a final draft of the task. Since I was concerned with how peer review affects revisions, students who did not submit the initial draft, peer review, and final draft were not included in our final analysis. After accounting for this, a total of (n=76) students were in the control group and (n=119) in the experimental group.

To investigate the effect that peer review structure has on the type of feedback and revisions students make, I needed students to meaningfully engage with the peer review process and provide content-based feedback. To accomplish this, students need scaffolding regarding what constitutes good feedback and instruction on how to provide feedback (Finkenstaedt-Quinn et al., 2019). Therefore, a peer review template was created that included an introduction about the importance of peer review in science and instructed students to provide 2-3 sentences of feedback. Students in both the scaffolded group and non-scaffolded group were given a peer review template, shown in Figure 1 and Figure 2. In addition to this, students in the scaffolded group were given criteria about what should be provided in an experimental design. The criteria included the following: (1) discussion of acid-base theory and how it frames their hypothesis, (2) what data will be gathered and recorded, (3) how much data is needed to support valid conclusions, (4) limitations in the precision of the data that will be collected, (5) identification of the independent, dependent, and control variables, (6) consideration of possible confounding variables, and (7) possible conclusions they will be able to draw from the data they collect. These criteria align with the NGSS standards for evaluating proposed experiments (NGSS Lead States, 2013).

Peer Review Rubric

Peer review is an essential part of science and is a way for scientists to evaluate the validity of experiments, claims, and reasoning. You are going to have an opportunity to review your peers' responses to the task. Your review should be substantive, constructive, and respectful.
Substantive: Your review should focus on the content of their experimental design, not the grammar or stylistic choices.
Constructive: Your review should prompt the person you are reviewing to think about something in a new way, consider another variable, or revise their experiment to better articulate what they mean.
Respectful: Your review should be kind and should never question someone's character or intelligence.

Provide a brief review of their explanation in question 1 that will help them improve their response (minimum 2-3 sentences). Their response should include an assertion about what volume of DCMA should be used, data that supports their assertion, and an explanation for how the data supports their assertion.

[text box]

Provide a one paragraph review of the experimental design in question 2 that will help them improve their experiment (minimum 4-5 sentences).
Each proposed experiment should include the following:

- Discussion of acid-base theory and how it frames their hypothesis
- What data will be gathered and recorded
- How much data is needed to support valid conclusions
- Limitations in the precisions of the data that will be collected
- Identification of the independent, dependent, and control variables
- Consideration of possible confounding variables
- Possible conclusions they will be able to draw from the data they collect

[text box]

Figure 1: Scaffolded Peer Review Template

Peer Review Rubric

Peer review is an essential part of science and is a way for scientists to evaluate the validity of experiments, claims, and reasoning. You are going to have an opportunity to review your peers' responses to the task. Your review should be substantive, constructive, and respectful.
Substantive: Your review should focus on the content of their experimental design, not the grammar or stylistic choices.
Constructive: Your review should prompt the person you are reviewing to think about something in a new way, consider another variable, or revise their experiment to better articulate what they mean.
Respectful: Your review should be kind and should never question someone's character or intelligence.

Provide a brief review of their explanation in question 1 that will help them improve their response (minimum 2-3 sentences).

[text box]

Provide a one paragraph review of the experimental design in question 2 that will help them improve their experiment (minimum 4-5 sentences).

[text box]

Figure 2: Non-Scaffolded Peer Review Template

3.3 Task Development

To collect peer review, I first developed a task that targeted the science practice of designing an experiment. First, a literature search was done to find a context or phenomenon that aligned with topics taught in the course. This offers an opportunity to enrich students' understanding of a topic that they are already covering in class (Zagallo, Meddleton, & Bolger, 2016). The phenomenon chosen for this study was the use of the switchable solvent, DMCA (N,N-dimethyl cyclohexylamine), in the extraction of benz[a]anthracene from water samples (Lasarte-Aragonés, Lucena, Cárdenas, & Valcárcel, 2015). DMCA is a hydrophobic solvent that when combined with CO_2 in water will change to hydrophilic. When the molecule is "switched", it becomes hydrophobic again and separates from the water. The process of switching can be utilized to remove analytes (in this case benz[a]anthracene) from water. Once the Benz[a]anthracene is removed, fluorescence is used to determine the concentration of benz[a]anthracene that was in the water sample. The mechanism of switching DMCA relies on acid and base theories taught in organic chemistry I, making it a good fit for this task. Lasarte-Aragonés et al. (2015) investigated several variables affecting the switchable solvents mechanism and the effectiveness of multiple solvents. For our task, I simplified the context to look at just the final step of the mechanism when the extraction occurs, removing details that would distract or confuse students. In the task, students designed two experiments to test two theories about how the extraction occurs and determine which method would be the most effective. This approach is like that of Zagallo et al. (2016), who created a model for designing and teaching data interpretation with real-world data.

The final version of the task that was given to students is shown in Figure 3 and Figure 4. Previously created science practice tasks from our group informed the structure of this study's task. Data collection with the tasks showed that including an introduction of the topic, scaffolded questions, and explicit instructions about what should be included in their answer, engaged students with the practice most effectively. Therefore, the task for this study consisted of an introduction to switchable solvent extractions using benz[a]anthracene, an initial question to help students with the concepts related to the context, and a final question prompting them to design an experiment investigating the two theories. The first question asked students to use the graph and make an argument for which volume of DMCA should be used to extract the most analyte. The second question asked students to design two experiments that would investigate whether an acid or base is what switches the DMCA at the final step of extraction and which would be the most effective method for extracting the benz[a]anthracene. To assess the validity of the task and its alignment with the course, the instructor provided feedback about the context and task itself. They suggested clarifying the steps in the figure to show the process of switching the DMCA more clearly and suggested changing the format of the molecules to better align with representations students would have seen in class. The instructor's feedback was implemented in the final version of the task to make sure the concepts in the task were accessible to students.

The Use of Switchable Solvents in the Extraction of Benz[a]anthracene from Water Samples

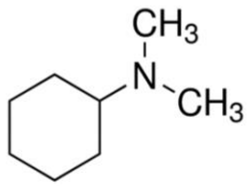


Figure 1. DMCA

In this task you are going to think critically about the "switchable" solvent, DMCA (N,N-dimethyl cyclohexylamine), and its use in extracting hydrophobic analytes. Figure 1 shows DMCA in its neutral form. In this form, DMCA is relatively hydrophobic. However, when DMCA is combined with CO_2, it becomes hydrophilic via the following reaction in water:

$$NR_3 + CO_2 \rightarrow NR_3H^+ + HCO_3^-$$

The charged DMCA molecule (NR_3H^+) is then miscible in water. When the molecule is "switched", it becomes hydrophobic again and separates from the water. It has been found that this method of "switching" DMCA can be used to extract hydrophobic analytes from water. In this case, DMCA will be used to extract benz[a]anthracene (hydrophobic) from water. The overall approach used to extract benz[a]anthracene is shown in Figure 2.

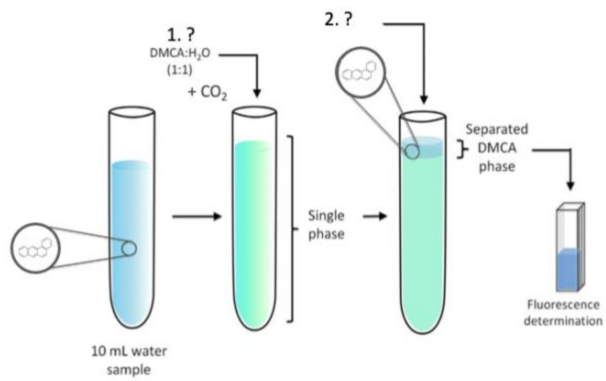


Figure 2. Schematic for DMCA analyte extraction

Step 1: DMCA and CO_2 are added to a 10 mL water sample with benz[a]anthracene to make a single phase (i.e., they mix together).

Step 2: A reagent is added to the mixture to "switch" the DMCA to make the layers separate. The DMCA extracts the benz[a]anthracene from the water layer during this process (i.e., the benz[a]anthracene is now in the separate DMCA layer).

Step 3: The DMCA layer is extracted from the sample and the fluorescence is measured to determine the concentration of benz[a]anthracene that was in the 10 mL sample of water.

There are two parts of this process that you will figure out.
1. The optimal volume of DMCA to use
2. A way to determine the optimal reagent for switching the DMCA back in order to separate the layers.

Figure 3: Experimental Design Task Introduction Page

1. One affordance of benz[a]anthracene is its native fluorescence. This means that the intensity of the fluorescence correlates to the concentration of benz[a]anthracene extracted from the solution (higher intensity = higher concentration). Here is data showing the fluorescence intensity of benz[a]anthracene for three different volumes of DMCA. What volume of DMCA will you use to carry out the extraction? Be sure to clearly explain the reasoning for your choice.

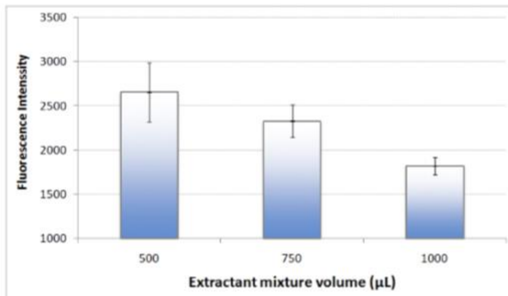


Table 1: The fluorescence intensity of benz[a]anthracene extracted using different volumes of DMCA. Each volume of DMCA was used in 10 mL of water.

