The language of organic chemistry: is fluency the key to success?

Ahmed Youssef

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree of
Master of Science in Chemistry

Department of Chemistry and Biomolecular Sciences
Ottawa-Carleton Chemistry Institute
Faculty of Science
University of Ottawa

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AR	Augmented reality
CLT	Cognitive load theory
EDG	Electron-donating group
EEG	Electroencephalography
EPF	Electron-pushing formalism
EWG	Electron-withdrawing group
IPT	Information processing theory
IUPAC	International Union of Pure and Applied Chemistry
LDA	Lithium diisopropylamide
LO	Learning outcome
NLG	Normalized learning gain
PX	Participant X (X to be replaced with a number between 1–36)
QX	Question X (X to be replaced with a number between 1–8)
REB	Research Ethics Board
RQ	Research question
TEPR	Task-evoked pupillary response

ABSTRACT

People around the world use language to exchange ideas and connect with each other. We use the language of organic chemistry similarly, but in addition to words, we use symbols to represent chemical phenomena. One such example of symbols used in chemistry is the electron-pushing formalism (EPF), which depicts electron movement during chemical reactions. Students spend a large amount of time decoding and often make no sense of chemistry's symbolic language. This expenditure increases students' working memory load, limiting their abilities to learn new concepts, as they can only process as much information as their working memory allows. As such, students often rely on rote memorization, employ heuristics, and adopt a product-oriented thinking approach, resulting in limited reasoning abilities that hinder their learning progress.

In this study, we explored how fluency in the electron-pushing formalism (EPF) relates to second-year organic chemistry students' cognitive load, reasoning ability and chemistry argumentation skills. Additionally, we looked to reveal students' misinterpretations after learning about chemical factors affecting stability. We hypothesized that participants more fluent in the EPF would exhibit lower pupil measures due to decreased cognitive load. We also hypothesized that participants more fluent in the EPF would demonstrate a greater reasoning ability since they would have a greater cognitive capacity to engage in chemical reasoning. We employed a pre/post-experimental design separated by a learning phase with a treatment group focusing on EPF mastery and a control group focusing on acid–base chemistry mastery. The pre- and post-tests assessed participants' ability to solve problems related to the EPF and evaluated their scientific reasoning ability. To measure cognitive load, we used eye-tracking technology to capture changes in participants' average and maximum pupil diameters. Eye-tracking data was analyzed using a custom-made *Python* program, and participants' reasoning ability was analyzed by categorizing their arguments based on complexity and examining connections between chemical factors.

The findings revealed significant decreases in the average and maximum changes in pupil diameter for participants in the treatment group from the pre-test to the post-test (Z = 2.098, p = 0.036), indicating a lower cognitive load. However, post-test pupil diameters were not significantly different between groups when controlling for pre-test diameters. A significant relationship was also established between pre-test EPF scores and average pupil diameter change ($\rho(31) = 0.322$, p = 0.039). The treatment group exhibited increases in causal and overall reasoning ability compared to the control group and both groups demonstrated a more structured use of arguments in the post-test compared to the pre-test, indicating enhanced thought organization. Furthermore, we identified five common errors made by students, including misinterpreting the direction of an inductive effect, using carbocation degree of substitution as a heuristic, focusing solely on product stability without considering the feasibility of the reaction mechanism, relying heavily on electronegativity as evidence for leaving group ability, and claiming that products formed quickly are the most stable products.

This research highlights the importance of teaching in a way that minimizes cognitive load with respect to symbolism and facilitates effective information processing. Reducing cognitive load can be achieved by ensuring students develop a foundational understanding of the EPF and incorporating causal mechanistic reasoning questions to encourage chemical reasoning. Further research with larger sample sizes would provide valuable insights into these relationships and contribute to improving organic chemistry education.

ACKNOWLEDGEMENTS

I would like to take this time to acknowledge the countless individuals who have supported me throughout my academic journey these past two years. Completing my degree would have been much more difficult without their guidance, patience, and support.

To all the graduate students in the Flynn group during my time: Dr. Myriam McKenna, Jacky Deng, Alisha Szozda, Nick Streja, Samira Behroozi, and Denzel Huang: I am grateful for your guidance, support, and encouragement. These past two years have been nothing short of spectacular, and I owe a significant portion of that to you. It will be weird not hearing everyone's check-ins and updates after being used to hearing them every week for the past two years. I would also like to extend my gratitude to the undergraduate students I have had the pleasure of working and attending group meetings with, including Matthew, Allison, Sarah, Jacob, Lauren, Jade, Laurence, Mann, Hossein, Serina, Roshan, Patrick, Zahra, and Gisele. I hope that each of you is finding joy in whatever you're up to. Special shoutout to Dr. Danielle Skropeta and Dr. Siegbert Schmid for providing valuable feedback while on a sabbatical from Australia.

