Examining the role of assignment design and peer review on student responses and revisions to an organic chemistry writing-to-learn assignment

SCHOLARONE™
Manuscripts

ARTICLE

Examining the role of assignment design and peer review on student responses and revisions to an organic chemistry writing-to-learn assignment

Field M. Watts,[a] Solaire A. Finkenstaedt-Quinn [b] and Ginger V. Shultz*[b]

Research on student learning in organic chemistry indicates that students tend to focus on surface level features of molecules with less consideration of implicit properties when engaging in mechanistic reasoning. Writing-to-learn (WTL) is one approach for supporting students' mechanistic reasoning. A variation of WTL incorporates peer review and revision to provide opportunities for students to interact with and learn from their peers, as well as revisit and reflect on their own knowledge and reasoning. However, research indicates that the rhetorical features included in WTL assignments may influence the language students use in their responses. This study utilizes machine learning to characterize the mechanistic features present in second-semester undergraduate organic chemistry students' responses to two versions of a WTL assignment with different rhetorical features. Furthermore, we examine the role of peer review on the mechanistic reasoning captured in students' revised drafts. Our analysis indicates that students include both surface level and implicit features of mechanistic reasoning in their drafts and in the feedback to their peers, with slight differences depending on the rhetorical features present in the assignment. However, students' revisions appeared to be primarily connected to the peer review process via the presence of surface features in the drafts students read (as opposed to the feedback received). These findings indicate that further scaffolding focused on how to utilize information gained from the peer review process (i.e., both feedback received and drafts read) and emphasizing implicit properties could help support the utility of WTL for developing students' mechanistic reasoning in organic chemistry.

Introduction

Mechanisms in organic chemistry are one of the more challenging topics for students to learn and instructors to teach (Dood and Watts, 2022, 2023). Research focused on supporting students with mechanistic reasoning has emphasized the value of tasks that require students to articulate their reasoning (Dood and Watts, 2022, 2023), such as constructed response tasks (Stowe and Cooper, 2017; Dood *et al.*, 2018, 2020; Yik *et al.*, 2021, 2023; Frost *et al.*, 2023) and case comparison activities (Caspari, Kranz, *et al.*, 2018; Graulich and Schween, 2018; Bode *et al.*, 2019; Caspari and Graulich, 2019; Graulich *et al.*, 2019; Watts *et al.*, 2021; Asmussen *et al.*, 2022; Kranz *et al.*, 2023). The present study focuses on the use of writing-to-learn (WTL) assignments with peer review. Our prior research on WTL demonstrates the value of WTL assignments for eliciting students' mechanistic reasoning (Watts *et al.*, 2020; Watts, Dood, *et al.*, 2022; Watts, Park, *et al.*, 2022); furthermore, studies demonstrate how the knowledge students articulate can be influenced by the peer review process (Finkenstaedt-Quinn *et al.*, 2017; Halim *et al.*, 2018; Moon, Zotos, *et al.*, 2018;

Finkenstaedt-Quinn *et al.*, 2019; Schmidt-McCormack *et al.*, 2019; Finkenstaedt-Quinn, Halim, *et al.*, 2020; Finkenstaedt-Quinn, Polakowski, *et al.*, 2021; Gupte *et al.*, 2021; Petterson *et al.*, 2022; Watts, Park, *et al.*, 2022; Finkenstaedt-Quinn *et al.*, 2024). In this study, we examine how minor differences in WTL assignment design may influence the reasoning students exhibit, and we explore the role of peer review in influencing students' mechanistic reasoning as elicited by WTL.

Mechanistic reasoning in organic chemistry
One of the central learning goals for introductory organic chemistry involves supporting students' development of mechanistic and causal reasoning in the context of organic reaction mechanisms. In general, mechanistic reasoning involves understanding the changes that occur during the course of a reaction (i.e., the bonds that are broken and made; the transformation of functional groups) alongside understanding *how* those changes occur by accounting for the underlying electron movement. Causal reasoning entails similarly understanding the changes that occur during the course of a reaction, alongside considering *why* those changes occur by accounting for the chemical properties of interacting species (i.e., acidity and basicity or nucleophilicity and electrophilicity) and the energetics of the reaction (Dood and Watts, 2022).

For the present article, we focus on the mechanistic framework originally described by Machamer et al. (2000) and

[a.] Department of Chemistry & Biochemistry, University of Wisconsin – Milwaukee, Milwaukee, WI 53211
[b.] Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109; E-mail: gshultz@umich.edu

elaborated upon by Russ et al. (2008). The framework focuses on how students explain phenomena by accounting for the underlying *entities*, the *activities* they undergo to effect change, and the *properties of entities* which guide the activities (Machamer *et al.*, 2000; Russ *et al.*, 2008). This framework has been operationalized in the context of the organic chemistry education research literature with the understanding that *entities* are the electrons, atoms, and molecules involved in a given reaction; that *activities* are the movements of electrons which involve the breaking and making of bonds; and that *properties of entities* include the chemical properties such as basicity or nucleophilicity which provide explanations for why entities interact in predictable ways (Caspari, Weinrich, *et al.*, 2018; Keiner and Graulich, 2020, 2021; Watts *et al.*, 2020). Previous research in organic chemistry settings has used this framework to characterize student responses to WTL assignments (Watts *et al.*, 2020), case comparisons (Caspari, Kranz, *et al.*, 2018), and tasks intended to support students with connecting laboratory procedures to particulate-level explanations of phenomena (Keiner and Graulich, 2020, 2021).

A review of studies on students' mechanistic reasoning in organic chemistry indicates that a prominent theme involves the dichotomy between students' focus on *surface features* and *implicit properties* when explaining why reactions occur (Dood and Watts, 2023). In this context, reasoning based on surface features has been previously defined as justifying the result of a given chemical transformation by focusing on the explicit properties of interacting entities. For example, a common finding in the literature is that students often explain why molecules interact by focusing on formal charges (which are an explicit property, since they are visible on the page when drawing a reaction mechanism; Anzovino and Lowery Bretz, 2015; Galloway *et al.*, 2017; Finkenstaedt-Quinn *et al.*, 2020; Petterson *et al.*, 2020). In other words, reasoning based on surface features tends to focus more on a descriptive explanation of what occurs during a reaction without appealing to the underlying, implicit chemical properties. In contrast, reasoning based on implicit properties utilizes the chemical properties of interacting species to guide the explanation of a given transformation; these chemical properties, such as basicity or nucleophilicity, are considered implicit and require students to access their chemistry knowledge (Strickland *et al.*, 2010; Cartrette and Mayo, 2011; Cruz-Ramírez de Arellano and Towns, 2014; Anzovino and Bretz, 2016; Wilson and Varma-Nelson, 2019; Deng and Flynn, 2021). Reviews of the literature on students' reasoning in organic chemistry demonstrate that students are capable of reasoning about reaction mechanisms using both surface features and implicit properties (Dood and Watts, 2022, 2023). Supporting students with moving from focusing on surface features towards engaging in reasoning about implicit properties is one of the central goals of many novel assessments and interventions in organic chemistry instruction, including the WTL assignments central to this study.

Writing-to-learn and peer review in chemistry

Writing is a versatile practice that can be used to support various learning goals. Within undergraduate chemistry courses, the types of writing pedagogies reported in the literature expand beyond the traditional form of laboratory reports to include writing that supports students' conceptual understanding (e.g, Rhoad, 2017; Cox *et al.*, 2018; Logan and Mountain, 2018), disciplinary thinking (e.g., Hand *et al.*, 2004, 2007; Greenbowe *et al.*, 2007; Yaman, 2021), and affective learning experiences (e.g., Finkenstaedt-Quinn, Garza, *et al.*, 2022). Many of these assignments fit into the genre of WTL—assignments where writing is used as a tool to support students' thinking about concepts and their development of reasoning skills key to the discipline. Meta-analyses of WTL research indicate that assignments which are effective at supporting the goals of WTL incorporate meaning-making tasks, interactive writing processes, clear writing expectations, and structures to support metacognition (Anderson *et al.*, 2015; Klein, 2015; Gere *et al.*, 2019). In alignment with these characteristics, assignments have been developed that provide students a context to which they apply their knowledge (e.g., Balgopal and Wallace, 2013; McDermott and Hand, 2016; Rootman-le Grange and Retief, 2018) or scaffold the writing process (e.g., Hand *et al.*, 2004, 2007; Greenbowe *et al.*, 2007; Grimberg and Hand, 2009; Cox *et al.*, 2018). In addition to WTL, the use of peer review to support writing pedagogies is gaining traction. Not only can peer review mitigate barriers to using writing, such as providing feedback in large classes (Moon, Gere, *et al.*, 2018; Finkenstaedt-Quinn, Gere, *et al.*, 2022), but it can also lead students to think more critically about a topic as they reflect on their own response and critique their peers' work (Topping, 2009). Reports within the chemistry education literature detailing the use of peer review in undergraduate courses primarily focus on Calibrated Peer Review (Russell, 2013; Cox *et al.*, 2018) or peer review used to complement WTL (Finkenstaedt-Quinn *et al.*, 2023).

The MWrite program combines both the elements of effective WTL and the benefits of peer review by working with instructors to develop WTL assignments in which students submit initial drafts in response to a contextualized prompt, undergo content-focused peer review, and submit revised drafts (Finkenstaedt-Quinn, Petterson, *et al.*, 2021). We have studied how students respond to these assignments across a series of disciplines (e.g., materials science, biology, statistics) and courses (e.g., general chemistry, organic chemistry, introductory physical chemistry); an analysis across our studies indicates that the assignments successfully engage students with the targeted content (Finkenstaedt-Quinn *et al.*, 2023). Specifically, analyses of students' drafts indicate that these assignments support students with describing challenging concepts, applying their content knowledge, and engaging in complex reasoning in various STEM courses, including introductory organic chemistry courses (e.g., Finkenstaedt-Quinn *et al.*, 2017; Finkenstaedt-Quinn, Halim, *et al.*, 2020; Watts *et al.*, 2020; Brandfonbrener *et al.*, 2021). In addition, examination of students' revisions and peer review comments

indicates that students constructively participate in the peer review and revision processes associated with these assignments (e.g., Halim *et al.*, 2018; Finkenstaedt-Quinn *et al.*, 2019; Finkenstaedt-Quinn, Halim, *et al.*, 2020; Finkenstaedt-Quinn, Polakowski, *et al.*, 2021).

The present study seeks to extend our understanding of how students engage with WTL, specifically in the context of a second-semester organic chemistry WTL assignment intended to support students' mechanistic reasoning. In our prior work focused on WTL in an organic chemistry course context, we examined how the assignments could support students' mechanistic reasoning, understanding of acid-base chemistry, and representational competence (Schmidt-McCormack *et al.*, 2019; Watts *et al.*, 2020; Watts, Dood, *et al.*, 2022; Watts, Park, *et al.*, 2022; Finkenstaedt-Quinn *et al.*, 2024). Through qualitative and quantitative analysis of students' drafts, a subset of these studies identified evidence of mechanistic reasoning within students' responses to WTL assignments in alignment with the aforementioned mechanistic reasoning framework (Russ *et al.*, 2008; Watts *et al.*, 2020; Watts, Dood, *et al.*, 2022). In another study, we identified that students' revisions to a WTL assignment were largely influenced by the drafts they read during the peer review process, and that the peer review comments received were more influential for students who demonstrated inaccurate chemical reasoning (Watts, Park, *et al.*, 2022). Additionally, we have identified that students engage with peer review to different degrees, and that specific features of the drafts students read or the comments received can be connected to specific revisions identifiable within students' writing (Finkenstaedt-Quinn *et al.*, 2024).

In addition to the importance of peer review and revision, the research on WTL also indicates the importance of the rhetorical aspects of the assignments, which includes providing a context for the writing task, an audience for students to write to, and a genre to guide how students structure their response (Gere *et al.*, 2019). Specifically, studies capturing students' experiences with the assignments in organic chemistry demonstrate how the context provided in the assignments can support students to make connections between concepts and support their affect about the assignments (e.g., Gupte *et al.*, 2021; Petterson *et al.*, 2022; Zaimi *et al.*, 2024). Furthermore, students described the audience and genre as influencing both the language they used and the amount of target content they included in their responses (e.g., Gupte *et al.*, 2021; Petterson *et al.*, 2022; Zaimi *et al.*, 2024). One of these studies found that different rhetorical contexts (e.g., a grant proposal *vs.* a news article) can pose challenges for students when the rhetorical aspects are misaligned with the learning goals (Zaimi *et al.*, 2024). Because the presence of rhetorical features requires students to balance the level of detail provided in their response with the expectations for a given audience and genre, it is necessary to investigate how the presence of rhetorical features influences students' responses and revisions throughout the WTL process.