[text box]

2. There are two ideas for switching the DMCA back and completing the extraction.
   a. Acids will remove the carbonate (CO_3^{2-}) and bicarbonate (HCO_3^-) from the solution in the form of CO_2 and therefore may induce phase separation, or
   b. Bases induce a change in the chemical state of the DMCA to its neutral form and therefore may induce phase separation.

   Propose *two* experiments/tests you will carry out to determine what you would use to switch back, either an acid or base. Be sure to describe your experiments in detail. For each experiment, include the a) the hypothesis you are testing, b) what variables will be controlled and changed, and c) the possible outcomes and the conclusions you can make based on those outcomes. For each part, justify your decisions with sufficient reasoning to convince your peers of your experimental design.

   [text box]

Figure 4: Experimental Design Task Questions

3.4 Data Analysis

I used multiple statistical tests to determine if there were differences in the frequency of feedback and revisions between the scaffolded group and the non-scaffolded group. To be able to run these tests, I transformed the qualitative data into quantitative data by assigning a score or category to each draft and feedback comment. An experimental design rubric was created to evaluate and assign scores to students' initial and final drafts. Another rubric was created to categorize the type of feedback that students gave during the peer review process. Interrater reliability was calculated for each rubric using three researchers until an acceptable value for Krippendorff's alpha was obtained. Krippendorff's alpha is a reliability coefficient that measures agreement among coders when assigning categories or values to data (Krippendorff, 2011). Typically, an acceptable agreement between coders ranges from $0.667 \leq \alpha \leq 0.823$, with anything lower than 0.667 considered to be unacceptable (Shabankhani, Charati, Shabankhani, & Cherati, 2020). The following section will discuss the development and reliability of each rubric.

3.4.1 Experimental Design Analysis

I developed a rubric to evaluate students' hypotheses, variable manipulations, and outcomes and conclusions in their initial and final drafts, shown in Tables 1, to determine if the peer review template structure had any effect on the revisions students made (RQ2). Previously discussed literature showed these three areas to be difficult for students to consider when designing an experiment (De Jong & Van Joolingen, 1998; Kallery et al., 2017a; Lawson, 2002). Therefore, I wanted to capture improvements students made in these three areas. Each area was scored out of three points. The higher score, the more complete a student's response was. Once the coding was finished, students were given an overall score on their experimental design by adding up their scores in each area. The highest score a student could receive was a 9 and the lowest was a 0.

Based upon the NGSS guidelines for designing an experiment, a hypothesis should include a discussion of the related theory, make a prediction, and be testable (NGSS Lead States, 2013). I defined testability as being able to be proven through scientific investigation. The rubric for the hypothesis is shown in Table 1. Students who included theory in their hypothesis had all three elements of a hypothesis and were given a score of three. Students who made a prediction in their hypothesis and wrote a testable hypothesis were given a score of two. If a hypothesis was only considered testable and did not make a prediction or include theory, it was given a score of two. Students who did not include a hypothesis or included one with none of the above criteria were given a zero.

Table 1: Hypothesis Scoring Rubric

Score	Description	Student Example
3	Student has a testable hypothesis, makes a prediction, and includes acid-base theory.	*"Hypothesis- If sulfuric acid (H_2SO_4) were added to remove the carbonate (CO_3^{2-}) and bicarbonate (HCO_3^-), then CO_2 would be formed, and a phase separation would be induced in order to switch the DMCA back to complete the extraction to test for fluorescence intensity"*
2	Student has a testable hypothesis and makes a prediction but does not include acid-base theory.	*"In my first experiment my hypothesis would be that by adding acid stronger than bicarbonate to DMCA solution then a reaction would occur with carbonate that would switch DMCA from being hydrophilic back to hydrophobic"*
1	Student has a testable hypothesis, but it does not make a prediction or include acid-base theory.	*"In order to induce separation, acid can be added to DMCA to complete the extraction."*
0	No characteristics of a hypothesis.	*"Because of the reactions between HCO_3^- and acids, this would be an effective method of separating the two substances."*

For the variable manipulation rubric, shown in Table 2, a student's score was based on the number of accurately defined variables in their experimental design. If a student correctly labeled a variable as independent, dependent, control, or confounding then it was considered accurate. Meaning, if a student labeled a variable as being the control and treated it as a dependent variable in their design, it was not considered accurate and was not counted. Students who accurately defined three or more variables were given a score of three. If they only accurately defined two, they were scored a two. And if they only accurately defined one variable, they were scored a one. If students did not have any accurate variables defined or did not include any variables at all, they were given a score of zero.

Table 2: Variable Scoring Rubric

Score	Description	Student Example
3	Student identifies three types of variables. (Independent, dependent, control, or confounding variable)	*"The variables being changed will be the acid used as a reagent in the separation. The acids used will be hydrochloric acid, hydrogen iodide, acetic acid, carbonic acid and water as a control group. Furthermore, the dependent variable will be the fluorescence intensity. The control variables will be volume of acid used, volume of DMCA, and volume of* CO_2*"*
2	Student identifies two types of variables. (Independent, dependent, control, or confounding variable)	*"The variables that will be controlled are the starting amounts of DMCA, water, and* CO_2*. The variable that will be changed is the reagent's property (acid versus base)."*
1	Student identifies one type of variable. (Independent, dependent, control, or confounding variable)	*"Independent variable: amount of HBr added (mL), dependent variables: carbonate and bicarbonate* (Note: not the dependent variables)"
0	No mention of variables.	*"The environmental factors, such as temp, will remain constant. The only variable that is being changed is adding 500 uL HCl to the mixture."*

For the last rubric, shown in Table 3, I scored student responses based on whether they included possible outcomes and conclusions for their experiment. I defined outcomes as being directly observable from the data. Whereas a conclusion is what a student could conclude regarding the observation. Students who included both an outcome and conclusion were given a score of three. Given these definitions, students who included a conclusion but not a defined outcome were considered more sophisticated and were given a score of two. Students who only included outcomes were given a score of one. Lastly, if a student did not include either in their design, they were given a score of zero.

Table 3: Outcomes and Conclusions Scoring Rubric

Score	**Description**	**Student Example**
3	Student identifies potential outcomes and conclusions.	*"The phases will separate, and it can be concluded that bases induced a change in the chemical state of the DMCA to its neutral form and induced phased separation. The phases will not separate, and it can be concluded that bases did not induce a change in the chemical state of the DMCA to its neutral form or induce phased separation"*
2	Student identifies potential conclusions.	*"Compare the fluorescence results obtained when acid or base was used. The result with high fluorescence intensity is preferred for the extraction of benz[a]anthracene since it yields more product in various conditions9different temperatures)"*
1	Student identifies potential outcomes.	*"The possible outcomes would be a range of effectiveness of bases inducing a switch of DMCA from effective to negligible."*

0	No mention of potential outcomes or conclusions.	*"The first experiment I would attempt is a distillation. Distillation is used to separate different mixtures of a solution by boiling points. It takes into account that the mixtures have different boiling points and eventually one of the substances will boiling off before the other. DMCA has a boiling point of 320 ºF ,while benz[a]anthracene has a boiling point of 820.4 ºF. Thus, in the end of the experiment DMCA will boil off completely leaving us with just benz[a]anthracene."*

After an initial rubric for each was created, interrater reliability was calculated with three researchers. Each person coded a set of responses (composed of 10% of our initial and final drafts) that included initial and final drafts from students in the scaffolded group and non-scaffolded groups. After the first round, the agreement between researchers was found to be α=0.554 for the hypothesis rubric, α=0.478 for the outcomes and conclusion rubric, and α=0.505 for the variable rubric. After the first round, we discussed how we were interpreting the rubric and clarifications that needed to be made in the rubric. From this discussion, I clarified what constituted a discussion of acid-base theory in the hypothesis, that making a prediction was signaled by future tense, and that if a variable was identified incorrectly then it was not counted. Once these clarifications were added to the rubric, researchers coded the data again. After the second round of coding, the agreement between researchers increased to α=0.783 for the hypothesis rubric, α=0.745 for the outcomes and conclusion rubric, and α=0.785 for the variable rubric. These values are within the acceptable agreement range for coders; therefore no more changes were made to the rubric and the head researcher coded the rest of the student drafts.

3.4.2 Peer Review Analysis

The feedback students gave each other was categorized using the rubric in Table 2 to determine if the structure of the peer review template had any effect on the type of feedback students gave (RQ1). I adapted a rubric created by Patchan et al. (2015) who investigated how reviewer ability affected the types of comments they provided their peers, shown in Table 4. They used a theoretical model of feedback to develop a rubric that codes the type of feedback and the focus of the feedback (Patchan & Schunn, 2015). I chose to use the feedback rubric created by Patchan et al. (2015) because it has been used throughout the peer review literature (Patchan & Schunn, 2015; Patchan, Schunn, & Clark, 2018; Patchan et al., 2016) and I wanted to connect our study to the broader peer review community. However, the rubric was meant for longer written assignments (such as writing-to-learn assignments), so I modified the rubric slightly to better fit our data. The first modification I made was removing the localization code and substance code. The localization code captures instances when students specify where the problem is that they are talking about in the feedback. Since student responses to our task were much shorter than most written assignments, reviewers did not need to localize their comments. The substance code is meant to capture feedback that points out content that a student is missing in their response. I found that the other codes in the rubric accounted for this already in the student responses (most likely again due to the shorter nature of their responses), so it was also removed from our final rubric. The final modification I made was transforming the codes into categories. In the present study, students were instructed to provide two to three sentences of feedback. Whereas in the other studies that utilized the original rubric, students provided much more feedback due to the length of the written assignments they were reviewing. Therefore, I used the 'type of feedback' codes and 'focus of feedback' codes to create nine groups to categorize the overall nature of the feedback students gave, shown in Table 5.