To my colleagues and friends who helped me pilot, test, and validate various components of my study, including Jon, Mike, Allison, Nelson, Petra, Sarah, Eryn, Elly, and Elexa, I am grateful for your time. Remember when some of you claimed you stink at organic chemistry? Well, you proved yourselves wrong by doing well on that test! Seriously, you all need to own your knowledge and believe in yourselves more.

I am deeply grateful to Mohamed Bachrouch, a good friend of mine who took time out of his busy schedule to help me write the *Python* code (which I appropriately dubbed the "*Bach Filter*") used to analyze the eye-tracking data and really showed me the true power of coding. I can confidently say that you helped me through the greatest challenge of my degree. Without your help, I probably would still be scratching my head trying to figure it all out, so thank you!

I would like to thank my thesis committee members, Dr. David Trumpower and Dr. Natalie Goto. I am grateful for your time and contributions to my research. I would like to acknowledge Dr. Kathy Focsaneanu for her assistance in validating the questions used in my pre-/post-test, and I would also like to thank Dr. Claudia El Nachef for allowing me to recruit

participants from her Organic Chemistry II course sections. I would like to give a shoutout to Chloé, Annette, Hajar, Linda, Annik, Victoria, and the various academic support staff in the department who had to deal with my requests, inquiries, and questions. You are all treasured members of the department, and I appreciate you looking out for us.

I am deeply grateful to the *Social Sciences and Humanities Research Council* for recognizing my research's importance and awarding me a *CGS-M* scholarship. I would also like to thank the Office of the Vice-Provost for nominating me as a finalist for the *uOGRADflix* competition and all my friends who came out to support me. I would like to thank all the students who participated in my study. Thank you for taking time from your loaded undergraduate schedules to help a random graduate student.

Finally, I would like to extend a special thanks to Dr. Alison Flynn. You have been an incredible mentor from day one. Your support and patience have inspired me to pursue excellence in my work and have been a constant source of motivation. You fostered a great working environment, and I know you've probably heard it from me before, but I want to re-emphasize it all starts at the top. Despite your talent for collecting USBs like souvenirs and your efforts to persuade me of winter's likability (even though my opinion remains unchanged for the most part), I still feel incredibly fortunate to have had the opportunity to work with you. Thank you… a thousand times.

As an undergraduate student, I vividly recall completing my Organic Chemistry II final exam with about 40 minutes to spare and just sitting at my desk with a smile on my face. The TAs were confused and thought I was up to something, and despite telling me that I could leave, I just sat at my empty desk, smiling. How fitting that almost six years later, I am sitting at my desk, as I type this, with the same smile.

INTRODUCTION

The role of symbolism in chemistry education

Language is an integral part of society as it is the primary means of communication among individuals within a community. People can express ideas, share information, and understand others' perspectives. Chemistry possesses its own unique language: a set of symbols that are used to represent macroscopic and submicroscopic phenomena (Mathewson, 2005; Gkitzia *et al.*, 2011; Sim and Daniel, 2014). Whether writing out a chemical formula in a general chemistry course, deriving a mathematical equation in an analytical chemistry course, or looking at a phase diagram in a physical chemistry course, symbolism is heavily used in a chemistry classroom setting. Many studies in chemistry education look at students' learning approaches (Grove and Bretz, 2012; Graulich, 2015a), the amount of effort required to learn (Cranford *et al.*, 2014; Milenković, Segedinac, and Hrin, 2014), and the ability to think like a chemist (Kraft *et al.*, 2010; Dood *et al.*, 2020). In some cases, the curricula of institutions have been redesigned to help address the consequences associated with these factors (Grove *et al.*, 2008; Lipton, 2020). Yet not many studies explore how fluency in the language of chemistry may influence these factors.

Symbols are very prominent in chemistry to the point that they constitute one-third of Johnstone's Triangle, a framework that presents the various modes of representing chemistry (Johnstone, 1982, 2000). Macroscopic chemistry (that which is observable by the eye) and submicroscopic chemistry (that which occurs on an atomic/molecular level) make up the remaining two-thirds of the framework (**Figure 1**). Though Johnstone's Triangle paints the picture that the three levels of chemistry are equivalent, the symbolic level predominates in the classroom. Nyachwaya and Wood (2014) evaluated the chemical representations in physical chemistry textbooks and found that 85% of those representations were of the symbolic level. Taber recognized that the symbolic level could act as a bridge between the other two groups and help build knowledge that explains chemical phenomena (2013). Yet a paper published by Gkitzia *et al.* (2020) found that third-year undergraduate chemistry students' abilities to transfer between the three chemistry levels was unsatisfactory. A heavy emphasis on teaching

the symbols without incorporating elements of the macroscopic and submicroscopic levels may lead to "levels confusion," where students mistakenly think that characteristics from one of the levels apply to another (Stieff *et al.*, 2013).