The goal of the present study is to further investigate the nature of students' writing and revisions, as influenced by the rhetorical aspects of an assignment and the peer review

process, in the context of mechanistic reasoning. Specifically, we employ previously developed machine learning (ML) models (Watts, Dood, *et al.*, 2022) to automatically analyze the evidence of mechanistic reasoning present in students' responses to two variations of the same WTL assignment and to identify students' revisions at the sentence level. We additionally use a similar ML approach to analyze students' peer review comments at scale in order to quantitatively investigate the influence of the peer review process on students' revisions. By leveraging ML methods, this work demonstrates the value of combining automated analysis with human interpretation to facilitate insights related to learning environments (Martin and Graulich, 2023). Specifically, this study not only replicates findings from prior research by analyzing the peer review process at a larger scale, but also provides nuance and improved understanding on the way assignment design and peer review influences students' writing about organic chemistry reaction mechanisms.

Theoretical framework

This study was guided by the theory of distributed cognition (Nardi, 1996; Klein and Leacock, 2012) and the cognitive process theory of writing (Flower and Hayes, 1981; Hayes, 1996). Distributed cognition provides grounding for how students can learn from interacting with their peers during the course of responding to the WTL assignment, whereas the cognitive process theory of writing speaks to how students go through the writing process. Specifically, distributed cognition considers knowledge as contained by both internal and external representations and how an individual's knowledge can develop as they interact with others directly or with external representations of knowledge (called artifacts; Nardi, 1996; Klein and Leacock, 2012). Regarding external representations of knowledge, distributed cognition states that as a person interacts with an artifact, they can draw information from it. For example, in reading a textbook a student may learn more about the subject matter described in the text. Furthermore, external representations can support both working and long-term memory and reduce the cognitive effort required for a task. The cognitive process theory of writing (Flower and Hayes, 1981; Hayes, 1996) focuses more specifically on the individual and how the task environment affects the writing process. The task environment consists of the social elements related to the writing task and the text as it is composed. These influence the cognitive processes the individual engages in as they draw on their working and long-term memory, as well as affective elements that may impact the writing process. Together, this results in a cyclical process as the writer draws on their working and long-term memory to produce a text that they can then revise in response to the task environment and internal processes. Importantly, after the cyclical process of writing and revising, the features writers ultimately choose to include in their produced text reflect their internal representations of the knowledge relevant for the writing task.

In the context of this study, the assignment prompt (and the included chemical structures), peer review criteria, and revision

guidelines are artifacts that serve as representations of what students should be considering as they respond to the assignment. Furthermore, their peers' drafts that they read and the peer feedback they receive are artifacts that serve as external representations of their peers' knowledge or understanding on the topic covered by the assignment. As students engage in the process of writing, they create an initial response based on how they interpret the assignment prompt and their existing knowledge on the topic. Then, they engage with the artifacts resulting from the scaffolded social interactions of the peer review process. As students revise, they can reflect on their understanding of the assignment, the target content, and how to present the content. With the significance of the task environment for guiding students' responses, it is important to consider how changes to the task environment (e.g., different assignment prompts) may result in differences to both students' initial responses and how they interact with their peers' artifacts. Considered together, distributed cognition supports our thinking on how students can draw knowledge from both their peers and artifacts related to the assignment while the cognitive process theory of writing demonstrates how that knowledge can be used and incorporated into their final draft for the assignment. Furthermore, both theories indicate that students' written artifacts can serve as external representations of students' knowledge about a topic.

Research questions

To address the goals of this study, we focus on the following research questions in the context of students enrolled in a second-semester organic chemistry laboratory course:

1. What are the differences between students' initial and revised drafts for their written explanations of reaction mechanisms for two versions of the same WTL assignment (one with rhetorical components and one without)?
2. How are students' revised drafts influenced by peer review (both receiving feedback and reading peer responses) in the context of the two versions of the WTL assignment?

Methods

Course context

This study is situated within the context of a second-semester organic chemistry laboratory course at the University of Michigan. The course includes weekly, one-hour lectures, taught by faculty or postdoctoral instructors, and weekly, four-hour laboratory sessions, taught by graduate student instructors. The lectures cover topics and procedures relevant for the laboratory sessions. The lecture component of the course spans three sections of students ranging from approximately 150-350 students each, whereas the laboratory sessions span multiple sections with typically 15-17 students per laboratory. There are three WTL assignments within the course, the first of which is the focus of this study; altogether,

the WTL assignments account for thirty percent of students' final grades. The remainder of students' grades is made up of quiz performance and weekly completion of their laboratory notebooks. Students typically enroll in the course alongside the second-semester organic chemistry lecture, though students may enroll in the laboratory after completing the lecture. In addition to covering material relevant for the laboratory sessions, the lecture sequence covers the specific content that students engage with for the WTL assignments, including the racemization and hydrolysis mechanisms central to the WTL assignment described herein.

Writing-to-learn assignment and implementation

The WTL assignment central to this study is a revised version of the assignment reported in our prior research (Watts *et al.*, 2020; Watts, Dood, *et al.*, 2022). The assignment introduced students to the thalidomide molecule and provided historical background information about how thalidomide was used for treating morning sickness in pregnant women before the discovery that one of thalidomide's enantiomers causes birth defects. The assignment text described two reactions that affect thalidomide, racemization and hydrolysis, demonstrating the starting materials and products of these reactions as shown in Figure 1 and Figure 2.

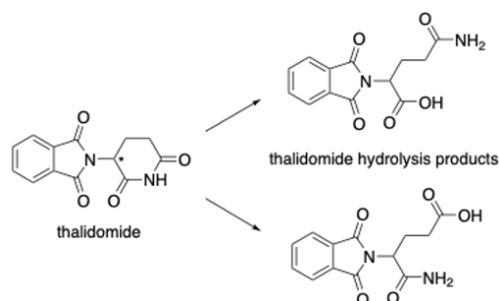Figure 1. The racemization of thalidomide.

Figure 2. Thalidomide and two hydrolysis products. The stereocenter is shown (*).

After providing context by discussing the interest in using thalidomide to treat nausea in cancer patients, the assignment discussed the desire to develop an analog of thalidomide that would prevent both racemization and hydrolysis. The writing task involved asking students to explain the mechanisms for both reactions and to propose a thalidomide analog. For this study, we focused on the component of the assignment wherein students were asked to explain the mechanisms. The text for this component of the assignment stated,

"Provide thorough descriptions of the mechanisms of both racemization and acid hydrolysis, highlighting the critical structural features of thalidomide and their role in these mechanisms.

a. When racemization occurs, what changes occur in the molecule?

b.   When hydrolysis occurs, what changes occur in the molecule?"

There were two versions of the assignment with rhetorical differences. The central writing tasks (i.e., describing the mechanisms and proposing an analog) remained the same between versions. The difference between the assignments related to the rhetorical framing of the prompt, specifically whether or not students were given a role, a specific audience, and a genre. In the traditional version, the assignment indicated that,

"*You are* an OB-GYN at the Mayo Clinic. *A colleague, who is* an oncologist at the University of Minnesota, *has approached you about* a potential collaboration on a human clinical trial… *As an organic expert in the chemical pathways that lead to birth defects, you are writing an email to your collaborator. Your goal will be* to propose a structural difference that will make the thalidomide analog unreactive toward both racemization and hydrolysis. You must provide descriptions of the structure and reactivity of thalidomide toward racemization and hydrolysis as well as descriptions of the structural differences in the proposed analog that will make it unreactive to both of these processes. *The oncologist is not an expert in organic chemistry. Therefore, carefully consider which organic chemistry terms to use and when to define or explain them. Use clear and concise language, striking a balance between organic jargon and oversimplified explanations.*" - Full (italics added)

In contrast, the other version indicated that,

"*An* OB-GYN at the Mayo Clinic *and* an oncologist at the University of Minnesota *are exploring* a potential collaboration on a human clinical trial… *Your assignment is* to propose a structural difference that will make the thalidomide analog unreactive toward both racemization and hydrolysis. You must provide descriptions of the structure and reactivity of thalidomide toward racemization and hydrolysis as well as descriptions of the structural differences in the proposed analog that will make it unreactive to both of these processes." - Pared (italics added)

The goal of implementing the two versions of the assignment was to investigate the influence of rhetorical aspects on the specificity of students' writing, specifically with respect to features related to mechanistic reasoning. Investigating these differences will provide insight regarding how the learning goals of the WTL assignment interact with the rhetorical aspects, which are in place to make the writing task meaningful (Finkenstaedt-Quinn, Petterson, *et al.*, 2021). The full text for both versions of the assignment are provided in Appendix 1. Hereafter, the two assignment versions will be referred to as the *full* and *pared* versions, respectively.

The assignment was the first of three implemented in the course. In addition to the meaning-making task in which students were expected to describe and explain the thalidomide mechanisms, the WTL implementation included additional structures to support students' learning with WTL (i.e., to provide clear expectations, include opportunities for interactive writing, and promote metacognition; Bangert-

Drowns et al., 2004; Anderson et al., 2015; Gere et al., 2019). Specifically, the assignment was provided in the learning management system alongside the evaluation rubric to clarify expectations. Students were given either the full or pared version of the assignment depending on the lecture section of the course in which they were enrolled. Students then had one week to write and submit their first drafts, after which they underwent peer review as a form of interactive writing. The peer review process was automated and double-blind, and students typically gave and received feedback to/from three of their peers within the same lecture section. Peer review entailed responding to content-focused criteria developed to elicit constructive feedback focused on the concepts within the assignment rather than the grammar or style of students' writing; the peer review criteria relevant to this study are included in Table 1, and the remaining criterion is provided in Appendix 1. Following peer review, students had three days to revise their response, which provided an opportunity for metacognition wherein students could reflect on what they incorporated into their initial response and why by considering the peer review feedback received and the responses they read during peer review.

Table 1. The peer review criteria to which students responded when providing peer review feedback. Only the criteria relevant to the study are shown.

How well does the author explain the process of racemization in thalidomide? Suggest some ways that the author could improve their mechanism description, including discussing what changes occur in the thalidomide molecule through the racemization mechanism.
How well does the author explain the process of hydrolysis in thalidomide? Suggest some ways that the author could improve their mechanism description, including discussing what changes occur in the thalidomide molecule through the hydrolysis mechanism.

Positionality statement

This study is part of a broader research effort to understand the effectiveness of WTL across STEM courses as implemented through a WTL program, called MWrite, at the University of Michigan. As such, it is important to engage in reflexivity and acknowledge our positionality with respect to this work (Lincoln and Guba, 1985). The corresponding author (GVS) was the instructor for the second-semester organic chemistry laboratory course from which we collected data. The remaining authors are a recent doctoral graduate (FMW) and the MWrite program manager (SFQ), who are both deeply familiar with the MWrite program and have conducted prior studies on WTL and peer review within the context of MWrite. This study was developed as part of a collaboration that combined the first and second authors' research interests, specifically by applying ML to characterize mechanistic reasoning (FMW) and characterizing the role of peer review in supporting students' content-focused revisions (SFQ). All authors contributed to the design of the WTL assignment for this study, and our familiarity with the MWrite program and second-semester laboratory course influenced the goals of this study and our approach for data analysis.

Participants and data collection

Participants for this study include the 632 students enrolled in the second-semester laboratory course who received a final grade and consented to participate in the study. The University's Institutional Review Board granted approval for this study with exempt status, and ethical procedures for human subjects research were followed by collecting data from students who consented to participate and anonymizing student responses. The data collected includes students' initial and final drafts to the two versions of the WTL assignment ($n =$ 300 initial and $n =$ 300 revised drafts for the full assignment; $n =$ 332 initial and $n =$ 331 revised drafts for the pared assignment), along with the associated peer review comments ($n =$ 1808 comments for the full assignment; $n =$ 1966 comments for the pared assignment).

Data analysis

An overview of the data analysis process is provided in Figure 3. In short, for RQ1 we utilized machine learning to identify features of mechanistic reasoning within students' initial and revised drafts, followed by statistical analysis to identify (1) differences between prompts and (2) differences between initial and revised drafts. For RQ2, we used machine learning to analyze students' peer review comments (informed by the findings from RQ1), followed by using linear regression to investigate the relationship between revisions and the peer review process for the different prompt versions.

Analysis of students' initial and revised drafts. In alignment with the goal for this study to examine students' initial responses and revisions at scale to uncover trends related to the slight differences between the two assignment versions, we employed ML methods to analyze student writing about the reaction mechanisms at the sentence level. As summarized in Figure 4, students' initial and revised drafts were automatically analyzed using the ML models described and deployed in our prior studies (Watts, Dood, et al., 2022; Watts, Dood, and Shultz, 2023; Watts, Dood, Shultz, et al., 2023). This involved taking the full corpus of student responses (both initial and revised drafts) and using natural language processing techniques to split each response into individual sentences, applying a ML model to determine whether each sentence included a mechanistic description, then using a set of previously described ML models to predict whether sentences included specific mechanistic reasoning features. We performed statistical analyses on the output data (which
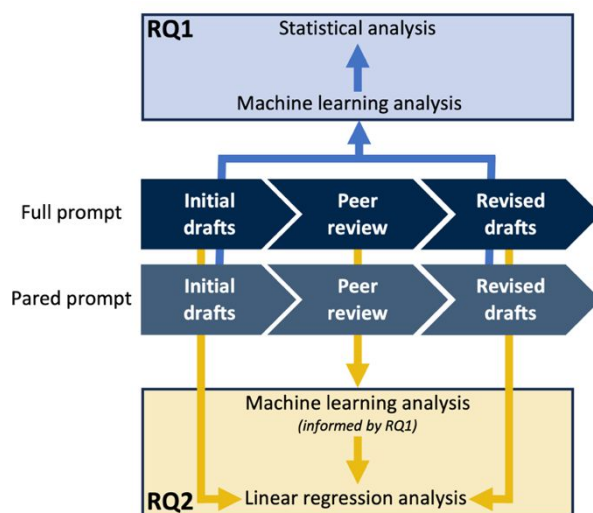

Figure 3. Overview of data analysis for both research question 1 (RQ1) and research question 2 (RQ2).

captured the number of sentences including each feature for each response) to identify trends.