Table 4: Adapted from Patchan and Schunn (2015)

Category	Definition	Example
All comments		
Praise	A positive feature of the paper	"*It was a good job explaining differences between the MSNBC article and the article from the scientific journal*"
Problem	Something wrong with the paper	"*The writer did not offer insight into casual and correlational relationships*"
Solution	How to fix a problem or improve the quality of the paper	"*Also, I would suggest writing a stronger conclusion to the end of the paper*"
Criticism comments only (i.e., problems and solutions)		
Localization	Where the issue occurred	
Low Prose	An issue dealing with the literal text choice-usually at a word level	"*Why you say, 'the hypotheses and whether those hypotheses were proven', I think you would say 'that hypothesis' or 'the hypothesis' because it's just one hypothesis*"
High Prose	High-level writing issues (e.g., clarity, use of transitions, strength of arguments, provision of support and counter-arguments, insight)	"*I do not understand what the argument is as it isn't very clear. 'Another peer suggested 'use your own voice in order to capture the reader's attention*"
Substance	An issue with missing, incorrect, or contradictory content	"*I don't see where you stated the independent and dependent variables*"

Table 5: Present Study's Peer Review Coding Scheme

Category	Description	Student Examples

Praise or Summary	The review consists of positive statements and points out no concerns or suggestions for improvement, or the review only summarizes what the argument said.	*"Your two experiments sound good and are very organized."*
Low Prose Problem	The review vaguely describes a flaw within the argument that is not one of the criteria given.	*"The user stated which volume they believe to be the optimal volume. They clearly explained why, higher the concentration of benz[a]anthracene, they higher the fluorescence will be.* ***It was not stated that this information was pulled from the graph shown above the question****."*
High Prose Problem	The review specifically points out a flaw within the argument, offers a counterargument to the student, or asks a question about the student's argument. This includes pointing out a problem with one of the criteria given.	*"The discussion of acid-base theory is slight, and does not explain how it frames the experiment. There is no clear hypothesis, but can see the reasoning of how the base or acid would rect. The methods of the experiment are clear, concise, and detailed. The statement at the end of the paragraph shows that there are possible conclusions that can be drawn from the data they collect."*
Low Prose Solution	The recommendation made by the reviewer is vague, superficial, or stylistic, and no specific flaw within the argument is pointed out. It does not have them fix one of the criteria given. Future tense signals a solution.	*"This is very specific! I would suggest going back and look at the grammar."*

High Prose Solution	The recommendation indicates the author should add a specific component or consider additional information to improve the argument. This includes changes to the arguments that could change the meaning of argument, including changing one of the criteria given. Future tense signals a solution.	*"I love how organized your experiments are. The first experiment is very good and well stated.* ***My only suggestion would be to change the possible outcome to something that talks about comparing bases to acid rather than using the base as a baseline to determine how well acid is working.*** *Your second experiment is very well written. I would not change anything. Great job!"*
Low prose problem and low prose solution	The suggestion that the reviewer offers is vague, the problem is also vaguely stated.	*"Your explanation was good. It hit all the points that were requested: which volume should be used, data that supports your reasoning, and an explanation.* ***The only thing that I would suggest adding is including DMCA in your labels. This will make sure your intention is very clear.****"*
Low prose problem and high prose solution	The recommendation that the reviewer offers indicates a specific component the writer should add to the argument, the **problem stated vaguely**.	*"It would be very helpful if you wrote out more of a step-by-step procedure. By saying that you are adding acid it doesn't exactly explain the chemistry behind how the compound would change properties. I think if you just explained more of the procedure and why each step is important individually, it* ***would help the reader better understand*** *the methods. This same advice would go for both experiment 1 and 2."*

High prose problem and high prose solution	The recommendation that the reviewer offers indicates the writer should add to the argument or consider additional information, the problem is clearly stated.	*"The experiment was broken down concisely to ensure that each detail was covered. I would recommend going into more detail about exactly how the experimental variables will be manipulated. What concentrations of acids and bases will be tested and how will the experiment be set up? What roles do the control variables play in the experiment? Other than experimental set up, the hypotheses are good and offer an explanation of what the experiment will test."*
High prose problem and low prose solution	The reviewer notes a specific flaw within the argument or provides a counterargument, the reviewer offers a **vaguely stated suggestion.**	*"Is the 500uL the highest volume? I don't think that it is.* ***Maybe you could reword that****. Also are we looking for a greater chance of a higher outcome of fluorescence intensity? The way that the question is worded there should be something to do with an extraction."*

After an initial rubric was created, two additional researchers were trained on how to use the rubric and interrater reliability was performed. During the first round of interrater reliability, researchers individually categorized a set of peer reviews (composed of 10% of our total peer review received) from the control and experimental group. After the first round, the agreement between researchers was α=0.595. After discussing differences in how we were interpreting the categories, several clarifications were made in the rubric. First, it was decided that a peer review would only be considered high prose if it pointed out a problem or solution with one of the criteria used to evaluate experiments: (1) discussion of acid-base theory and how it frames their hypothesis, (2) what data will be gathered and recorded, (3) how much data is needed to support valid conclusions, (4) limitations in the precisions of the data that will be collected, (5) identification of the independent, dependent, and control variables, (6) consideration of possible confounding variables, and (7) possible conclusions they will be able to draw from the data they collect (NGSS Lead States, 2013). Second, it was clarified that if a student used future tense in their peer review, this signaled a solution and not a problem. Once these clarifications were made to the rubric, the researchers categorize another set of peer review data. After the second round, the interrater reliability between research was α=0.770. These values are within the acceptable agreement range for Krippendorff's alpha; therefore no more changes were made to the rubric, and the head researcher categorized the rest of the peer review.

3.4.3 Statistical Analysis

3.4.3.1 Research Question One: How does the structure of the peer review template impact the feedback students give for an experimental design task?

To test if there was a difference in the frequency of feedback students gave, a chi-square test of independence was done first to determine if the type of feedback given was influenced by the peer review template. Independent chi-square tests determine whether or not two categorical variables are related to each other, in this case, if the type of feedback given was independent of the group membership (Mchugh, 2013). To compute chi-square, a contingency table with the frequencies of each combination of group membership and type of feedback given was created. After computing the chi-square test of independence, adjusted standardized residuals were calculated to guide what post-hoc testing would be done. The adjusted standardized residual calculations determine what cells of the contingency table (i.e., combinations of group membership and feedback type) significantly contributed to the chi-square value. This is done by comparing the expected and observed values for each combination or cell. The larger the adjusted standardized residual is, the more impact it had on the chi-square value. Lastly, the results from the adjusted standardized residuals were used to guide post-hoc comparison testing. If a combination of group membership and feedback type was found to be significant, then a comparison test was done to determine if the frequencies of that feedback were significantly different between the two groups (i.e., the scaffolded group and the non-scaffolded group). Pearson's chi-square was used for the comparison testing as it can be used to determine the homogeneity of the data when comparing one variable across two groups. A significant Pearson's chi-square indicated that the two groups (i.e., scaffolded, and non-scaffolded) had significantly different frequencies of that type of feedback given. If multiple comparisons are done, this increases the potential for Type 1 error in the calculations (i.e., false positives) (Goldman, 2008). Therefore, when more than one comparison test was performed, a Bonferroni correction was used to limit the Type 1 error (Rupert Jr, 2012).

3.4.3.2 Research Questions Two: How does the structure of the peer review template impact students' revisions for an experimental design task?

Before statistical tests were run on the experimental design results, students were categorized based on how their scores changed from the initial to final draft. If students submitted the same draft for the initial and the final submission, then they were placed in the "no change" group. If students submitted different drafts, but there was no improvement in their score they were categorized as "no change in score". For example, if a student only fixed grammatical or formatting issues, they would be placed in the "no change in score" category. Lastly, if a student's score improved from the initial to final draft, then they were placed in the "score improvement" category. To test if there was a difference in the frequency of revisions students made, a chi-square test of independence was first done to determine if the type of revision made was independent of group membership. Independent chi-square tests determine whether or not two categorical variables are related to each other, in this case, if the category or type of revision was independent of the group membership (i.e., the scaffolded and non-scaffolded group) (Mchugh, 2013). After computing the chi-square test of independence, adjusted standardized residuals were calculated to inform what post-hoc testing would be done. The adjusted standardized residual calculations determine what cells of the contingency table (i.e., combinations of group membership and type of revision) significantly contributed to the chi-square value. Then, post hoc comparison tests were computed for combinations that significantly contributed to the chi-square value. Pearson's chi-square was again used to determine the homogeneity of the frequencies across the scaffolded and non-scaffolded groups. A significant Pearson's chi-square indicated that the two groups had significantly different frequencies of that type of revision made. When more than one comparison test was performed, a Bonferroni correction was used to limit Type 1 error (Rupert Jr, 2012).

CHAPTER 4: RESULTS

4.1 Research Question One: How does the structure of the peer review template impact the feedback students give for an experimental design task?

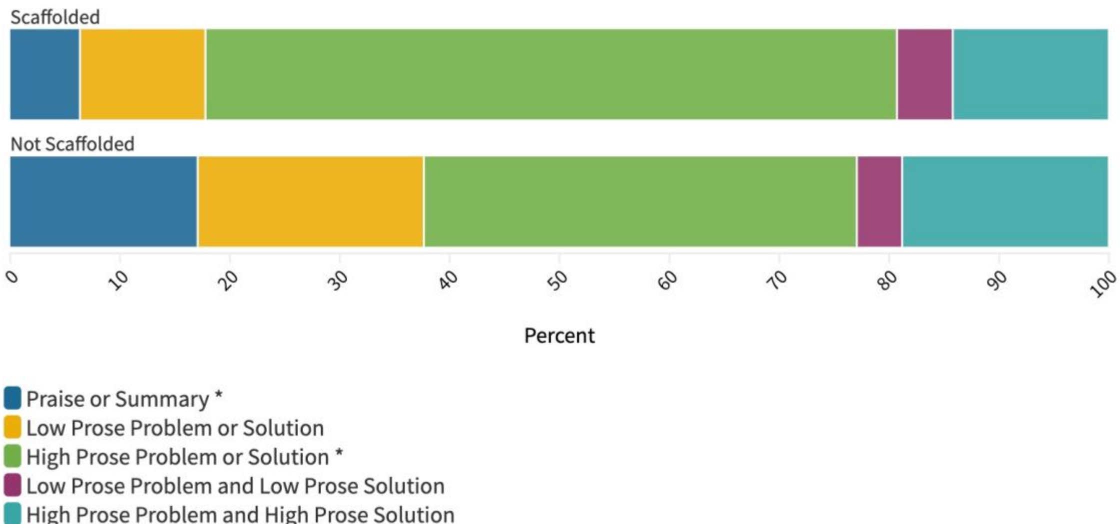
4.1.1 Chi-Square Test of Independence

I calculated the chi-square test of independence to determine if the type of feedback students gave was independent of the type of peer review template used. An alpha value of 0.05 was selected as the threshold for determining significance. I used the contingency table, shown in Table 6, for our chi-square test of independence and residual analysis. The distribution of peer review feedback for the scaffolded and non-scaffolded group are shown in Figure 2. To perform a chi-square test of independence, a minimum frequency of 5 is needed in each category. For this reason, I collapsed some of our original peer review categories before running the test. I combined the "high prose problem" and "high prose solution" categories, and then combined the "low prose problem" and "low prose solution" categories. Additionally, I removed the "high prose problem and low prose solution" and "low prose problem and high prose solution" categories because they fell below the cutoff for the scaffolded group and non-scaffolded group and combining them did not reach the minimum.

Table 6: Peer Review Contingency Table

	Scaffolded	Non-Scaffolded	Totals
praise or summary	25	29	54
low prose problem or solution	45	35	80
high prose problem or solution	148	67	215
low prose problem and low prose solution	10	7	17
high prose problem and high prose solution	56	32	88
Totals	284	170	454

Chi-Square Test: P=0.0242



*Significant threshold of ±1.96

Figure 5: Distribution of Peer Review Given

It was concluded that the type of feedback given was not independent of the peer review template (scaffolded and non-scaffolded). The results of the chi-square test of independence were P=0.0242. Since the p-value is less than 0.5, the null hypothesis was rejected. Our hypotheses for the chi-square test of independence were as follows:

H_o: The type of feedback given is independent of the peer review template.

H_a: The type of feedback given is not independent of the peer review template.

4.1.2 Adjusted Residuals

After testing for independence, adjusted standardized residuals (referred to as adjusted residuals from here on) were calculated for each cell in the contingency table. The chi-square of independence provides insight into the relationship between the variables, whereas the adjusted residuals give insight into what is driving that relationship (Agresti, 1990). The adjusted residual identifies what cells in the contingency table made the greatest contribution to the chi-square test of independence by comparing the observed value and expected value. The cells that have an absolute value of 1.96 or more indicate a lack of fit with the null hypothesis, meaning they contributed significantly to the rejection of the null hypothesis. The sign of the adjusted residual signals whether the actual frequency observed in the contingency table was higher or lower than the expected frequency.

Two cells significantly contributed to the results: the frequency of praise feedback and high prose problem or solution feedback from the non-scaffolded group. The amount of praise feedback from the non-scaffolded group was significantly higher than the expected frequency. The amount of high prose problem or solution feedback from the non-scaffolded group was also significantly lower than the expected frequency. The residual analysis results are shown in Table 7.

Table 7: Peer Review Adjusted Residuals

	Scaffolded	Non-Scaffolded
praise or summary	-1.672	2.128*
low prose problem or solution	-0.833	1.051
high prose problem or solution	1.913	-2.287*
low prose problem and low prose solution	-0.201	0.258
high prose problem and high prose solution	0.152	-0.192

*Significant threshold of ±1.96

4.1.3 Post-hoc Testing

I used the results from the residual analysis to inform what cells post-hoc testing would be performed on. Adjusted residuals provide information about how the observed values compare to the expected values, whereas post-hoc testing compares the conditions to determine if they are significantly different (Franke, Ho, & Christie, 2012). The utility in calculating adjusted residuals first is that they can be used to direct what post hoc testing is done (Sharpe, 2015). Since the amount of praise feedback from the non-scaffolded group was significantly higher than expected, a comparison test was done to determine if the amount of praise feedback from the non-scaffolded group was significantly different from the amount of praise feedback given by the scaffolded group. A second comparison test was done to determine if the amount of high prose problem or solution feedback from the non-scaffolded group was significantly different from the amount of high prose problem or solution feedback given by the scaffolded group. The threshold for significance, after the Bonferroni correction, was α=.025. The contingency tables for the post-hoc testing are shown in Table 8 and Table 9.

It was concluded that there was a significant difference in the frequency of praise feedback given when using the scaffolded template versus the non-scaffolded template. The Pearson's chi-square test result was P=0.0086. This was less than our corrected alpha value, so the null hypothesis was rejected. The null hypothesis was as follows:

H_o: There is no difference in the frequency of praise between the scaffolded and non-scaffolded groups. elicited by the peer review templates.

H_a: There is a difference in the frequency of praise between the scaffolded and non-scaffolded groups.

Table 8: Praise or Summary Contingency Table.

	Scaffolded	Non-Scaffolded	Total
Praise or Summary	25	29	54
Other	259	142	401

It was also concluded that there was a significant difference in the frequency of high prose problem or solution feedback given when using the scaffolded template versus the non-scaffolded template. The Pearson's chi-square test for this comparison was

P=0.0087. This was less than our corrected alpha value, so the null hypothesis was rejected. The null hypothesis was as follows:

H_o: There is no difference in the frequency of high prose problems or solutions between the scaffolded and non-scaffolded groups.

H_a: There is a difference in the frequency of high prose problems or solutions between the scaffolded and non-scaffolded groups.

Table 9: High Prose Problem or Solution Contingency Table

	Scaffolded	Non-Scaffolded	Total
High Prose Problem or Solution	148	67	215
Other	136	103	239

4.2 Research Question Two: How does the structure of the peer review template impact students' revisions for an experimental design task?

4.2.1 Chi-Square Test of Independence

I used the contingency table, shown in Table 10 for our chi-square test of independence and residual analysis. The Sankey diagram in Figure 6 and Figure 7 illustrates how students' scores changed from their initial draft to the final draft in the scaffolded group and the non-scaffolded group. A Sankey diagram is a type of flow diagram used to visualize data. I calculated the chi-square test of independence to determine if the frequency of student revision type was independent of the type of peer review template used. The three revision categories were (1) revision with score improvement, (2) revision with no score improvement, and (3) no revision. An alpha value of 0.05 was selected as the threshold for determining significance.

It was concluded that the type of revision made was not independent of the peer review template (scaffolded and non-scaffolded). The result of the chi-square test of independence was less than 0.001. Since the p-value is less than 0.05, the null hypothesis was rejected. Additionally, the non-scaffolded group had no students that received a score of nine and there were no scaffolded students in the lower third of scores after engaging in peer review. Our hypotheses for the chi-square test of independence were as follows:

H_o: The type of revision is independent of the peer review template.

H_a: The type of revision is not independent of the peer review template.

Table 10: Revisions Contingency Table

	Non-Scaffolded	Scaffolded	Totals
No Revision	24	16	40
Revision with No Score Improvement	33	33	66
Revision with Score Improvement	19	70	89
Totals	76	119	195