The model which identified whether individual sentences in a response contained text relevant to a mechanistic explanation was trained using 3027 sentences which either included ($n =$ 1243) or did not include ($n =$ 1784) descriptions of mechanisms. The data was split into 67.5%, 22.5%, and 10% sets for training, validation, and testing respectively. The training and validation sets were used to train a convolutional neural network which performed with 88.6% accuracy and 0.762 Cohen's κ on the testing set; these human-machine agreement values were deemed acceptable due to being above the recommended value of $\kappa > 0.70$ for using ML in assessment (Williamson et al., 2012). After identifying sentences with relevant text, we used a set of previously reported ML models to identify the presence of specific mechanistic reasoning features (Watts, Dood, et al., 2022). The models capture features necessary for mechanistic reasoning, originally derived from Russ et al.'s (2008) framework for discourse analysis for capturing students' mechanistic descriptions and explanations; the alignment between the ML models and the Russ et al. framework is shown in Figure 4. The priorly reported models exhibited strong performance with accuracies and Cohen's κ between 88.4–99.7% and 0.738–0.993, respectively (Watts, Dood, et al., 2022). Using this analysis process, we automatically evaluated the

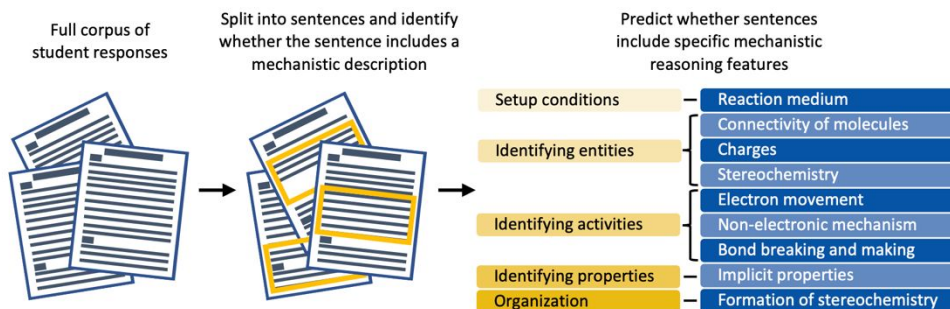Figure 4. Overview of the automated analysis process for students' writing.

presence of mechanistic reasoning features in students' responses to both versions of the assignment.

Next, we performed statistical analyses to investigate differences between responses for the two versions of the WTL assignment. We first conducted chi-square tests of independence to determine whether student responses differed between the full *versus* pared assignments; the chi-square tests of independence specifically sought to identify differences in whether students incorporated the mechanistic reasoning features *at least once* in their response (Sheskin, 2011). We then sought to compare the frequency with which students included each mechanistic reasoning feature. We conducted Shapiro-Wilk tests and determined that the distributions for the total number of sentences and for the number of sentences including each feature were non-normally distributed; as such, we used Mann-Whitney U tests to compare the frequency with which features appeared between groups. The distributions for the number of relevant sentences were normally distributed, so these were compared using t-tests (Sheskin, 2011). The analyses involved comparing between the full and pared prompt for both students' initial and revised drafts, along with comparing between students' initial and revised drafts for each version of the prompt. For all statistical analyses, we set alpha = 0.05 and corrected p-values using Bonferroni's method to account for family-wise Type 1 error rates (Sheskin, 2011). To calculate effect sizes, we used phi for the chi-square tests of independence, r for the Mann-Whitney U tests, and Cohen's d for the t-tests (Fritz *et al.*, 2012).

Analysis of peer review comments. The goal of the next stage of analysis was to identify features of peer review which may have influenced students' revised drafts. Hence, the peer review comments associated with the initial drafts included in the study were analyzed with a focus on the features of interest identified during the previous analysis of students' revised drafts. The specific features we sought to identify were whether the comments discussed (1) formal charges or (2) implicit properties. We deductively developed a coding scheme, derived from the analytical scheme used in the previous analysis stage, to analyze the peer comments for these features specifically (Table 2). For developing the coding scheme, two researchers (FMW & SFQ) independently coded 202 comments, compared the applied codes, and calculated the agreement measures shown in Table 2. The agreement measures reflected strong to almost perfect agreement (Watts and Finkenstaedt-Quinn, 2021), so one researcher then coded additional comments to reach 1010 total peer review comments analyzed. The comments analyzed were a randomly selected, stratified subset based on the prompt students responded to (with n = 488 comments from the full prompt and n = 522 comments from the pared prompt) and based on the degree of revisions associated with the authors receiving the comments (i.e., selecting comments received by students who revised their writing to include a high, medium, and low number of additional sentences incorporating the mechanistic reasoning features of interest).

ARTICLE

Table 2. Coding scheme for analyzing the peer review comments, with definitions and exemplars for each code along with human-human agreement measures, the type of machine learning model used to automate the analysis, and human-computer agreement measures.

Code	Definition	Exemplars	Human-human agreement measures		Machine learning model	Human-computer agreement measures	
			%	κ		%	κ
Charge	The peer review comment included mention of the formal charges (i.e., positive, negative, or neutral charge) of atoms or molecules.	"Good explanation of the hydrolysis reaction in thalidomide however it could be more detailed including details on what causes the molecule to split (the nitrogen being positively charged)...."	98.5	0.935	Convolutional neural network	96.5	0.803
Implicit properties	The peer review comment included mention of implicit properties (i.e., acidity/basicity, nucleophilicity/electrophilicity, electronegativity, resonance, etc.).	"...In my mechanism of racemization, I had a few more steps (I started with a protonation of the carbonyl, making the carbon of the carbonyl more electrophilic, then I did the deprotonation of the stereocenter's hydrogen, and formed the double bond), but I'm not sure whether mine is completely correct, so take that with a grain of salt..."	95.5	0.831	SciBERT model	96.0	0.829
Other	The peer review comment included anything else.	"The racemization is hinted at through the considerations of the R and S enantiomers. To improve the mechanism description, it may be beneficial to describe the steps of racemization, instead of stating that both enantiomers are produced..."	96.5	0.910	Convolutional neural network	97.0	0.910

After coding the initial 1010 peer review comments, we trained three ML models (one for each code) to automatically analyze the remaining peer review comments in the dataset. To train the ML models, we randomly split the 1010 human-analyzed peer review comments into a training and testing dataset using a 64% training, 16% validation, and 20% testing split. We evaluated the performance of several traditional ML algorithms (naive Bayes, linear regression, support vector machines) and deeper ML algorithms (convolutional neural networks and transformer models). The models with the highest human-computer agreement measures on the testing set were then used for further analysis (see Table 2). All models used for further analysis exhibited human-computer agreement measures with near-perfect agreement and exceeded the recommended value of $\kappa > 0.70$ for using ML models for automated assessment (Williamson *et al.*, 2012). With these models, we automatically analyzed the remaining peer review comments, for a total of 3774 comments analyzed (1808 from students providing feedback to the full version of the prompt; 1966 from students providing feedback to the pared version of the prompt). We then conducted chi-square tests of independence to identify whether the frequency with which students commented on different features differed between students responding to the full or pared versions of the assignment, correcting the *p*-values using Bonferroni's method (Sheskin, 2011).

Analysis of the influence of peer review on students' revised drafts. After analyzing the content of the peer review

comments for the two features of interest (charges and implicit properties), we investigated the relationship between peer review and revisions by performing two manual, stepwise sequential linear regressions for each feature of interest. For both regressions, the frequency of sentences containing the feature in students' revised drafts was the dependent variable (*charges_d2* and *implicit_d2*), and we sequentially added the following independent variables: a dummy variable for the prompt students were responding to (*prompt_dummy*; coded as 0 = Pared and 1 = Full); the frequency of the feature in students' initial drafts (*charges_d1* and *implicit_d1*, respectively); the number of peer review comments received which included the feature of interest (*charges_pr* and *implicit_pr*, respectively); the number of drafts read which included the feature of interest (*charges_dr* and *implicit_dr*, respectively). A final model incorporated all independent variables from both sets of regressions.

Results

RQ1. What are the differences between students' initial and revised drafts for their written explanations of reaction mechanisms for two versions of the same WTL assignment (one with rhetorical components and one without)?

To address research question one, we first conducted chi-square tests of independence to identify the relationship between the two different versions of the assignment and whether or not students incorporated each mechanistic reasoning feature in their response (Figure 5). We first compared the presence of mechanistic reasoning features between the two versions of the prompt for the initial and

revised drafts separately (Figure 5 left and right, respectively). We found that students who responded to the pared prompt were more likely to incorporate charges within their initial draft; however, this difference between assignment versions had a small effect size (phi = 0.149) and was not present in students' revised drafts.

Next, we used Mann-Whitney U tests to compare the frequency of sentences in which students included each mechanistic reasoning feature between the two versions of the assignment (Figure 6). From this comparison, we identified significant differences between the full and pared version of the assignment for the number of sentences in which students included charges, non-electronic mechanisms, and implicit properties for both the initial and revised drafts. For each of these features, students who responded to the pared version of the prompt included more sentences in their writing, with small effect sizes (*r* ranging from -0.135 to -0.222). Notably, there was not a significant difference between the total number of sentences or for the total number of relevant sentences for the two versions of the prompt, for both the initial and revised drafts ($p = 1.000$ and $p = 1.000$ for total number of sentences and $p = 0.133$ and $p = 0.124$ for total number of relevant sentences, for initial and revised drafts, respectively).

In addition to comparing between the full and pared versions of the assignment, we also ran a parallel analysis comparing students' initial and revised drafts for both the full version of the assignment (Figure 7, a and c) and the pared version of the assignment (Figure 7, b and d). These comparisons indicate that, for both versions of the prompt, more students included charges and electron movement within their response upon revision (Figure 7, a and b), with trivial effect sizes (see Appendix 2, Table 8). Additionally, for both
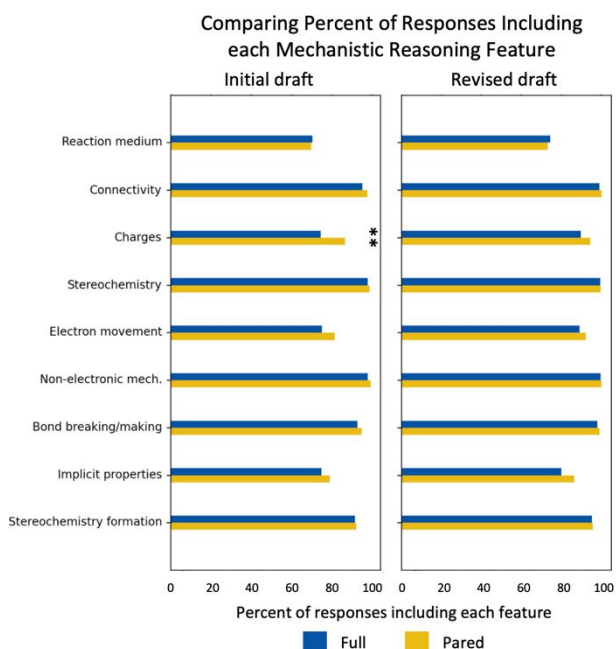
Figure 5. Percentages of students incorporating each mechanistic reasoning feature within their response, compared between the full and pared versions of the prompt for students' initial drafts (left) and revised drafts (right). Tabular results are provided in Appendix 2, Table 6. ** $p < 0.01$.

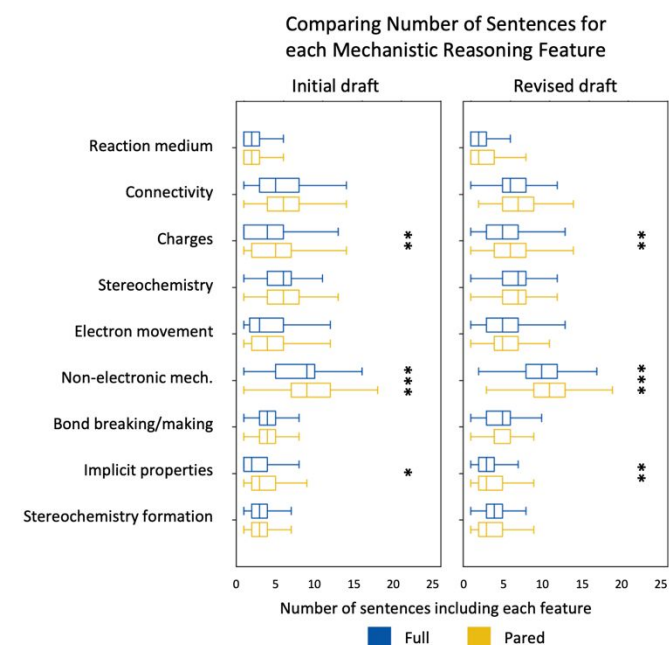Figure 6. The distribution of the number of sentences for each mechanistic reasoning feature, compared between the full and pared versions of the prompt for students' initial drafts (left) and revised drafts (right). Tabular results are provided in Appendix 2, Table 7. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

assignments, students included significantly more sentences for the same set of mechanistic reasoning features: connectivity, charges, stereochemistry, electron movement, non-electronic mechanisms, and bond breaking/making (Figure 7, c and d), with small effect sizes (see Appendix 2, Table 9).