Chi-Square Test: P=0.000013

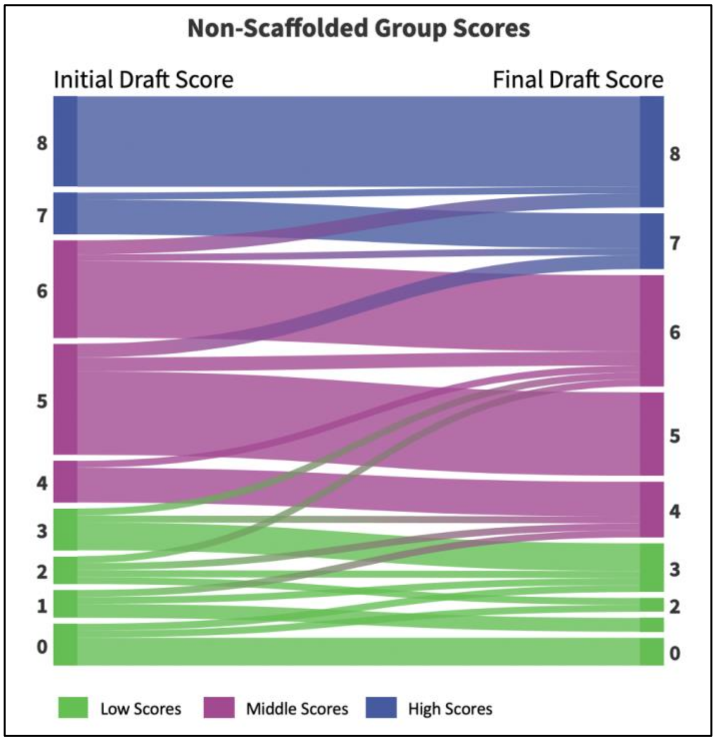


Figure 6: Non-Scaffolded Group Initial and Final Scores

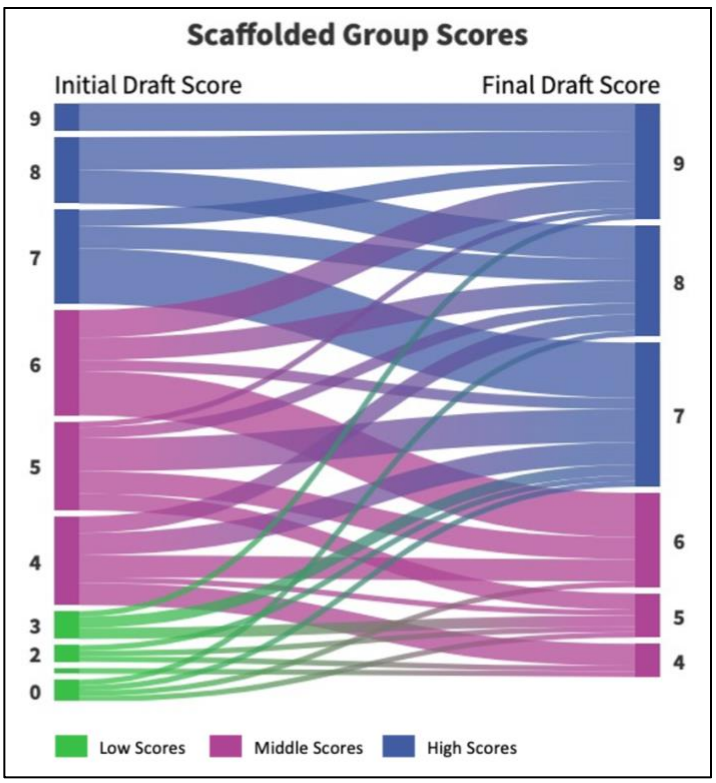


Figure 7: Scaffolded Group Initial and Final Score

4.2.2 Adjusted Residuals

After testing for independence, adjusted residuals were calculated for each cell in the contingency table. The adjusted residual identifies what cells in your contingency table made the greatest contribution to the chi-square test of independence by comparing the observed value and expected value. The cells that have an absolute value of 1.96 or more indicate a lack of fit with the null hypothesis, meaning they contributed significantly to the rejection of the null hypothesis.

Four cells significantly contributed to the results: the frequency of "revision with score improvement" for the scaffolded group, "no revision" for the scaffolded group, "revision with score improvement" for the non-scaffolded group, and "no revision" for the non-scaffolded group. The sign of the adjusted residual signals whether the actual frequency observed in the contingency table was higher or lower than the expected frequency. The amount of "revision with score improvement" from the scaffolded group was significantly higher than the expected frequency. The amount of "no revision" from the scaffolded group was significantly lower than the expected frequency. The amount of "revision with score improvement" from the non-scaffolded group was significantly lower than the expected frequency. The amount of "no revision" from the non-scaffolded group was significantly higher than the expected frequency. The residual analysis results are shown in Table 11.

Table 11: Revisions Adjusted Residuals

	Non-Scaffolded	Scaffolded
No Revision	2.490*	-2.041*
Revision with No Score Improvement	1.893	-1.582
Revision with Score Improvement	-3.984*	3.399*

*Significant

4.2.3 Post-hoc Testing

I again used the results from the residual analysis to inform what cells post-hoc testing would be performed on. A comparison test was done to determine if the amount of “revision with score improvement” from the scaffolded group was significantly different from the amount of “revision with score improvement” given by the non-scaffolded group. A second comparison test was done to determine if the amount of “no revision” made by the non-scaffolded group was significantly different from the amount of “no revision” made by the scaffolded group. The threshold for significance, after the Bonferroni correction, was α=.025. The contingency tables for the post-hoc testing are in Table 12 and Table 13.

It was concluded that there was a significant difference in the frequency of "no revision" for those in the scaffolded group compared to those in the non-scaffolded group. The Pearson’s chi-square test result was P=0.0022. This was less than our corrected alpha value, so the null hypothesis was rejected. The null hypothesis was as follows:

H_o: There is no difference in the frequency of "no revision" between the scaffolded and non-scaffolded group.

H_a: There is a difference in the frequency of "no revision" between the scaffolded and non-scaffolded group.

Table 12: “No Revision” Contingency Table

	Non-Scaffolded	Scaffolded	Total
No Revision Made to Draft	24	16	40
Other	52	103	155

It was also concluded that there was a significant difference in the frequency of "revision with score improvement" for those in the scaffolded group compared to those in the non-scaffolded group. The Pearson's chi-square test for this comparison was less than 0.001. This was less than our corrected alpha value, so the null hypothesis was rejected. The null hypothesis was as follows:

H_o: There is no difference in the frequency of "revision with score improvement" between the scaffolded and non-scaffolded group.

H_a: There is a difference in the frequency of "revision with score improvement" between the scaffolded and non-scaffolded group.

Table 13: "Revision with Score Improvement" Contingency Table

	Non-Scaffolded	Scaffolded	Total
Revision with Score Improvement	19	70	89
Other	57	49	106

4.3 Score Revision

As shown in Figure 3, the non-scaffolded group had no students that received a score of nine (highest possible score), and there were no scaffolded students in the lower third of scores after engaging in peer review. The only students who improved their score to an 8 in the non-scaffolded group had a score of 6 or higher on their initial draft. Whereas in the scaffolded group, 16 students improved their scores to a 9. Students with initial draft scores from the lower, middle, and high tier improved their score to 9 in the scaffolded group. Table 14 shows a student in the scaffolded group who improved their score from a 3 to a 9. In the non-scaffolded group, the highest score improvement seen for a lower-tier score was from a 2 to a 6, shown in Table 15.

Table 14: Scaffolded Student Example

	Initial Draft	Final Draft
Hypothesis	No Hypothesis	"***Acids can theoretically remove carbonate and bicarbonate in the form of CO_2 in order to induce phase separation.*** [theory and prediction] *This experiment will test this hypothesis and* ***determine whether or not an acid would be a good reagent for the extraction of benz[a]anthracene***. [testable]" Score: 3
Variables	No Variables	"*Mix two solutions of DMCA, CO_2 and water,* ***holding the volumes of each component constant*** [Control Variable] *as well as the temperature. In one of the solutions, add an acid such as HCl.* ***The presence of the acid is the independent variable***. [independent variable] *Test for the presence of CO_2 in both solutions, using a method such as a reaction with lime water*. ***The presence of CO_2 is the dependent variable.***" [dependent variable] Score: 3
Outcome and Conclusion	"***If there is CO2*** [outcome], *it can be concluded that the idea that acids remove carbonate and bicarbonate from the solution is correct, and this would induce layer separation.* ***Therefore, an acid would be a good reagent.***" [conclusion] Score: 3	"*If there is CO_2 in the solution to which the acid was added, it can be concluded that the idea that acids remove carbonate and bicarbonate from the solution is correct, and this would induce layer separation. Therefore, an acid would be a good reagent.* ***If there is no CO_2 present in either solution*** [acid or base solution], ***then there can be no conclusion drawn and there is likely an issue with the method used to test for the presence of CO_2***." Score: 3

Here we see that the scaffolded student went from having no hypothesis or variables in their initial draft to including them in their final draft. The student received the highest score possible in both these categories in their final draft. Their final draft hypothesis was testable, made a prediction, and included theory about how the switching occurs. They also correctly identified three different variables after identifying none in their initial draft. Lastly, the student also made changes to their outcomes and conclusion. Even though there were no score changes in the outcomes and conclusions from the initial to final draft, the student added an additional outcome and conclusion about if there was no carbon dioxide seen with an acid or base.

Table 15: Non-Scaffolded Student Example

	Initial Draft	Final Draft
Hypothesis	"*For this experiment, it is predicted that liquid chromatography with* ***UV/fluorescence will help determine and switch back to acid or base*** *because this process uses both qualitive and quantitative data for organic compound such as: high resolution, sensitivity, and selectivity.*" [UV does not switch the compound. The acid and base are not what is being switched.] Score: 0	"*The bases will induce a change in the chemical state of the DMCA to its natural form and might induce phase separation.*" [theory, prediction, and testable] Score: 3
Variables	"*Changed: various wavelength*" [Independent variable] Score: 1	"*Controlled: Amount of water in DMCA* [Control Variable] *Changed: Amount of the base added* [Independent Variable]" Score: 2

Outcome and Conclusion	"*Overall, possible outcomes that can arise from this experiment could be: ecofriendly, fast extraction, and high efficiency."* [possible outcomes] Score: 1	"*Possible outcomes: Phase separation will increase if base is increased*." [possible outcome] Score: 1

In the non-scaffolded student's initial draft, they misunderstood the concepts and theories presented in the task about how the switching occurs, what is switching, and how ultraviolet-visible spectroscopy (UV-vis spectroscopy) is utilized in determining the amount of benz[a]anthracene extracted. They base their initial hypothesis on UV-vis spectroscopy switching the acids and bases, which is not how the mechanism occurs. This also does not provide insight into the two theories I tasked them to investigate about the utility of acids and bases for switching the DMCA. Both factors contributed to the student receiving a score of zero on their initial hypothesis as it is not testable given the mechanism presented in the task. Even though their hypothesis received a score of 0, the student did receive 1 point for correctly identifying an independent variable for the hypothesis they wrote. If UV-vis spectroscopy was responsible for the switching, then different wavelengths would be an appropriate independent variable for the said experiment. Lastly, they received a score of 1 for including potential outcomes for the experiment they described. In their final draft, the student resolved confusion about the mechanism and provided a testable hypothesis which included a prediction and theory. They also improved their variable score by including a control variable and a new independent variable. However, they did not include a dependent or compounding variable, leading to a score of 2 for the variable category. Finally, we see no score improvement for the outcomes and conclusions. However, they did change the outcome from their initial draft to align with their final draft hypothesis.

CHAPTER 5: DISCUSSION

The current study provided new information about the effects peer review structure has on the type of feedback students give and the kinds of revisions that students make on an experimental design task. This study suggests that providing students with evaluative criteria (given in the scaffolded group) can increase the amount of quality feedback students give. The results showed that the peer review structure affected the rate of praise feedback and high prose feedback students gave. When students used a scaffolded peer review template, they gave significantly more high prose feedback to their peers than students in the non-scaffolded group. A concern with peer feedback has been that students may not be able to provide quality feedback to their peers due to being novices on the topic (Cheng, Liang, & Tsai, 2015; Kaufman & Schunn, 2011). However, I found that no matter the initial draft score, students in both groups were able to provide high prose feedback. The difference in high prose feedback rate between the scaffolded and non-scaffolded group is most likely explained by the added structure in the peer review template. Multiple studies have had similar results that show providing well-structured, student-friendly templates can help students give quality feedback (Patchan et al., 2016; Wang, 2014).

Our study showcases the potential for peer review to support students when designing experiments, supporting previous findings in the literature (Basso, 2020; J. Walker et al., 2012). Students in the non-scaffolded group gave significantly more praise feedback than the scaffolded group with the scaffolded group giving significantly higher prose feedback. It is desirable for students to give high prose feedback as it is more likely to improve the quality of a students' response (Patchan et al., 2016). Patchan et al. (2016) found that when students implemented high prose feedback, it was more likely to improve the quality of their paper than when students implemented low prose, less substantive feedback. However, students were less likely to implement high prose feedback in their revisions compared to low prose and localized feedback. Our study found that the scaffolded group made more revisions with score improvement than the non-scaffolded group, suggesting that high prose feedback contributed to the score improvements seen and did not impede student revisions. Wu and Schunn (2020) had similar findings that showed feedback quality is a predictor for student implementation of comments when making revisions.

Score improvements were seen in both the scaffolded and non-scaffolded group, but no students in the non-scaffolded group were able to raise their score above an 8. Every student who scored an 8 in the non-scaffolded group received a 2 in the variable category, meaning they only correctly identified two variables. This aligns with other findings that have shown that students struggle to identify variables, especially the independent and dependent variables (De Jong & Van Joolingen, 1998; Lawson, 2002). When comparing the scaffolded group and non-scaffolded group, I found that students in the scaffolded group with low initial scores improved their scores more than the non-scaffolded group. Students in the scaffolded group with low scores were all able to reach mid (4-6) and high (7-9) scores on their final draft. Whereas no students with low-level scores in the non-scaffolded group were able to reach a high score. Additionally, many students in the non-scaffolded group with low-level scores stayed in the low score range compared to the scaffolded group which had no students in the low score range on their final draft. This further suggests that peer review played a role in students improving their scores, with a more scaffolded peer review template providing more support to students than the non-scaffolded peer review template.

I observed two kinds of revisions commonly made by students: (1) the addition of more components (i.e., adding theory to their hypothesis, identifying additional variables, and incorporating potential conclusions) and (2) completely revising their experimental design. Students in the scaffolded group added more to their initial drafts than students in the non-scaffolded group. In the scaffolded example, Table 14, the student added a hypothesis, additional variables, and an additional outcome and conclusion to their initial draft. This resulted in a 6-point increase in their score from the initial to final draft. Students in the non-scaffolded group made similar revisions to this one, however, students did not add as much to their initial draft as did in the scaffolded group. This is shown in Figure 6, where there were no students in the non-scaffolded group receiving a score higher than 8. The second kind of revision made by students involved completely revising their experimental design to better align with the theory described in the task. I observed this happening in both groups where students would design experiments not informed by the mechanism in the task and then later resolve this issue in their final draft. In the non-scaffolded example, Table 15, the student misunderstood the mechanism by which the switching occurs. However, in the final draft, the student completely revised their original experiment and aligned their hypothesis, variables, and outcomes with the mechanism presented in the paper.

The revisions I observed are similar to that of Berg et al. (2021) who modeled four distinct pathways students took during the generation, evaluation, and revision of a data interpretation task. After the simulated peer review, they found that students who improved their response either adopted a new stance or added to their original response. During the simulated peer review, students generated internal feedback that informed the kinds of changes (or lack thereof) that students made by comparing their work to the sample responses. According to Nicol (2021), when making comparisons students gather new information for evaluating their own work and then modify their approach to the given task. I cannot say for sure that the students in our study followed similar routes of generating internal feedback to make changes to their drafts. However, the following peer review comments left by students suggest that the generation of internal feedback may have been what contributed to the revisions students made.

> "***I believe that the things I posted for this question are not correct,*** *but I tried my best. I think that you explained the main hypothesis, the variables, and the conclusions perfectly*." -Student 113
>
> "*You successfully discussed acid-base theory and how it frames your hypothesis. I liked how you described the method for experiment B;* ***it showed me what I need to change in my own experiment!***" -Student 201

In the peer review comments, the students recognized gaps in their responses after reviewing their peers' responses. According to the model developed by Berg et al. (2020), this is one of the steps that leads to students improving upon their initial response, the same types of improvements I observed in our study. The significant differences I observed in the rate of revisions made between the scaffolded and non-scaffolded group suggest that scaffolded peer review could better support students in generating internal feedback on experimental design tasks. With the present study, I cannot be sure if the score improvements are from students generating internal feedback or from students implementing feedback given by their peers. However, it is clear that a more scaffolded peer review template leads to more students making revisions and improving their designs.

5.1 Implications for Research and Teaching

Our investigation of peer review clearly showed that a scaffolded prompt that includes evaluative criteria led to more students giving high-quality feedback. Further, our findings showed that a scaffolded prompt led to more students revising and improving their experimental designs. Students in the scaffolded group were able to obtain higher scores on the final draft and no students scored in the lower tier on their final draft. The scaffolding provided extra support to lower-level students, similar to other findings (van Riesen et al., 2018). Teachers are therefore encouraged to provide scaffolded peer review when employing experimental design activities in the classroom. Providing criteria for students in the peer review template helps them evaluate their peers' work, leading them to give more high-quality feedback. Peer review also gives students an opportunity to compare their work to others. This could help students generate helpful internal feedback, leading them to revise their work.

The present study provides insight into how peer review can be used to support students when designing experiments. Scaffolding in the peer review template led to more students improving their initial draft scores and no students receiving a low score on the final draft. As previously mentioned, I am not able to determine what about the peer review process led to more students making revisions on their experimental designs. According to Wu and Schunn (2020), feedback quality is a predictor in the implementation of feedback by a student. However, other studies would suggest that engaging in peer review triggers the generation of internal feedback, leading to revision (Berg & Moon, 2022; Nicol, 2019). The feedback given by students in our study suggests that students compared their work to their peers with students, evaluating the correctness of their own work. Whether students were implementing feedback or generating feedback, the significant gains made by students in the scaffolded group provide evidence that scaffolded peer review supports students when designing experiments. Future research should investigate the ability of peer review to support students engaged with other science practices.

References

Agresti, A. (1990). Inference for two-way contingency tables. *Categorical Data Analysis*, 36–78.

Alexander, P. A., & Judy, J. E. (1988). The Interaction of Domain-Specific and Strategic Knowledge in Academic Performance. *Review of Educational Research*, *58*(4), 375–404. https://doi.org/10.3102/00346543058004375

Anker-Hansen, J., & Andrée, M. (2015). More Blessed to Give Than Receive – A Study of Peer-assessment of Experimental Design. *Procedia - Social and Behavioral Sciences*, *167*, 65–69. https://doi.org/10.1016/j.sbspro.2014.12.643

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

Azevedo, R., & Bernard, R. M. (1995). A Meta-Analysis of the Effects of Feedback in Computer-Based Instruction. *Journal of Educational Computing Research*, *13*(2), 111–127. https://doi.org/10.2190/9lmd-3u28-3a0g-ftqt

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*, *61*(2), 213–238. https://doi.org/10.3102/00346543061002213

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Barzilai, S., & Chinn, C. A. (2018). On the goals of epistemic education: Promoting apt epistemic performance. *Journal of the Learning Sciences*, *27*(3), 353–389. https://doi.org/10.1080/10508406.2017.1392968

Basso, A. (2020). Results of a Peer Review Activity in an Organic Chemistry Laboratory Course for Undergraduates. *Journal of Chemical Education*, *97*(11), 4073–4077. https://doi.org/10.1021/acs.jchemed.0c00373

Beishuizen, J., Wilhelm, P., & Schimmel, M. (2004). Computer-supported inquiry learning: effects of training and practice. *Computers & Education*, *42*(4), 389–402.