Considering both perspectives of the data, the primary findings with respect to research question one are that (1) more students include charges at least once in their initial response to the pared version of the prompt and (2) students include more sentences describing charges, non-electronic mechanisms, and implicit properties for both their initial and revised drafts in response to the pared version of the prompt. However, when comparing between initial and revised drafts for each prompt, the trends are similar between the two prompts, in that students included significantly more sentences for several of the same mechanistic reasoning features upon revision. Although the effect sizes for significant differences were small, we would not necessarily expect large effect sizes due to the minor variation between the prompts.

RQ2. How are students' revised drafts influenced by peer review (both receiving feedback and reading peer responses) in the context of the two versions of the WTL assignment?

We chose to focus our analysis of the peer review process on two specific mechanistic reasoning features, charges and implicit properties, due to the differences observed between the full and pared versions of the assignment (as indicated in Figure 6 above). These two features represent common reasoning approaches presented in the literature: surface-level reasoning and deeper reasoning, respectively. Focusing the analysis on these features allows us to identify possible differences within the peer review process for these two types of reasoning.

To address this research question, we first analyzed students' peer review comments, and then conducted linear regressions to identify the influence of different aspects of peer review on students' revisions. First, the frequency with which each feature was commented on within the peer review process is provided in Table 3. As indicated within the table, students were significantly more likely to comment on charges when responding to the pared prompt compared to the full prompt. However, for both versions of the prompts, significantly more comments focused on implicit properties with small effect sizes
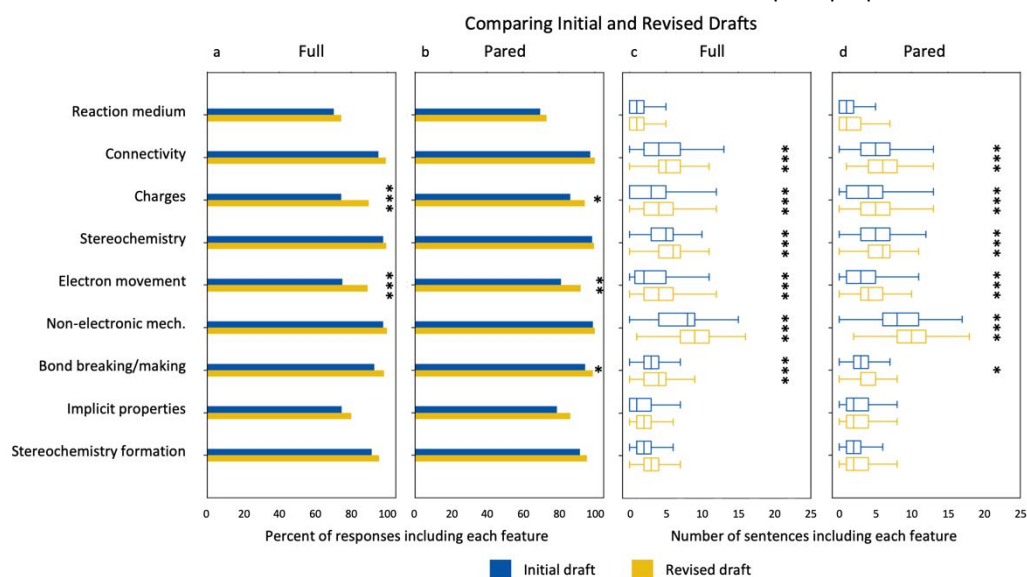
Figure 7. Comparison between students' initial and revised drafts for the full version of the assignment (a, c) and the pared version of the assignment (b, d); for each assignment version, comparisons between the percentage of students including each feature (a, (a, b) and the frequency of sentences in which each feature appeared (c, d) are shown. Tabular results are provided in Appendix 2, Table 8 and Table 9. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(Table 3).

ARTICLE

Table 3. The frequency of peer review comments pertaining to charges, implicit properties, both, or other comments. Resulting *p*-values from chi-square tests of independence are shown, corrected with Bonferroni coefficient of 7.

Feature	Frequency of comments pertaining to each feature			*p*-value (between Full and Pared)	Effect size (phi)
	Full (N = 1808)	Pared (N = 1966)	All (N = 3774)		
Charge	153 comments (8.5%)	262 comments (13.3%)	415 comments (11.0%)	< 0.001 ***	0.077
Implicit properties	229 comments (12.7%)	301 comments (15.3%)	530 comments (14.0%)	0.155	0.037
Both charge and implicit properties	43 comments (2.4%)	86 comments (4.4%)	129 comments (3.4%)	0.007 **	0.053
Other	1469 comments (81.3%)	1489 comments (75.7%)	2958 comments (78.4%)	< 0.001 ***	0.066
p-value (between Charge and Implicit Properties)	< 0.001 ***	< 0.001 ***	< 0.001 ***		
Effect size (phi)	0.138	0.189	0.171		

After characterizing the peer review comments, we performed linear regressions to investigate the relationship between peer review and students' revised drafts. The two sequential linear regressions are provided for each feature of interest (charges and implicit properties) in Table 4 and Table 5, respectively. Descriptive statistics for the variables included in the linear regressions are provided in Appendix 2, Table 10.

For the set of regressions focused on charges (Table 4), the only two significant independent variables across the five models are *charges_d1* and *charges_dr*. The first significant variable, *charges_d1*, indicates that the frequency of sentences related to charges in students' initial draft is a significant predictor of the student including additional sentences related to charges in their revised draft. The second significant variable, *charges_dr*, indicates that, in terms of the influence of the peer review process, the frequency of drafts read which included charges significantly influences students' revisions to include more sentences with charges. The version of the prompt (*prompt_dummy*), the frequency of peer review comments related to charges (*charges_pr*), and the independent variables related to implicit properties (*implicit_d1*, *implicit_pr*, *implicit_dr*) did not significantly influence the frequency of sentences pertaining to charges in students' revisions.

The set of regression models focused on implicit properties are presented in Table 5. As indicated in Table 5, the only

feature which significantly predicted the inclusion of sentences pertaining to implicit properties in students' revisions (*implicit_d2*) was the frequency with which students included sentences pertaining to implicit properties in their initial draft (*implicit_d1*). None of the other variables, including the frequency of comments or drafts read relating to implicit properties during the peer review process (*implicit_pr*, *implicit_dr*), the variables related to charges (*charges_d1*, *charges_pr*, *charges_dr*), or the version of the prompt (*prompt_dummy*) significantly predicted the frequency with which students included implicit properties in their revisions.

Considering the analyses pertaining to our second research question, we saw that (1) students gave more feedback related to implicit properties than charges for both versions of the assignment and (2) students commented more on charges in response to the pared version of the assignment compared to the full version. When examining the results of the regressions, we identified that (1) the frequency of sentences including charges in students' initial draft and the discussion of charges in the drafts they read served as predictors for students' revisions to incorporate additional sentences with charges, but (2) the frequency of sentences including implicit properties in their initial draft was the only predictor for students' revisions to incorporate additional sentences with implicit properties in their revised drafts.

1
2
3
4
5
6
7

ARTICLE

8
9
10
11
12
13
14
15
16
17

Table 4. The set of linear regressions for investigating the influence of the peer review process on students' revisions to include charges.

Dependent variable	Revised drafts – charges (*charges_d2*)	Model 1a coeff. (st. err.)	Model 1b coeff. (st. err.)	Model 1c coeff. (st. err.)	Model 1d coeff. (st. err.)	Model 1e coeff. (st. err.)
Independent variables	Prompt version (*prompt_dummy*)	-0.9586 (0.239) ***	-0.2271 (0.160)	-0.2038 (0.162)	-0.0907 (0.164)	-0.0985 (0.165)
	Initial drafts – charges (*charges_d1*)	-	0.7564 (0.027) ***	0.7518 (0.027) ***	0.7541 (0.027) ***	0.7579 (0.028) ***
	Peer review comments – charges (*charges_pr*)	-	-	0.0926 (0.095)	0.0941 (0.094)	0.0991 (0.099)
	Drafts read – charges (*charges_dr*)	-	-	-	0.3205 (0.098) **	0.3827 (0.110) **
	Initial drafts – implicit properties (*implicit_d1*)	-	-	-	-	-0.0218 (0.037)
	Peer review comments – implicit properties (*implicit_pr*)	-	-	-	-	-0.0023 (0.089)
	Drafts read – implicit properties (*implicit_dr*)	-	-	-	-	-0.1294 (0.108)
	Intercept	5.0815 (0.166) ***	1.9801 (0.155) ***	1.9229 (0.166) ***	1.2089 (0.274) ***	1.3770 (0.305) ***
	R-squared	0.026	0.576	0.576	0.584	0.585

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5. The set of linear regressions for investigating the influence of the peer review process on students' revisions to include implicit properties.

Dependent variable	Revised drafts – implicit properties (*implicit_d2*)	Model 2a coeff. (st. err.)	Model 2b coeff. (st. err.)	Model 2c coeff. (st. err.)	Model 2d coeff. (st. err.)	Model 2e coeff. (st. err.)
Independent variables	Prompt version (*prompt_dummy*)	-0.7122 (0.191) ***	-0.1544 (0.128)	-0.1591 (0.128)	-0.1535 (0.128)	-0.1458 (0.133)
	Initial drafts – implicit properties (*implicit_d1*)	-	0.7739 (0.028) ***	0.7814 (0.029) ***	0.7815 (0.029) ***	0.7806 (0.029) ***
	Peer review comments – implicit properties (*implicit_pr*)	-	-	-0.0624 (0.068)	-0.0613 (0.068)	-0.0461 (0.071)
	Drafts read – implicit properties (*implicit_dr*)	-	-	-	0.0381 (0.078)	-0.0005 (0.087)
	Initial drafts – charges (*charges_d1*)	-	-	-	-	-0.0036 (0.022)
	Peer review comments – charges (*charges_pr*)	-	-	-	-	-0.0540 (0.079)
	Drafts read – charges (*charges_dr*)	-	-	-	-	0.0900 (0.088)
	Intercept	3.0125 (0.132) ***	0.9044 (0.115) ***	0.9427 (0.122) ***	0.8643 (0.201) ***	0.7923 (0.245) **
	R-squared	0.022	0.574	0.575	0.575	0.576

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

ARTICLE

Discussion

For this study, we analyzed student responses, revisions, and peer review comments for two versions of an organic chemistry WTL assignment with minor differences in rhetorical context, which is theorized to inform the writing process (Flower and Hayes, 1981; Hayes, 1996). Because students can learn from one another through peer review and revision (Nardi, 1996; Klein and Leacock, 2012), we also sought to investigate the interplay between the rhetorical context of the assignment and peer review. Altogether, the results indicate minor differences between student responses to the WTL assignment with modest differences in prompting. From research question one, the findings suggest that students responding to the pared version of the prompt include more references to charges, non-electronic mechanistic descriptions, and implicit properties in both their initial and revised drafts. The modest rhetorical variations in the two prompts appeared to elicit small differences for specific features related to justifications for *why* mechanistic steps occur (i.e., charges and implicit properties). This finding extends previous findings in the literature related to supporting students' mechanistic reasoning, where researchers have identified that larger differences in prompting can influence the nature of students' exhibited reasoning (e.g., differences in amount of chemical information provided or differences in problem modality (DeCocq and Bhattacharyya, 2019; Finkenstaedt-Quinn, Watts, *et al.*, 2020; Petterson *et al.*, 2020; Petritis *et al.*, 2021; Zaimi *et al.*, 2024). Nevertheless, many features necessary for mechanistic reasoning showed no differences between the prompts (e.g., electron movement, bond breaking and making). These features reflect students' descriptions of *how* the mechanisms occur, indicating that both versions of the prompt support students with providing similarly descriptive accounts of the mechanisms in alignment with the learning goals. Furthermore, comparisons between initial and final drafts for both versions of the prompts reveal no apparent differences in trends for students' revisions. This finding suggests that, despite the slight differences in prompting, the WTL process similarly supports students with responding and revising based on the learning goals of the assignment.

The differences between students' responses to the two versions of the WTL assignment pertain to features of mechanistic reasoning that reflect both surface-level (i.e., charges and non-electronic mechanistic descriptions) and deeper (i.e., implicit properties) reasoning, perhaps suggesting that students responding to the pared version of the prompt included slightly more detailed mechanistic explanations in general relative to students responding to the full version of the prompt. Prior studies indicate that the audience can influence the degree of students' exhibited knowledge; for example, students in a statistics course exhibited different degrees of explanation for an assignment in which the audience was their grandparents in comparison to an assignment in which the audience was a sports team trainer (Gere *et al.*, 2018). The present study extends our understanding of the interaction between rhetorical aspects of WTL assignments and the learning goals for assignments to promote students' reasoning by presenting a direct comparison where the rhetorical situation of the assignment differed by altering only nine sentences (out of 38) between the two versions of the assignment. Specifically, the assignment versions differed only in that the full version included an explicitly stated audience and role for the students to assume, whereas the pared version removed the explicit references to the rhetorical situation. The small but significant differences in students' responses to such small variations demonstrate how minor differences in prompting may guide students to go into more or less mechanistic detail (e.g., by only writing for the implicit audience of their instructor, students responding to the pared prompt did not have to balance mechanistic detail with understandability for a target audience). This finding aligns with how the audience, as part of the task environment, is thought to influence the writing process as described by the cognitive process theory of writing (Flower and Hayes, 1981; Hayes, 1996). The differences in students' responses extends the findings from prior studies on writing pedagogies in STEM courses in which students described how balancing content expectations for different audiences (e.g., the audience given in the assignment *vs.* the instructor or grader) posed them challenges (Gere *et al.*, 2018; Gupte *et al.*, 2021; Finkenstaedt-Quinn, Garza, *et al.*, 2022; Zaimi *et al.*, 2024). Furthermore, while the observed differences between responses to the assignment versions might be small (e.g., a difference of one sentence), the cognitive process theory of writing supports the notion that such differences may reflect increased engagement with the specific ideas that represent students' understanding of the assignment content (Flower and Hayes, 1981; Hayes, 1996).

Research question two sought to further investigate the role of the peer review process on students' revisions pertaining to charges and implicit properties specifically. These two features appeared with different frequencies between students' responses for the two versions of the prompt, although the two features appeared relatively infrequently overall (representing 11-14% of comments received across both assignment versions). We found that students commented upon implicit

properties more often than charges for both versions of the assignment. This suggests that even when students emphasize charges (which reflect surface-level reasoning; e.g., Anzovino and Lowery Bretz, 2015; Galloway *et al.*, 2017) in their own responses, they can provide feedback to their peers related to implicit properties (which reflects deeper reasoning; e.g., Anzovino and Bretz, 2016; Deng and Flynn, 2021). Other studies on WTL and peer review in chemistry have demonstrated similar findings, where examination of the comments students provided their peers indicates that they can provide feedback on higher order concepts (Moon, Zotos, *et al.*, 2018; Finkenstaedt-Quinn *et al.*, 2019, 2024; Finkenstaedt-Quinn, Halim, *et al.*, 2020). Additionally, students commented on charges more often for the pared version than the full version, suggesting a similar trend related to the role of the audience as identified by examining students' responses for research question one.

To investigate which aspects of the entire WTL review process influenced students' revisions, we performed regression analyses with the frequency of sentences pertaining to each feature (charges or implicit properties) in students' revisions as the dependent variable and the prompt, frequency of sentences in students' initial drafts, frequency of peer review comments received, and frequency of drafts read pertaining to each feature as independent variables. The findings from the regressions indicated no apparent differences between the two versions of the prompt for students' engagement in peer review. Additionally, for both charges and implicit properties, the most important independent variable for predicting the frequency of sentences related to each feature within students' revisions was whether the feature was included in students' initial drafts. However, the two regressions indicated nuanced differences between the two features with respect to the influence of the peer review process on revisions. Specifically, for charges, students' revisions were also significantly influenced by the number of drafts read which included charges; this finding corroborates prior research examining students' responses to WTL assignments which indicates that reading drafts and forming feedback often plays a more significant role in students' revisions compared to receiving peer review comments (Finkenstaedt-Quinn, Polakowski, *et al.*, 2021; Watts, Park, *et al.*, 2022). However, the trend was not evident for the implicit properties feature. The finding that the influence of drafts read on revisions may be different for implicit properties furthers our understanding of the WTL process; particularly, it appears that the benefit of reading might be evident for more accessible content (such as charges, which are a surface feature of reaction mechanisms) compared to more challenging content (such as implicit properties, which require deeper reasoning). Additionally, as there was a lower frequency of implicit properties in students' initial drafts, compared to charges, students may have also gained less exposure to how to incorporate implicit properties into their responses or the importance of implicit properties for mechanistic reasoning. Students' engagement with aspects of the peer review process may also influence the type and extent of their revisions. In a study examining students' responses to a

similar assignment, Finkenstaedt-Quinn et al. (2024) found that students viewed reading their peers drafts to be more helpful than receiving peer feedback. This distinction may be exacerbated when students consider features related to surface-level versus deeper reasoning. Altogether, viewed through the lens of distributed cognition (Nardi, 1996; Klein and Leacock, 2012), the findings from this study indicate the way students' knowledge (as represented by their writing) can develop through the process of reading and providing feedback to their peers' drafts.

Limitations

The present study is limited in the claims we can make based on the data collected and our approach to data analysis. Our analysis identified trends in students' responses, revisions, and interactions during peer review for two variations of one WTL assignment within a single undergraduate organic chemistry course at a single institution; as such, findings may not be directly generalizable to other populations of students or to other types of writing assignments. The WTL assignment analyzed was the first implemented in the course, so the results may be influenced by the fact that students were in the process of gaining familiarity with writing about organic reaction mechanisms and engaging with peer review and revision. While the automated analysis of students' initial responses and revisions allowed for understanding student revisions at scale, the analysis does not allow for us to make claims about the quality of students' initial drafts, peer review comments, or revisions. Furthermore, our analysis of the influence of the peer review process allowed us to uncover broad trends across the large dataset of student writing; however, we cannot definitively pinpoint all factors which may have influenced students' revision processes. Additionally, with respect to the application of machine learning methods, we recognize that the quantification and automated analysis of qualitative data (student writing and peer review comments) allows for biases that may have influenced the qualitative data analysis to affect the quantitative and statistical analyses. With this in mind, we emphasize that the findings presented herein may not reflect individual students' reasoning and experiences; rather, the findings suggest overall trends in students' responses, peer review, and revisions at the classroom level. Lastly, the project sought to understand potential differences in students' writing and revisions based on the presence of rhetorical factors within the assignment prompt; while we identified slight differences in students writing, the data we collected does not afford insight regarding potential effects of the different prompts on students' perceptions and affective experiences of the WTL process.

This work is also limited by our inability to account for *chaining*, a key aspect of Russ et al.'s framework for mechanistic reasoning (Russ *et al.*, 2008). *Chaining* entails reasoning about one step of a mechanism based on what has happened previously (backward chaining) or what will happen next (forward chaining). As discussed in our previous work detailing the development of the analytical framework used in this study

(Watts *et al.*, 2020), chaining does not appear distinctly in students' WTL responses due to the nature of providing a written description when given the opportunity to refer to outside sources and engage in the peer review process. While this precludes us from being able to make claims regarding how students construct a full mechanistic account, the present analysis does allow for us to explore variations in how students incorporate different features of organic reaction mechanisms (such as charges and implicit properties) which reflect how students engage in mechanistic reasoning.

Conclusion

In this study, we analyzed students' responses, revisions, and peer review comments in response to two versions of the same WTL assignment focused on eliciting students' mechanistic reasoning. In general, students' responses to the pared version of the prompt included more sentences for several features of mechanistic reasoning. For both versions of the prompt, students generally revised their writing to include more of the mechanistic reasoning features targeted by the assignment. Within the peer review comments, students frequently included comments related to implicit properties, though students' revisions tended to include more discussion of charges rather than implicit properties. Students' revisions were often more influenced by drafts read during the peer review process; however, this trend was only apparent for charges (a surface feature) and not for implicit properties.

Implications for instruction

The findings indicate that slight differences in prompting can influence the responses elicited by students, even for aspects of the assignment description not directly related to the elicited content knowledge. Hence, instructors should carefully consider the prompts for their assessments and ensure alignment with learning goals (e.g., such as incorporating rhetorical aspects which reinforce, rather than detract from, the learning goals). The results also provide further evidence suggesting the nuanced benefits of peer review and revision. Providing students the opportunity to engage with each other (through processes like peer review) and revisit their responses to assessments (through opportunities such as revision) can allow students additional opportunities to demonstrate and refine the knowledge intended to be elicited through assessments. In the context of WTL specifically, the present study provides further evidence that students' revisions are often connected to reading other student drafts and providing feedback rather than from receiving peer review comments. However, the findings suggest that this trend is not always the case for more challenging aspects of reasoning, such as considering implicit properties. Hence, instructors should consider approaches for eliciting and supporting students' engagement with more challenging reasoning. For example, instructors could emphasize or provide examples which demonstrate the more challenging aspects of reasoning within assignment materials. Instructors could also engage students in

activities such as written reflections about their revisions and the reasons for revising (including how their thinking may have been influenced by the drafts read or peer review comments). Such revision reflections may promote students' metacognition about whether and how they can use various aspects of the peer review process to improve their drafts.

Implications for research

This study uses both qualitative and quantitative methodologies to replicate and provide further nuance to findings related to how students engage with WTL pedagogy. Utilizing diverse methods (such as qualitative artifact analysis and ML) to investigate phenomena of interest can provide various perspectives; doing so can contribute to the reproducibility of research findings while also affording new insights as seen here. Additionally, further research can continue moving beyond the application of ML to classify student work and towards leveraging ML as a research tool. Beyond the possibilities for broad research directions building upon the methodology of the present study, the findings from this work also suggest directions for future research on WTL or other writing pedagogies and on eliciting students' mechanistic reasoning. As for writing pedagogies, the findings from this study suggest that students' writing is influenced by the presence of rhetorical prompt features. Within WTL pedagogy, these rhetorical aspects are intended to support students' positive affective engagement with the assignment; however, we note instead that they have the potential to shift the content students focus on. As such, there is a need for further research investigating WTL assignment design and the interplay between the affective and cognitive goals of the assignments. Lastly, this study identified nuanced differences within the peer review process depending on specific features of students' mechanistic reasoning (i.e., charges *vs* implicit properties); while this result corroborates the existing research, the finding suggests the need for future research exploring the nuances of pedagogical approaches for eliciting and supporting students' reasoning in organic chemistry.

Author Contributions

FMW - conceptualization, methodology, validation, formal analysis, investigation, writing - original draft, writing - review & editing, visualization; SAFQ - conceptualization, methodology, validation, formal analysis, investigation, writing - original draft, writing - review & editing, funding acquisition; GVS - investigation, resources, writing - review & editing, funding acquisition

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Appendices

Appendix 1. Full text of the two versions of the WTL assignment and the peer review guidelines

Developing a Therapeutic Analog for Thalidomide (Version A)

Thalidomide was widely used after World War II as a sedative and later as a treatment for morning sickness. Unfortunately, after its widespread use, it was discovered that thalidomide causes very serious side effects—in particular, birth defects such as phocomelia (limb malformation). The drug was banned in 1962, and these events resulted in important changes to the way the FDA approves drugs. Now, despite the inherent dangers, thalidomide is used for treatment of nausea related to chemotherapy, where benefit of treatment outweighs the inherent dangers.

It is understood that thalidomide exists as two enantiomers; one is a teratogen that causes birth defects, while the other has therapeutic properties. Rapid racemization occurs at neutral pH, so both enantiomers are formed at roughly an equal mixture in the blood, which means that, even if only the therapeutic isomer is used, both will form once introduced in the body. The racemization is illustrated below in Figure 1.

Furthermore, both enantiomers are subject to acid hydrolysis once in the stomach at lower pH, which could produce products that are teratogens. The structure of thalidomide and two thalidomide hydrolysis products are shown below in Figure 2. For these reasons, it is important to prevent both the racemization and the subsequent hydrolysis of thalidomide.

You are an OB-GYN at the Mayo Clinic. A colleague, who is an oncologist at the University of Minnesota, has approached you about a potential collaboration on a human clinical trial. This trial will propose and test the efficacy of thalidomide analogs for the treatment of nausea in cancer patients. (See note on the third page for an explanation of an analog.)

As an organic expert in the chemical pathways that lead to birth defects, you are writing an email to your collaborator. Your goal will be to propose a structural difference that will make the thalidomide analog unreactive toward both racemization and hydrolysis. You must provide descriptions of the structure and reactivity of thalidomide toward racemization and hydrolysis as well as descriptions of the structural differences in the proposed analog that will make it unreactive to both of these processes. The oncologist is not an expert in organic chemistry. Therefore, carefully consider which organic chemistry terms to use and when to define or explain them. Use clear and concise

language, striking a balance between organic jargon and oversimplified explanations.

Your email should be approximately between 500-700 words (1-2 pages) in length. It should address the following points:

1. Provide thorough descriptions of the mechanisms of both racemization and acid hydrolysis, highlighting the critical structural features of thalidomide and their role in these mechanisms.
    a. When racemization occurs, what changes occur in the molecule?
    b. When hydrolysis occurs, what changes occur in the molecule?
2. Propose a thalidomide analog (one compound) that would not undergo racemization or hydrolysis. Explain what structural features are in place that would inhibit or prevent these processes.