Berg, S. A., & Moon, A. (2022). Prompting hypothetical social comparisons to support chemistry students' data analysis and interpretations. *Chemistry Education Research and Practice*, *23*(1), 124–136. https://doi.org/10.1039/d1rp00213a

Bienstock, J. L., Katz, N. T., Cox, S. M., Hueppchen, N., Erickson, S., & Puscheck, E. E. (2007). To the point: medical education reviews-providing feedback. *American Journal of Obstetrics and Gynecology*, *196*(6), 508–513. https://doi.org/10.1016/j.ajog.2006.08.021

Butler, D. L., & Winne, P. H. (1995). Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research*, *65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment and Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Carlo, D. Del, & Flokstra, B. (2017). *Sesame Street Picnic Sesame Street Picnic: An Introductory Activity to Claims, Evidence, and Rationale* (Vol. 46).

Cheng, K.-H., Liang, J.-C., & Tsai, C.-C. (2015). Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *The Internet and Higher Education*, *25*, 78–84.

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology*, *103*(1), 73.

Choi, A., Hand, B., & Greenbowe, T. (2013). Students' Written Arguments in General Chemistry Laboratory Investigations. *Research in Science Education*, *43*(5), 1763–1783. https://doi.org/10.1007/s11165-012-9330-1

Collison, C. G., Kim, T., Cody, J., Anderson, J., Edelbach, B., Marmor, W., … Niziol, J. (2018). Transforming the Organic Chemistry Lab Experience: Design, Implementation, and Evaluation of Reformed Experimental Activities - REActivities. *Journal of Chemical Education*, *95*(1), 55–61. https://doi.org/10.1021/acs.jchemed.7b00234

Cooper, M. M., Caballero, M. D., Ebert-May, D., Fata-Hartley, C. L., Jardeleza, S. E., Krajcik, J. S., … Underwood, S. M. (2015). Challenge faculty to transform STEM learning. *Science*, *350*(6258), 281–282. https://doi.org/10.1126/science.aab0933

Council, N. R., Education, D. B. S. S., Education, B. S., & Assessment, N. C. S. E. S. (1996). *National Science Education Standards*. National Academies Press. Retrieved from https://books.google.com/books?id=WprSjvDW0dAC

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. https://doi.org/10.1187/cbe.13-09-0192

De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*(2), 179–201. https://doi.org/10.3102/00346543068002179

Dijkstra, P., Kuyper, H., Van Der Werf, G., Buunk, A. P., & Van Der Zee, Y. G. (2008). Social comparison in the classroom: A Review. *Review of Educational Research*, *78*(4), 828–879. https://doi.org/10.3102/0034654308321210

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's Argument Pattern for studying science discourse. *Science Education*, *88*(6), 915–933. https://doi.org/10.1002/sce.20012

Faize, F. A., Husain, W., & Nisar, F. (2018). A critical review of scientific argumentation in science education. *Eurasia Journal of Mathematics, Science and Technology Education*. https://doi.org/10.12973/ejmste/80353

Finkenstaedt-Quinn, S. A., Snyder-White, E. P., Connor, M. C., Gere, A. R., & Shultz, G. V. (2019). Characterizing Peer Review Comments and Revision from a Writing-to-Learn Assignment Focused on Lewis Structures. *Journal of Chemical Education*, *96*(2), 227–237. https://doi.org/10.1021/acs.jchemed.8b00711

Franke, T. M., Ho, T., & Christie, C. A. (2012). The Chi-Square Test: Often Used and More Often Misinterpreted. *American Journal of Evaluation*, *33*(3), 448–458. https://doi.org/10.1177/1098214011426594

Gardner, D. P. (1983). A Nation At Risk: The Imperative For Educational Reform. An

Open Letter to the American People. A Report to the Nation and the Secretary of Education.

Goldman, M. (2008). Why is multiple testing a problem ? The Bonferroni correction The positive False Discovery Rate, 1–5.

Goodman, J. S., & Wood, R. E. (2004). Feedback specificity, learning opportunities, and learning. *Journal of Applied Psychology*, *89*(5), 809–821. https://doi.org/10.1037/0021-9010.89.5.809

Goodman, J. S., Wood, R. E., & Chen, Z. (2011). Feedback specificity, information processing, and transfer of training. *Organizational Behavior and Human Decision Processes*, *115*(2), 253–267. https://doi.org/10.1016/j.obhdp.2011.01.001

Goodman, J. S., Wood, R. E., & Hendrickx, M. (2004). Feedback Specificity, Exploration, and Learning. *Journal of Applied Psychology*, *89*(2), 248–262. https://doi.org/10.1037/0021-9010.89.2.248

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

Henderson, M., Ryan, T., & Phillips, M. (2019). The challenges of feedback in higher education. *Assessment & Evaluation in Higher Education*.

Hesketh, E. A., & Laidlaw, J. M. (2002). Developing the teaching instinct. *Medical Teacher*, *24*(3), 245–248. https://doi.org/10.1080/014215902201409911

Hmelo, C. E., Holton, D. L., & Kolodner, J. L. (2014). Designing to learn about complex systems. *Design Education: A Special Issue of the Journal of the Learning Sciences*, *9*(2000), 247–298. https://doi.org/10.1207/S15327809JLS0903

Hosbein, K. N., Lower, M. A., & Walker, J. P. (2021). Tracking Student Argumentation Skills across General Chemistry through Argument-Driven Inquiry Using the Assessment of Scientific Argumentation in the Classroom Observation Protocol. *Journal of Chemical Education*, *98*(6), 1875–1887. https://doi.org/10.1021/acs.jchemed.0c01225

Ion, G., Sánchez Martí, A., & Agud Morell, I. (2019). Giving or receiving feedback: which is more beneficial to students' learning? *Assessment and Evaluation in Higher Education*, *44*(1), 124–138. https://doi.org/10.1080/02602938.2018.1484881

James, N. M., & Ladue, N. D. (2021). Pedagogical Reform in an Introductory Chemistry Course and the Importance of Curricular Alignment. *Journal of Chemical Education*, *98*(11), 3421–3430. https://doi.org/10.1021/acs.jchemed.1c00688

Jiménez-Aleixandre, M. P., & Crujeiras, B. (2017). Epistemic practices and scientific practices in science teaching. *Science Education*, 69–80.

Kallery, M., Psillos, D., & Tselfes, V. (2017a). Students' Experimental Design Activities: Do They Promote Scientific Thinking? *Esera*, (July 2019), 93–100.

Kallery, M., Psillos, D., & Tselfes, V. (2017b). Students' Experimental Design Activities: Do They Promote Scientific Thinking? *Esera*, (January), 93–100.

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.

Kennedy, B., Tyson, A., & Funk, C. (2022). Americans' Trust in Scientists, Other Groups Declines. *Pew Research Center*. Retrieved from https://www.pewresearch.org/science/2022/02/15/americans-trust-in-scientists-

other-groups-declines/
Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science*, *36*(5), 757–798. https://doi.org/10.1111/j.1551-6709.2012.01245.x
Krippendorff, K. (2011). Computing Krippendorff ' s Alpha-Reliability. *Departmental Papers (ASC)*, 12. Retrieved from http://repository.upenn.edu/asc_papers
Kuhn, D., Arvidsson, T. S., Lesperance, R., & Corprew, R. (2017). Can Engaging in Science Practices Promote Deep Understanding of Them? *Science Education*, *101*(2), 232–250. https://doi.org/10.1002/sce.21263
Lasarte-Aragonés, G., Lucena, R., Cárdenas, S., & Valcárcel, M. (2015). Use of switchable solvents in the microextraction context. *Talanta*, *131*, 645–649. https://doi.org/10.1016/j.talanta.2014.08.031
Laverty, J. T., Underwood, S. M., Matz, R. L., Posey, L. A., Carmel, J. H., Caballero, M. D., … Cooper, M. M. (2016). Characterizing college science assessments: The three-dimensional learning assessment protocol. *PLoS ONE*, *11*(9), 1–21. https://doi.org/10.1371/journal.pone.0162333
Lawson, A. E. (2002). Sound and faulty arguments generated by preservice biology teachers when testing hypotheses involving unobservable entities. *Journal of Research in Science Teaching*, *39*(3), 237–252. https://doi.org/10.1002/tea.10019
Lefkos, I., Psillos, D., & Hatzikraniotis, E. (2011). Designing experiments on thermal interactions by secondary-school students in a simulated laboratory environment. *Research in Science and Technological Education*, *29*(2), 189–204. https://doi.org/10.1080/02635143.2010.533266
Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, *18*(1), 30–43. https://doi.org/10.1016/j.jslw.2008.06.002
Matthews, M. R. (n.d.). *History, philosophy and science teaching : new perspectives*.
Maydeu-Olivares, A. (2009). *The SAGE Handbook of Quantitative Methods in Psychology*. United Kingdom: SAGE Publications.
Mccomas, W. F. (2011). The History of Science And The Future of Science Education. In *Adapting Historical Knowledge Production to the Classroom*. SensePublishers. https://doi.org/https://doi.org/10.1007/978-94-6091-349-5_3
McConlogue, T. (2015). Making judgements: investigating the process of composing and receiving peer feedback. *Studies in Higher Education*, *40*(9), 1495–1506. https://doi.org/10.1080/03075079.2013.868878
Mchugh, M. L. (2013). The Chi-square test of independence Lessons in biostatistics. *Biochemia Medica*, *23*(2), 143–149. Retrieved from http://dx.doi.org/10.11613/BM.2013.018
Molloy, E., & Boud, D. (2014). Feedback Models for Learning, Teaching and Performance. *Handbook of Research on Educational Communications and Technology: Fourth Edition*, 1–1005. https://doi.org/10.1007/978-1-4614-3185-5
Mory, E. H. (1996). Feedback research revisited. *Most*, 745–784.
Mory, E. H. (2004). *Handbook of Research on Educational Communications and Technology*. (D. H. Jonassen, Ed.). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Murphy, P. K., Greene, J. A., Allen, E., Baszczewski, S., Swearingen, A., Wei, L., & Butler, A. M. (2018). Fostering high school students' conceptual understanding and argumentation performance in science through Quality Talk discussions. *Science Education*, *102*(6), 1239–1264. https://doi.org/10.1002/sce.21471

National Academies of Sciences, Engineering, and M. (2010). Rising above the gathering storm, revisited: Rapidly approaching category 5.