You can and should include figures of schemes, structures, or mechanisms, if that supports your response. We suggest that you have the figure(s) in front of you—ready to color-code or mark-up in various ways—and that you use your visible thinking to guide your audience through your explanation. Any images that you include in your response, including the figures in this prompt or those that you draw in ChemDraw or on paper, must have the original source cited using either ACS or APA format. Given your audience, your written response should suffice so that the explanations can be understood without the figures. You will be graded only on your written response.

An analog is a compound that is very similar to but has small structural differences from the pharmaceutical target. For example, m-cresol (shown in Figure 8 below) is an analog of phenol.
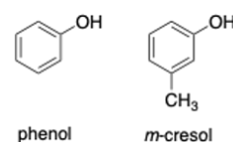

phenol   m-cresol

Figure 8. Phenol and m-cresol, an analog of phenol.

Developing a Therapeutic Analog for Thalidomide (Version B)

Thalidomide was widely used after World War II as a sedative and later as a treatment for morning sickness. Unfortunately, after its widespread use, it was discovered that thalidomide causes very serious side effects—in particular, birth defects such as phocomelia (limb malformation). The drug was banned in 1962, and these events resulted in important changes to the way the FDA approves drugs. Now, despite the inherent dangers, thalidomide is used for treatment of nausea related to chemotherapy, where benefit of treatment outweighs the inherent dangers.

It is understood that thalidomide exists as two enantiomers; one is a teratogen that causes birth defects, while the other has therapeutic properties. Rapid racemization occurs at neutral pH, so both enantiomers are formed at roughly an equal mixture in the blood, which means that, even if only the therapeutic isomer is used, both will form once introduced in the body. The racemization is illustrated below in Figure 1.

Furthermore, both enantiomers are subject to acid hydrolysis once in the stomach at lower pH, which could produce products that are teratogens. The structure of thalidomide and two thalidomide hydrolysis products are shown below in Figure 2. For these reasons, it is important to prevent both the racemization and the subsequent hydrolysis of thalidomide.

An OB-GYN at the Mayo Clinic and an oncologist at the University of Minnesota are exploring a potential collaboration on a human clinical trial. This trial will propose and test the efficacy of thalidomide analogs for the treatment of nausea in cancer patients. (See note on the third page for an explanation of an analog.)

Your assignment is to propose a structural difference that will make the thalidomide analog unreactive toward both racemization and hydrolysis. You must provide descriptions of the structure and reactivity of thalidomide toward racemization and hydrolysis as well as descriptions of the structural differences in the proposed analog that will make it unreactive to both of these processes.

Your response should be approximately between 500-700 words (1-2 pages) in length. It should address the following points:

1. Provide thorough descriptions of the mechanisms of both racemization and acid hydrolysis, highlighting the critical structural features of thalidomide and their role in these mechanisms.
    a. When racemization occurs, what changes occur in the molecule?
    b. When hydrolysis occurs, what changes occur in the molecule?
2. Propose a thalidomide analog (one compound) that would not undergo racemization or hydrolysis. Explain what structural features are in place that would inhibit or prevent these processes.

You can and should include figures of schemes, structures, or mechanisms, if that supports your response. We suggest that you have the figure(s) in front of you—ready to color-code or mark-up in various ways—and that you use your visible thinking to guide your explanation. Any images that you include in your response, including the figures in this prompt or those that you draw in ChemDraw or on paper, must have the original source cited using either ACS or APA format. Your written response should suffice so that the explanations can be understood without the figures. You will be graded only on your written response.

An analog is a compound that is very similar to but has small structural differences from the pharmaceutical target. For example, m-cresol (shown in Figure 8 below) is an analog of phenol.

Peer review guidelines

Print and read over your peer's essay to quickly get an overview of the piece.

- Read the essay more slowly keeping the rubric in mind.
- Highlight the pieces of texts that let you directly address the rubric prompts in your online responses.
- In your online responses, focus on larger issues (higher order concerns) of content and argument rather than lower order concerns like grammar and spelling.
- Be very specific in your responses, referring to your peer's actual language, mentioning terms and concepts that are either present or missing, and following the directions in the rubric.
- Use respectful language whether you are suggesting improvements to or praising your peer.

How well does the author explain the process of racemization in thalidomide? Suggest some ways that the author could improve their mechanism description, including discussing what changes occur in the thalidomide molecule through the racemization mechanism.

How well does the author explain the process of hydrolysis in thalidomide? Suggest some ways that the author could improve their mechanism description, including discussing what changes occur in the thalidomide molecule through the hydrolysis mechanism.

Does the author propose a reasonable thalidomide analog that would not undergo racemization or hydrolysis? To what extent does the author explain the specific structural features that are present in the thalidomide analog that would stop racemization and/or hydrolysis from occurring?

Appendix 2. Statistical analyses

ARTICLE

Table 6. Chi-squared tests of independence comparing the full and pared assignments. All *p*-values corrected with Bonferroni's method using coefficient 18.

Chi-squared tests of independence – D1 Full *versus* D1 Pared

Mechanistic reasoning feature	Frequency of responses including each feature		*p*-value	Effect size (phi)
	D1 Full (N = 300)	D1 Pared (N = 332)		
Reaction medium	211 (70.3%)	231 (69.6%)	1.000	0.005
Connectivity	285 (95.0%)	324 (97.6%)	1.000	0.061
Charges	223 (74.3%)	287 (86.4%)	0.003 **	0.149
Stereochemistry	293 (97.7%)	327 (98.5%)	1.000	0.019
Electron movement	225 (75.0%)	270 (81.3%)	1.000	0.073
Non-electronic mechanism	293 (97.7%)	329 (99.1%)	1.000	0.045
Bond breaking/making	278 (92.7%)	314 (94.6%)	1.000	0.033
Implicit properties	224 (74.7%)	262 (78.9%)	1.000	0.047
Stereochemistry formation	274 (91.3%)	305 (91.9%)	1.000	0.004

Chi-squared tests of independence – D2 Full *versus* D2 Pared

Mechanistic reasoning feature	Frequency of responses including each feature		*p*-value	Effect size (phi)
	D2 Full (N = 300)	D2 Pared (N = 331)		
Reaction medium	223 (74.3%)	242 (73.1%)	1.000	0.010
Connectivity	297 (99.0%)	331 (100%)	1.000	0.050
Charges	269 (89.7%)	313 (94.6%)	0.574	0.085
Stereochemistry	298 (99.3%)	330 (99.7%)	1.000	0.003
Electron movement	267 (89.0%)	305 (92.1%)	1.000	0.048
Non-electronic mechanism	299 (99.7%)	331 (100%)	1.000	0.002
Bond breaking/making	294 (98.0%)	328 (99.1%)	1.000	0.033
Implicit properties	240 (80.0%)	286 (86.4%)	0.726	0.082
Stereochemistry formation	286 (95.3%)	317 (98.5%)	1.000	0.003

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7. Mann-Whitney *U* tests comparing the full and pared assignments. All *p*-values corrected with Bonferroni's method using coefficient 22.

Mann-Whitney *U* tests – D1 Full *versus* D1 Pared				
Mechanistic reasoning feature	**Median number of sentences (mean, standard deviation)**		*p*-value	Effect size (*r*)
	D1 Full (N = 300)	**D1 Pared (N = 332)**		
Reaction medium	1 (1.54, 1.45)	1 (1.55, 1.51)	1.000	0.009
Connectivity	4 (4.66, 3.02)	5 (5.16, 2.87)	0.562	-0.088
Charges	3 (3.15, 2.86)	4 (4.05. 3.00)	0.002 **	-0.157
Stereochemistry	5 (4.91, 2.46)	5 (5.24, 2.39)	1.000	-0.070
Electron movement	2 (3.10, 3.05)	3 (3.51, 2.90)	0.338	-0.096
Non-electronic mechanism	8 (7.09, 3.60)	8 (8.39, 3.45)	< 0.001 ***	-0.175
Bond breaking/making	3 (3.05, 1.84)	3 (3.28, 2.26)	1.000	-0.026
Implicit properties	1 (2.00, 2.03)	2 (2.70, 2.48)	0.013 *	-0.135
Stereochemistry formation	2 (2.42, 1.48)	2 (2.33, 1.51)	1.000	0.045
Number of sentences	35 (36.09, 9.54)	34 (35.13, 8.89)	1.000	0.052
Number of relevant sentences	16 (16.49, 5.65)	18 (17.68, 5.26)	0.133 [a]	0.220 [b]
Mann-Whitney *U* tests – D2 Full *versus* D2 Pared				
Mechanistic reasoning feature	**Median number of sentences (mean, standard deviation)**		*p*-value	Effect size (*r*)
	D2 Full (N = 300)	**D2 Pared (N = 331)**		
Reaction medium	1 (1.63, 1.45)	1 (1.67, 1.52)	1.000	-0.004
Connectivity	5 (5.58, 2.63)	6 (6.16, 2.66)	0.228	-0.101
Charges	4 (4.14, 2.92)	5 (5.08, 3.00)	0.001 **	-0.158

ARTICLE

Stereochemistry	6 (5.89, 2.45)	6 (6.15, 2.48)	1.000	-0.045
Electron movement	4 (4.06, 2.93)	4 (4.45, 2.77)	0.834	-0.082
Non-electronic mechanism	9 (8.74, 2.98)	10 (10.20, 2.94)	< 0.001 ***	-0.222
Bond breaking/making	4 (3.66, 1.78)	3 (3.70, 1.92)	1.000	0.012
Implicit properties	2 (2.27, 2.09)	2 (3.05, 2.55)	0.003 **	-0.151
Stereochemistry formation	3 (2.81, 1.61)	2 (2.63, 1.50)	1.000	0.053
Number of sentences	41 (41.69, 9.47)	41 (40.85, 9.31)	1.000	0.039
Number of relevant sentences	19 (19.24, 5.18)	20 (20.34, 4.68)	0.124 [a]	0.222

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[a] p-values from t-tests with the Bonferroni correction for multiple hypothesis tests

[b] Effect sizes for t-tests are Cohen's d

Table 8. Chi-squared tests of independence comparing initial to final drafts for both assignment versions. All *p*-values corrected with Bonferroni's method using coefficient 18.

Chi-squared tests of independence – D1 Full *versus* D2 Full

Mechanistic reasoning feature	Frequency of responses including each feature		*p*-value	Effect size (phi)
	D1 Full (N = 300)	D2 Full (N = 300)		
Reaction medium	211 (70.3%)	223 (74.3%)	1.000	0.041
Connectivity	285 (95.0%)	297 (99.0%)	0.153	0.107
Charges	223 (74.3%)	269 (89.7%)	< 0.001 ***	0.195
Stereochemistry	293 (97.7%)	298 (99.3%)	1.000	0.055
Electron movement	225 (75.0%)	267 (89.0%)	< 0.001 ***	0.178
Non-electronic mechanism	293 (97.7%)	299 (99.7%)	1.000	0.073
Bond breaking/making	278 (92.7%)	294 (98.0%)	0.066	0.119
Implicit properties	224 (74.7%)	240 (80.0%)	1.000	0.060
Stereochemistry formation	274 (91.3%)	286 (95.3%)	1.000	0.073

Chi-squared tests of independence – D1 Pared *versus* D2 Pared

Mechanistic reasoning feature	Frequency of responses including each feature		*p*-value	Effect size (phi)
	D1 Pared (N = 332)	D2 Pared (N = 331)		
Reaction medium	231 (69.6%)	242 (73.1%)	1.000	0.036
Connectivity	324 (97.6%)	331 (100%)	0.233	0.097
Charges	287 (86.4%)	313 (94.6%)	0.011 *	0.133
Stereochemistry	327 (98.5%)	330 (99.7%)	1.000	0.048
Electron movement	270 (81.3%)	305 (92.1%)	0.001 **	0.155
Non-electronic mechanism	329 (99.1%)	331 (100%)	1.000	0.045
Bond breaking/making	314 (94.6%)	328 (99.1%)	0.035 *	0.120
Implicit properties	262 (78.9%)	286 (86.4	0.262	0.095
Stereochemistry formation	305 (91.9%)	317 (95.8%)	0.976	0.075

* *p* < 0.05, ** *p* < 0.01, *** *p* < 0.001

ARTICLE

Table 9. Mann-Whitney *U* tests comparing initial to final drafts for both assignment versions. All *p*-values corrected with Bonferroni's method using coefficient 22.

Mann-Whitney *U* tests – D1 Full *versus* D2 Full				
Mechanistic reasoning feature	Median number of sentences (mean, standard deviation)		*p*-value	Effect size (*r*)
	D1 Full (N = 300)	D2 Full (N = 300)		
Reaction medium	1 (1.54, 1.45)	1 (1.63, 1.45)	1.000	-0.034
Connectivity	4 (4.66, 3.02)	5 (5.58, 2.63)	< 0.001 ***	-0.167
Charges	3 (3.15, 2.86)	4 (4.14, 2.92)	< 0.001 ***	-0.176
Stereochemistry	5 (4.91, 2.46)	6 (5.89, 2.45)	< 0.001 ***	-0.197
Electron movement	2 (3.10, 3.05)	4 (4.06, 2.93)	< 0.001 ***	-0.192
Non-electronic mechanism	8 (7.09, 3.60)	9 (8.74, 2.98)	< 0.001 ***	-0.233
Bond breaking/making	3 (3.05, 1.84)	4 (3.66, 1.78)	< 0.001 ***	-0.168
Implicit properties	1 (2.00, 2.03)	2 (2.27, 2.09)	1.000	-0.070
Stereochemistry formation	2 (2.42, 1.48)	3 (2.81, 1.61)	0.100	-0.114
Number of sentences	35 (36.09, 9.54)	41 (41.69, 9.47)	< 0.001 ***	-0.293
Number of relevant sentences	16 (16.49, 5.65)	19 (19.24, 5.18)	< 0.001 *** [a]	0.508
Mann-Whitney *U* tests – D1 Pared *versus* D2 Pared				
Mechanistic reasoning feature	Median number of sentences (mean, standard deviation)		*p*-value	Effect size (*r*)
	D1 Pared (N = 332)	D2 Pared (N = 331)		
Reaction medium	1 (1.55, 1.51)	1 (1.67, 1.52)	1.000	-0.046
Connectivity	5 (5.16, 2.87)	6 (6.16, 2.66)	< 0.001 ***	-0.177
Charges	4 (4.05, 3.00)	5 (5.08, 3.00)	< 0.001 ***	-0.171

Stereochemistry	5 (5.24, 2.39)	6 (6.15, 2.48)	< 0.001 ***	-0.182
Electron movement	3 (3.51, 2.90)	4 (4.45, 2.77)	< 0.001 ***	-0.182
Non-electronic mechanism	8 (8.39, 3.45)	10 (10.20, 2.94)	< 0.001 *** [a]	0.564 [b]
Bond breaking/making	3 (3.28, 2.26)	3 (3.70, 1.92)	0.012 *	-0.132
Implicit properties	2 (2.70, 2.48)	2 (3.05, 2.55)	1.000	-0.075
Stereochemistry formation	2 (2.33, 1.51)	2 (2.63, 1.50)	0.132	-0.104
Number of sentences	34 (35.13, 8.89)	41 (40.85, 9.31)	< 0.001 ***	-0.307
Number of relevant sentences	18 (17.68, 5.26)	20 (20.34, 4.68)	< 0.001 *** [a]	0.532 [b]

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[a] p-values from t-tests with the Bonferroni correction for multiple hypothesis tests

[b] Effect sizes for t-tests are Cohen's d

Table 10. Descriptive statistics for variables included in the two sets of linear regressions.

Variable	Mean	St. dev.	Min.	Max.
Revised drafts – charges (*charges_d2*)	4.623	2.995	0	13
Revised drafts – implicit properties (*implicit_d2*)	2.672	2.381	0	13
Prompt version (prompt_dummy)	0.479	0.500	0	1
Initial drafts – charges (charges_d1)	3.637	2.977	0	13
Peer review comments – charges (charges_pr)	0.678	0.858	0	4
Drafts read – charges (charges_dr)	2.029	0.819	0	3
Initial drafts – implicit properties (implicit_d1)	2.379	2.314	0	14
Peer review comments – implicit properties (implicit_pr)	0.866	0.967	0	5
Drafts read – implicit properties (implicit_dr)	1.961	0.816	0	3

References

Anderson P., Anson C. M., Gonyea R. M., and Paine C., (2015), The Contributions of Writing to Learning and Development: Results from a

Large-Scale Multi-institutional Study. *Research in the Teaching of English*, **50**(2), 199–235.

Anzovino M. E. and Bretz S. L., (2016), Organic chemistry students' fragmented ideas about the structure and function of nucleophiles and electrophiles: a concept map analysis. *Chem. Educ. Res. Pract.*, **17**(4), 1019–1029, DOI: 10.1039/C6RP00111D.

Anzovino M. E. and Lowery Bretz S., (2015), Organic chemistry students' ideas about nucleophiles and electrophiles: the role of charges and mechanisms. *Chem. Educ. Res. Pract.*, **16**(4), 797–810, DOI: 10.1039/C5RP00113G.

Asmussen G., Rodemer M., Eckhard J., and Bernholt S., (2022), From Free Association to Goal-directed Problem-solving—Network Analysis of Students' Use of Chemical Concepts in Mechanistic Reasoning, in *Student Reasoning in Organic Chemistry*, Graulich N. and Shultz G. (eds.)., The Royal Society of Chemistry, pp. 90–109, DOI: 10.1039/9781839167782-00090.

Balgopal M. and Wallace A., (2013), Writing-to-Learn, Writing-to-Communicate, & Scientific Literacy. *The American Biology Teacher*, **75**(3), 170–175, DOI: 10.1525/abt.2013.75.3.5.

Bangert-Drowns R. L., Hurley M. M., and Wilkinson B., (2004), The Effects of School-Based Writing-to-Learn Interventions on Academic Achievement: A Meta-Analysis. *Review of Educational Research*, **74**(1), 29–58, DOI: 10.3102/00346543074001029.

Bode N. E., Deng J. M., and Flynn A. B., (2019), Getting Past the Rules and to the WHY: Causal Mechanistic Arguments When Judging the Plausibility of Organic Reaction Mechanisms. *Journal of Chemical Education*, **96**, 1068–1082, DOI: 10.1021/acs.jchemed.8b00719.

Brandfonbrener P. B., Watts F. M., and Shultz G. V., (2021), Organic Chemistry Students' Written Descriptions and Explanations of Resonance and Its Influence on Reactivity. *J. Chem. Educ.*, **98**(11), 3431–3441, DOI: 10.1021/acs.jchemed.1c00660.

Cartrette D. P. and Mayo P. M., (2011), Students' understanding of acids/bases in organic chemistry contexts. *Chem. Educ. Res. Pract.*, **12**(1), 29–39, DOI: 10.1039/C1RP90005F.

Caspari I. and Graulich N., (2019), Scaffolding the structure of organic chemistry students' multivariate comparative mechanistic reasoning. *International Journal of Physics and Chemistry Education*, **11**(2), 31–43, DOI: https://doi.org/10.12973/ijpce/211359.

Caspari I., Kranz D., and Graulich N., (2018), Resolving the complexity of organic chemistry students' reasoning through the lens of a mechanistic framework. *Chem. Educ. Res. Pract.*, **19**(4), 1117–1141, DOI: 10.1039/C8RP00131F.

Caspari I., Weinrich M. L., Sevian H., and Graulich N., (2018), This mechanistic step is "*productive*": organic chemistry students' backward-oriented reasoning. *Chem. Educ. Res. Pract.*, **19**(1), 42–59, DOI: 10.1039/C7RP00124J.

Cox C. T., Poehlmann J. S., Ortega C., and Lopez J. C., (2018), Using Writing Assignments as an Intervention to Strengthen Acid–Base Skills. *J. Chem. Educ.*, **95**(8), 1276–1283, DOI: 10.1021/acs.jchemed.8b00018.

Cruz-Ramírez de Arellano D. and Towns M. H., (2014), Students' understanding of alkyl halide reactions in undergraduate organic chemistry. *Chem. Educ. Res. Pract.*, **15**(4), 501–515, DOI: 10.1039/C3RP00089C.

DeCocq V. and Bhattacharyya G., (2019), TMI (Too much information)! Effects of given information on organic chemistry students' approaches to solving mechanism tasks. *Chem. Educ. Res. Pract.*, **20**(1), 213–228, DOI: 10.1039/C8RP00214B.

Deng J. M. and Flynn A. B., (2021), Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items. *Chem. Educ. Res. Pract.*, **22**(3), 749–771, DOI: 10.1039/D0RP00320D.

Dood A. J., Dood J. C., Cruz-Ramírez de Arellano D., Fields K. B., and Raker J. R., (2020), Analyzing explanations of substitution reactions using lexical analysis and logistic regression techniques. *Chem. Educ. Res. Pract.*, **21**(1), 267–286, DOI: 10.1039/C9RP00148D.

Dood A. J., Fields K. B., and Raker J. R., (2018), Using Lexical Analysis To Predict Lewis Acid–Base Model Use in Responses to an Acid–Base Proton-Transfer Reaction. *J. Chem. Educ.*, **95**(8), 1267–1275, DOI: 10.1021/acs.jchemed.8b00177.

Dood A. J. and Watts F. M., (2022), Mechanistic Reasoning in Organic Chemistry: A Scoping Review of How Students Describe and Explain Mechanisms in the Chemistry Education Research Literature. *J. Chem. Educ.*, **99**(8),

2864–2876, DOI: 10.1021/acs.jchemed.2c00313.

Dood A. J. and Watts F. M., (2023), Students' Strategies, Struggles, and Successes with Mechanism Problem Solving in Organic Chemistry: A Scoping Review of the Research Literature. *J. Chem. Educ.*, **100**(1), 53–68, DOI: 10.1021/acs.jchemed.2c00572.

Finkenstaedt-Quinn S. A., Garza N. F., Wilhelm C. A., Koutmou K. S., and Shultz G. V., (2022), Student Perceptions of Learning in Biochemistry Using a Science Communication Focused Writing Assignment. *J. Chem. Educ.*, **99**(10), 3386–3395, DOI: 10.1021/acs.jchemed.2c00171.

Finkenstaedt-Quinn S. A., Gere A. R., Dowd J. E., Thompson R. J., Halim A. S., Reynolds J. A., et al., (2022), Postsecondary Faculty Attitudes and Beliefs about Writing-Based Pedagogies in the STEM Classroom. *LSE*, **21**(3), ar54, DOI: 10.1187/cbe.21-09-0285.

Finkenstaedt-Quinn S. A., Halim A. S., Chambers T. G., Moon A., Goldman R. S., Gere A. R., and Shultz G. V., (2017), Investigation of the Influence of a Writing-to-Learn Assignment on Student Understanding of Polymer Properties. *J. Chem. Educ.*, **94**(11), 1610–1617, DOI: 10.1021/acs.jchemed.7b00363.

Finkenstaedt-Quinn S. A., Halim A. S., Kasner G., Wilhelm C. A., Moon A., Gere A. R., and Shultz G. V., (2020), Capturing student conceptions of thermodynamics and kinetics using writing. *Chem. Educ. Res. Pract.*, **21**(3), 922–939, DOI: 10.1039/C9RP00292H.

Finkenstaedt-Quinn S. A., Petterson M., Gere A., and Shultz G., (2021), Praxis of Writing-to-Learn: A Model for the Design and Propagation of Writing-to-Learn in STEM. *J. Chem. Educ.*, **98**(5), 1548–1555, DOI: 10.1021/acs.jchemed.0c01482.

Finkenstaedt-Quinn S. A., Polakowski N., Gunderson B., Shultz G. V., and Gere A. R., (2021), Utilizing Peer Review and Revision in STEM to Support the Development of Conceptual Knowledge Through Writing. *Written Communication*, **38**(3), 351–379, DOI: 10.1177/07410883211006038.

Finkenstaedt-Quinn S. A., Snyder-White E. P., Connor M. C., Gere A. R., and Shultz G. V., (2019), Characterizing Peer Review Comments and Revision from a Writing-to-Learn Assignment Focused on Lewis Structures. *J. Chem. Educ.*,

96(2), 227–237, DOI: 10.1021/acs.jchemed.8b00711.

Finkenstaedt-Quinn S. A., Watts F. M., Petterson M. N., Archer S. R., Snyder-White E. P., and Shultz G. V., (2020), Exploring Student Thinking about Addition Reactions. *J. Chem. Educ.*, **97**(7), 1852–1862, DOI: 10.1021/acs.jchemed.0c00141.

Finkenstaedt-Quinn S. A., Watts F. M., and Shultz G. V., (2024), Reading, receiving, revising: A case study on the relationship between peer review and revision in writing-to-learn. *Assessing Writing*, **59**, 100808, DOI: 10.1016/j.asw.2024.100808.

Finkenstaedt-Quinn S. A., Watts F. M., Shultz G. V., and Gere A. R., (2023), A Portrait of MWrite as a Research Program: A Review of Research on Writing-to-Learn in STEM through the MWrite Program. *International Journal for the Scholarship of Teaching and Learning*, **17**(1).

Flower L. and Hayes J. R., (1981), A Cognitive Process Theory of Writing. *College Composition and Communication*, **32**(4), 365, DOI: 10.2307/356600.

Fritz C. O., Morris P. E., and Richler J. J., (2012), Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, **141**(1), 2–18, DOI: 10.1037/a0024338.

Frost S. J. H., Yik B. J., Dood A. J., de Arellano D. C.-R., Fields K. B., and Raker J. R., (2023), Evaluating electrophile and nucleophile understanding: a large-scale study of learners' explanations of reaction mechanisms. *Chem. Educ. Res. Pract.*, **24**(2), 706–722, DOI: 10.1039/D2RP00327A.

Galloway K. R., Stoyanovich C., and Flynn A. B., (2017), Students' interpretations of mechanistic language in organic chemistry before learning reactions. *Chem. Educ. Res. Pract.*, **18**(2), 353–374, DOI: 10.1039/C6RP00231E.