National Research Council. (2007). Committee on Prospering in the Global Economy of the 21st Century: An Agenda for American Science and Technology, Committee on Science, Engineering, and Public Policy. National Academy of Sciences.

National Research Council. (2012). *Discipline based educational research: Understanding and Improving Learning in*. *National Research Council*.

Next Generation Science Standards. (2013). Next Generation Science Standards. Retrieved March 11, 2022, from https://www.nextgenscience.org/#:~:text=A goal for developing the,college%2C careers%2C and citizenship.

NGSS Lead States. (2013). Next Generation Science Standards: For States, by States (Appendix F – Science and Engineering Practices). *Achieve, Inc. on Behalf of the Twenty-Six States and Partners That Collaborated on the NGSS*, (November), 1–103. Retrieved from http://www.nextgenscience.org/next-generation-science-standards

Nicol, D. (2019). Reconceptualising feedback as an internal not an external process. *Italian Jornal of Education Research*, *9744*. https://doi.org/10.7346/SIRD-1S2019-P71

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

Nicol, D., & McCallum, S. (2021). Making internal feedback explicit: exploiting the multiple comparisons that occur during peer review. *Assessment and Evaluation in Higher Education*, *0*(0), 1–19. https://doi.org/10.1080/02602938.2021.1924620

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment and Evaluation in Higher Education*, *39*(1), 102–122. https://doi.org/10.1080/02602938.2013.795518

Olson, S., & Riordan, D. G. (2012). Engage to excel: producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. Report to the president. *Executive Office of the President*.

Patchan, M. M. (2011). *Learning From Revision Using Peer Feedback*. University of Pittsburgh.

Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: how students respond to peers' texts of varying quality. *Instructional Science*, *43*(5), 591–614. https://doi.org/10.1007/s11251-015-9353-x

Patchan, M. M., & Schunn, C. D. (2016). Understanding the effects of receiving peer feedback for text revision: Relations between author and reviewer ability. *Journal of Writing Research*, *8*(2), 227–265. https://doi.org/10.17239/jowr-2016.08.02.03

Patchan, M. M., Schunn, C. D., & Clark, R. J. (2018). Accountability in peer assessment: examining the effects of reviewing grades on peer ratings and peer feedback. *Studies in Higher Education*, *43*(12), 2263–2278.

https://doi.org/10.1080/03075079.2017.1320374

Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016). The nature of feedback: how peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology*, *108*(8), 1098–1120. https://doi.org/10.1037/edu0000103

Perez, S., Massey-Allard, J., Butler, D., Ives, J., Bonn, D., Yee, N., & Roll, I. (2017). Identifying productive inquiry in virtual labs using sequence mining. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10331 LNAI*(April 2021), 287–298. https://doi.org/10.1007/978-3-319-61425-0_24

Pomery, E. A., Gibbons, F. X., & Stock, M. L. (2012). Social Comparison. *Encyclopedia of Human Behavior: Second Edition*, 463–469. https://doi.org/10.1016/B978-0-12-375000-6.00332-3

Psillos, D., Tselfes, V., & Kariotoglou, P. (2004). An epistemological analysis of the evolution of didactical activities in teaching-learning sequences: The case of fluids. *International Journal of Science Education*, *26*(5), 555–578. https://doi.org/10.1080/09500690310001614744

Reiser, B.J. (2013). What Professional Development Strategies Are Needed for Successful Implementation of the Next Generation Science Standards? *Invitational Research Symposium on Science Assessment*, (September), 1–22.

Reiser, Brian J. (2018). Scaffolding Complex Learning: The Mechanisms of Structuring and Problematizing Student Work. *Scaffolding: A Special Issue of the Journal of the Learning Sciences*, *8406*, 273–304. https://doi.org/10.4324/9780203764411-2

Rupert Jr, G. (2012). *Simultaneous Statistical Inference*. Springer Science & Buisness Media.

Sadler, P. M., & Good, E. (2010). The Impact of Self- and Peer- Grading on Student Learning The Impact of Self- and Peer-Grading on Student Learning. *Science Education*, *7197*(January 2012), 37–41. https://doi.org/10.1207/s15326977ea1101

Sampson, V., Grooms, J., & Walker, J. P. (2011). Argument-Driven Inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study. *Science Education*, *95*(2), 217–257. https://doi.org/10.1002/sce.20421

Shabankhani, B., Charati, J. Y., Shabankhani, K., & Cherati, S. K. (2020). Survey of agreement between raters for nominal data using krippendorff ' s Alpha, *10*, 160–164.

Sharpe, D. (2015). Your chi-square test is statistically significant: Now what? *Practical Assessment, Research and Evaluation*, *20*(8), 1–10.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Simons, K. D., & Klein, J. D. (2007). *The impact of scaffolding and student achievement levels in a problem-based learning environment*. *Instructional Science* (Vol. 35). https://doi.org/10.1007/s11251-006-9002-5

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement : enabling students to make decisions about the quality of work Content courtesy of Springer Nature , terms of use apply . Rights reserved . Content courtesy of Springer Nature , terms of use apply . Rights reserved ., 467–481.

Tuovinen, J. E., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology*, *91*(2), 334–341. https://doi.org/10.1037//0022-0663.91.2.334

van Popta, E., Kral, M., Camp, G., Martens, R. L., & Simons, P. R. J. (2017). Exploring the value of peer feedback in online learning for the provider. *Educational Research Review*, *20*, 24–34. https://doi.org/10.1016/j.edurev.2016.10.003

van Riesen, S., Gijlers, H., Anjewierden, A., & de Jong, T. (2018). Supporting learners' experiment design. *Educational Technology Research and Development*, *66*(2), 475–491. https://doi.org/10.1007/s11423-017-9568-4

Walker, J., Sampson, V., Grooms, J., Anderson, B., & O. Zimmerman, C. (2012). *Argument-Driven Inquiry in undergraduate chemistry labs: The impact on students' conceptual understanding, argument skills, and attitudes toward science*. *Journal of college science teaching* (Vol. 41).

Walker, Joi P., & Wolf, S. F. (2017). Getting the Argument Started: A Variation on the Density Investigation. *Journal of Chemical Education*, *94*(5), 632–635. https://doi.org/10.1021/acs.jchemed.6b00621

Walker, Joi Phelps, & Sampson, V. (2013). Learning to argue and arguing to learn: Argument-driven inquiry as a way to help undergraduate chemistry students learn how to construct arguments and engage in argumentation during a laboratory course. *Journal of Research in Science Teaching*, *50*(5), 561–596. https://doi.org/10.1002/tea.21082

Walker, Joi Phelps, Sampson, V., & Zimmerman, C. O. (2011). Argument-driven inquiry: An introduction to a new instructional model for use in undergraduate chemistry labs. *Journal of Chemical Education*, *88*(8), 1048–1056. https://doi.org/10.1021/ed100622h

Walker, Joi Phelps, Van Duzor, A. G., & Lower, M. A. (2019). Facilitating Argumentation in the Laboratory: The Challenges of Claim Change and Justification by Theory. *Journal of Chemical Education*, (96), 435–444. https://doi.org/10.1021/acs.jchemed.8b00745

Wang, W. (2014). Students' perceptions of rubric-referenced peer feedback on EFL writing: A longitudinal inquiry. *Assessing Writing*, *19*, 80–96. https://doi.org/10.1016/j.asw.2013.11.008

Wheeler, L. B., Clark, C. P., & Grisham, C. M. (2017). Transforming a Traditional Laboratory to an Inquiry-Based Course: Importance of Training TAs when Redesigning a Curriculum. *Journal of Chemical Education*, *94*(8), 1019–1026. https://doi.org/10.1021/acs.jchemed.6b00831

Williams, L. C., & Reddish, M. J. (2018). Integrating Primary Research into the Teaching Lab: Benefits and Impacts of a One-Semester CURE for Physical Chemistry. *Journal of Chemical Education*, *95*(6), 928–938. https://doi.org/10.1021/acs.jchemed.7b00855

Wooley, R. S., Was, C. A., Schunn, C. D., & Dalton, D. W. (2011). The Effects of Feedback Elaboration on the Giver of Feedback. *Bulletin of Economic Research*, *63*(2), 177–199. https://doi.org/10.1111/j.1467-8586.2009.00345.x

Wu, Y., & Schunn, C. D. (2020). When peers agree, do students listen? The central role of feedback quality and feedback frequency in determining uptake of feedback. *Contemporary Educational Psychology*, *62*(July), 101897. https://doi.org/10.1016/j.cedpsych.2020.101897

Zagallo, P., Meddleton, S., & Bolger, M. S. (2016). Teaching real data interpretation with models (TRIM): Analysis of student dialogue in a large-enrollment cell and developmental biology course. *CBE Life Sciences Education*, *15*(2), 1–18. https://doi.org/10.1187/cbe.15-11-0239