Gere A. R., Limlamai N., Wilson E., MacDougall Saylor K., and Pugh R., (2019), Writing and Conceptual Learning in Science: An Analysis of Assignments. *Written Communication*, **36**(1), 99–135, DOI: 10.1177/0741088318804820.

Gere A. Ruggles., Knutson A. V., Limlamai N., McCarty R., and Wilson E., (2018), A Tale of Two Prompts: New Perspectives on Writing-to-Learn Assignments. *The WAC Journal*, **29**(1), 147–188, DOI: 10.37514/WAC-J.2018.29.1.07.

Graulich N., Hedtrich S., and Harzenetter R., (2019), Explicit *versus* implicit similarity – exploring relational conceptual understanding in organic chemistry. *Chem. Educ. Res. Pract.*, **20**(4), 924–936, DOI: 10.1039/C9RP00054B.

Graulich N. and Schween M., (2018), Concept-Oriented Task Design: Making Purposeful Case Comparisons in Organic Chemistry. *J. Chem. Educ.*, **95**(3), 376–383, DOI: 10.1021/acs.jchemed.7b00672.

Greenbowe T. J., Poock J. R., Burke K. A., and Hand B. M., (2007), Using the Science Writing Heuristic in the General Chemistry Laboratory To Improve Students' Academic Performance. *J. Chem. Educ.*, **84**(8), 1371, DOI: 10.1021/ed084p1371.

Grimberg B. I. and Hand B., (2009), Cognitive Pathways: Analysis of students' written texts for science understanding. *International Journal of Science Education*, **31**(4), 503–521, DOI: 10.1080/09500690701704805.

Gupte T., Watts F. M., Schmidt-McCormack J. A., Zaimi I., Gere A. R., and Shultz G. V., (2021), Students' meaningful learning experiences from participating in organic chemistry writing-to-learn activities. *Chem. Educ. Res. Pract.*, **22**(2), 396–414, DOI: 10.1039/D0RP00266F.

Halim A. S., Finkenstaedt-Quinn S. A., Olsen L. J., Gere A. R., and Shultz G. V., (2018), Identifying and Remediating Student Misconceptions in Introductory Biology via Writing-to-Learn Assignments and Peer Review. *LSE*, **17**(2), ar28, DOI: 10.1187/cbe.17-10-0212.

Hand B., Hohenshell L., and Prain V., (2007), Examining the effect of multiple writing tasks on Year 10 biology students' understandings of cell and molecular biology concepts. *Instr Sci*, **35**(4), 343–373, DOI: 10.1007/s11251-006-9012-3.

Hand B., Wallace C. W., and Yang E., (2004), Using a Science Writing Heuristic to enhance learning outcomes from laboratory activities in seventh-grade science: quantitative and qualitative aspects. *International Journal of Science Education*, **26**(2), 131–149, DOI: 10.1080/0950069032000070252.

Hayes J. R., (1996), A new framework for understanding cogni- tion and affect in writing, in *The Science of Writing: Theories, Methods, Individual Differences, and Applications*, Levy C. M. and Ransdell S. (eds.)., pp. 1–27.

Keiner L. and Graulich N., (2021), Beyond the beaker: students' use of a scaffold to connect observations with the particle level in the organic chemistry laboratory. *Chem. Educ. Res. Pract.*, **22**(1), 146–163, DOI: 10.1039/D0RP00206B.

Keiner L. and Graulich N., (2020), Transitions between representational levels: characterization of organic chemistry students' mechanistic features when reasoning about laboratory work-up procedures. *Chemistry Education Research and Practice*, **21**, 469–482, DOI: 10.1039/c9rp00241c.

Klein P. D., (2015), Mediators and Moderators in Individual and Collaborative Writing to Learn. *Journal of Writing Research*, **7**(1), 201–214, DOI: 10.17239/jowr-2015.07.01.08.

Klein P. D. and Leacock T. L., (2012), Distributed Cognition as a Framework for Understanding Writing, in *Past, present, and future contributions of cognitive writing research to cognitive psychology*, Berninger V. W. (ed.)., Psychology Press, pp. 133–152.

Kranz D., Schween M., and Graulich N., (2023), Patterns of reasoning – exploring the interplay of students' work with a scaffold and their conceptual knowledge in organic chemistry. *Chem. Educ. Res. Pract.*, **24**(2), 453–477, DOI: 10.1039/D2RP00132B.

Lincoln Y. S. and Guba E. G., (1985), *Naturalistic Inquiry*, Sage Publications.

Logan K. and Mountain L., (2018), Writing Instruction in Chemistry Classes: Developing Prompts and Rubrics. *J. Chem. Educ.*, **95**(10), 1692–1700, DOI: 10.1021/acs.jchemed.8b00294.

Machamer P., Darden L., and Craver C. F., (2000), Thinking about Mechanisms. *Philosophy of Science*, **67**(1), 1–25.

Martin P. P. and Graulich N., (2023), When a machine detects student reasoning: a review of machine learning-based formative assessment of mechanistic reasoning. *Chem. Educ. Res. Pract.*, **24**(2), 407–427, DOI: 10.1039/D2RP00287F.

McDermott M. A. and Hand B., (2016), Modeling Scientific Communication with Multimodal Writing Tasks: Impact on Students at Different Grade Levels, in *Using Multimodal Representations to Support Learning in the Science Classroom*, Hand B., McDermott M., and Prain V. (eds.)., Springer International

Publishing, pp. 183–211, DOI: 10.1007/978-3-319-16450-2_10.

Moon A., Gere A. R., and Shultz G. V., (2018), Writing in the STEM classroom: Faculty conceptions of writing and its role in the undergraduate classroom. *Sci. Ed.*, **102**(5), 1007–1028, DOI: 10.1002/sce.21454.

Moon A., Zotos E., Finkenstaedt-Quinn S., Gere A. R., and Shultz G., (2018), Investigation of the role of writing-to-learn in promoting student understanding of light–matter interactions. *Chem. Educ. Res. Pract.*, **19**(3), 807–818, DOI: 10.1039/C8RP00090E.

Nardi B. A., (1996), Studying Context: A Comparison of Activity Theory, Situated Action Models, and Distributed Cognition, in Nardi B. A. (ed.)., MIT Press, pp. 69–102.

Petritis S. J., Kelley C., and Talanquer V., (2021), Exploring the impact of the framing of a laboratory experiment on the nature of student argumentation. *Chem. Educ. Res. Pract.*, **22**(1), 105–121, DOI: 10.1039/D0RP00268B.

Petterson M. N., Finkenstaedt-Quinn S. A., Gere A. R., and Shultz G. V., (2022), The role of authentic contexts and social elements in supporting organic chemistry students' interactions with writing-to-learn assignments. *Chem. Educ. Res. Pract.*, **23**(1), 189–205, DOI: 10.1039/D1RP00181G.

Petterson M. N., Watts F. M., Snyder-White E. P., Archer S. R., Shultz G. V., and Finkenstaedt-Quinn S. A., (2020), Eliciting student thinking about acid–base reactions *via* app and paper–pencil based problem solving. *Chem. Educ. Res. Pract.*, **21**(3), 878–892, DOI: 10.1039/C9RP00260J.

Rhoad J. S., (2017), Written Assignments in Organic Chemistry: Critical Reading and Creative Writing. *J. Chem. Educ.*, **94**(3), 267–270, DOI: 10.1021/acs.jchemed.6b00402.

Rootman-le Grange I. and Retief L., (2018), Action Research: Integrating Chemistry and Scientific Communication To Foster Cumulative Knowledge Building and Scientific Communication Skills. *J. Chem. Educ.*, **95**(8), 1284–1290, DOI: 10.1021/acs.jchemed.7b00958.

Russ R. S., Scherr R. E., Hammer D., and Mikeska J., (2008), Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education*, **92**(3), 499–525, DOI: DOI 10.1002/sce.20264.

Russell A. A., (2013), The Evolution of Calibrated Peer Review™, in *ACS Symposium Series*, Holme T., Cooper M. M., and Varma-Nelson P. (eds.)., American Chemical Society, pp. 129–143, DOI: 10.1021/bk-2013-1145.ch009.

Schmidt-McCormack J. A., Judge J. A., Spahr K., Yang E., Pugh R., Karlin A., et al., (2019), Analysis of the role of a writing-to-learn assignment in student understanding of organic acid–base concepts. *Chem. Educ. Res. Pract.*, **20**(2), 383–398, DOI: 10.1039/C8RP00260F.

Sheskin D. J., (2011), *Handbook of Parametric and Nonparametric Statistical Procedures*, 5th ed. CRC Press.

Stowe R. L. and Cooper M. M., (2017), Practicing What We Preach: Assessing "Critical Thinking" in Organic Chemistry. *J. Chem. Educ.*, **94**(12), 1852–1859, DOI: 10.1021/acs.jchemed.7b00335.

Strickland A. M., Kraft A., and Bhattacharyya G., (2010), What happens when representations fail to represent? Graduate students' mental models of organic chemistry diagrams. *Chem. Educ. Res. Pract.*, **11**(4), 293–301, DOI: 10.1039/C0RP90009E.

Topping K. J., (2009), Peer Assessment. *Theory Into Practice*, **48**(1), 20–27, DOI: 10.1080/00405840802577569.

Watts F. M., Dood A. J., and Shultz G. V., (2023), Automated, content-focused feedback for a writing-to-learn assignment in an undergraduate organic chemistry course, in *LAK23: 13th International Learning Analytics and Knowledge Conference*,., ACM, pp. 531–537, DOI: 10.1145/3576050.3576053.

Watts F. M., Dood A. J., and Shultz G. V., (2022), Developing Machine Learning Models for Automated Analysis of Organic Chemistry Students' Written Descriptions of Organic Reaction Mechanisms, in *Student Reasoning in Organic Chemistry*, Graulich N. and Shultz G. (eds.)., The Royal Society of Chemistry, pp. 285–303, DOI: 10.1039/9781839167782-00285.

Watts F. M., Dood A. J., Shultz G. V., and Rodriguez J.-M. G., (2023), Comparing Student and Generative Artificial Intelligence Chatbot Responses to Organic Chemistry Writing-to-Learn Assignments. *J. Chem. Educ.*,

acs.jchemed.3c00664, DOI: 10.1021/acs.jchemed.3c00664.

Watts F. M. and Finkenstaedt-Quinn S. A., (2021), The current state of methods for establishing reliability in qualitative chemistry education research articles. *Chem. Educ. Res. Pract.*, **22**(3), 565–578, DOI: 10.1039/D1RP00007A.

Watts F. M., Park G. Y., Petterson M. N., and Shultz G. V., (2022), Considering alternative reaction mechanisms: students' use of multiple representations to reason about mechanisms for a writing-to-learn assignment. *Chem. Educ. Res. Pract.*, **23**(2), 486–507, DOI: 10.1039/D1RP00301A.

Watts F. M., Schmidt-McCormack J. A., Wilhelm C. A., Karlin A., Sattar A., Thompson B. C., et al., (2020), What students write about when students write about mechanisms: analysis of features present in students' written descriptions of an organic reaction mechanism. *Chem. Educ. Res. Pract.*, **21**(4), 1148–1172, DOI: 10.1039/C9RP00185A.

Watts F. M., Zaimi I., Kranz D., Graulich N., and Shultz G. V., (2021), Investigating students' reasoning over time for case comparisons of acyl transfer reaction mechanisms. *Chem. Educ. Res. Pract.*, **22**(2), 364–381, DOI: 10.1039/D0RP00298D.

Williamson D. M., Xi X., and Breyer F. J., (2012), A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, **31**(1), 2–13, DOI: 10.1111/j.1745-3992.2011.00223.x.

Wilson S. B. and Varma-Nelson P., (2019), Characterization of First-Semester Organic Chemistry Peer-Led Team Learning and Cyber Peer-Led Team Learning Students' Use and Explanation of Electron-Pushing Formalism. *J. Chem. Educ.*, **96**(1), 25–34, DOI: 10.1021/acs.jchemed.8b00387.

Yaman F., (2021), Examining students' quality and perceptions of argumentative and summary writing within a knowledge generation approach to learning in an analytical chemistry course. *Chem. Educ. Res. Pract.*, **22**(4), 985–1002, DOI: 10.1039/D1RP00060H.

Yik B. J., Dood A. J., Cruz-Ramírez de Arellano D., Fields K. B., and Raker J. R., (2021), Development of a machine learning-based tool to evaluate correct Lewis acid–base model use in written responses to open-ended formative assessment items. *Chem. Educ. Res. Pract.*, **22**(4), 866–885, DOI: 10.1039/D1RP00111F.

Yik B. J., Dood A. J., Frost S. J. H., Cruz-Ramírez de Arellano D., Fields K. B., and Raker J. R., (2023), Generalized rubric for level of explanation sophistication for nucleophiles in organic chemistry reaction mechanisms. *Chem. Educ. Res. Pract.*, **24**(1), 263–282, DOI: 10.1039/D2RP00184E.

Zaimi I., Dood A. J., and Shultz G. V., (2024), The evolution of an assignment: how a Writing-to-Learn assignment's design shapes organic chemistry students' elaborations on reaction mechanisms. *Chem. Educ. Res. Pract.*, 10.1039.D3RP00197K, DOI: 10.1039/D3RP00197K.