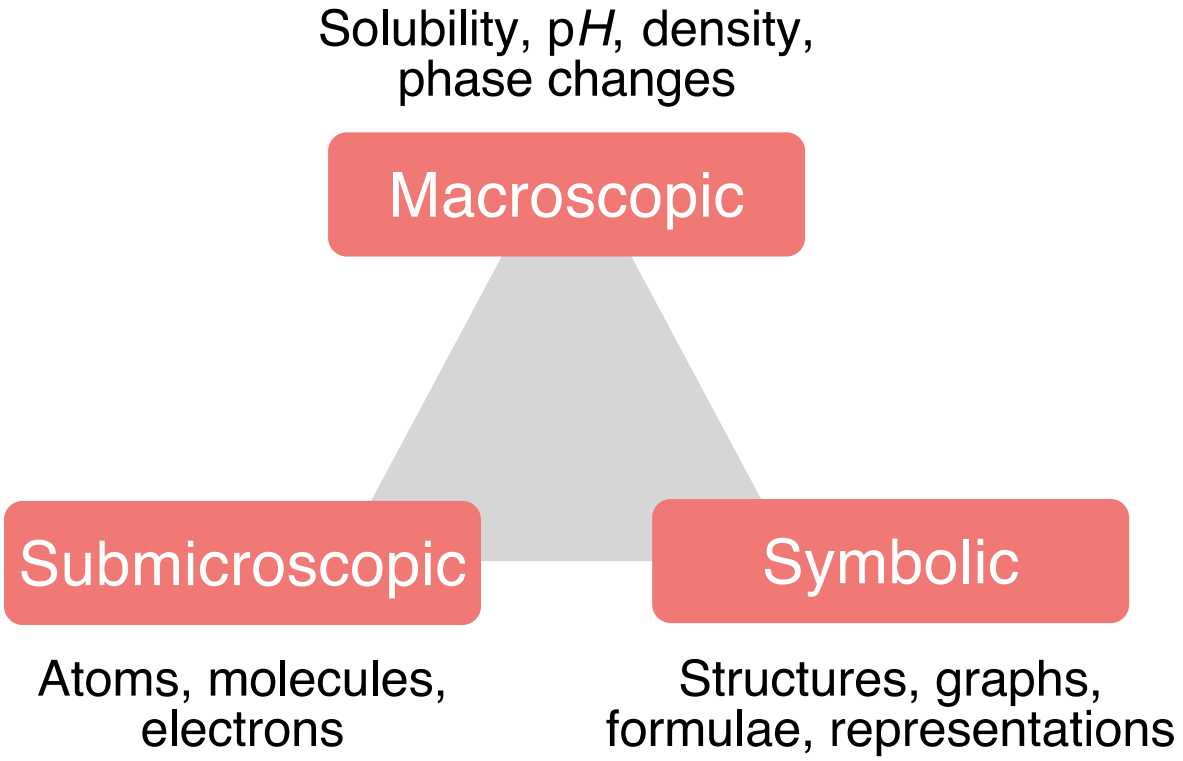


***Figure 1**. Composition of Johnstone's Triangle, also known as Chemistry's triplet.*

Regardless of the specific discipline, learning and interpreting these symbols, like any other language, can be challenging for novices (Talanquer, 2011; Taber, 2013; Taskin and Bernholt, 2014). Not only that, a study conducted in a physical chemistry classroom found that students spent approximately half the allotted time decoding symbols rather than solving problems (Becker *et al.*, 2015). If these participants had been more fluent in the symbolism, they might have been able to allocate more time and cognitive resources to problem solving.

Organic chemistry is taught primarily by symbols, and like other chemistry subjects, interpreting these symbols requires a lot of time and effort (Cooper *et al.*, 2012, 2013). One can argue that with its close partner, inorganic chemistry, organic chemistry may be the most symbol-heavy chemistry discipline (Erduran, 2019). The subject deals less with numbers that students have worked with all their life and focuses more on notations and representations used to convey the structure and reactions of organic compounds. On the other hand, the

symbols seen by participants in the study by Becker *et al.* (2015) were related to physical chemistry, which often involve numbers, constants, and other quantitative data. One important aspect of the symbolism in organic chemistry is the electron-pushing formalism (EPF). Organic chemists use this notation to communicate and explain how a chemical reaction proceeds (**Figure 2**). This formalism uses bond-breaking and bond-forming arrows to represent the movement of electrons. These one-sided arrows start at a pair of electrons (a lone pair or a bond) and point to an electron-deficient atom or bond. A collection of these steps that show a starting material (reactant) being converted into a product is known as a reaction mechanism. Researchers argue that the EPF is essential for students' development of conceptual understanding (Straumanis and Ruder, 2009). Mastery of organic chemistry hinges upon students' knowledge of the EPF and other symbols used in the subject, enabling them to use them alongside the principles of reactivity and related concepts (Goodwin, 2008; Graulich, 2015b).

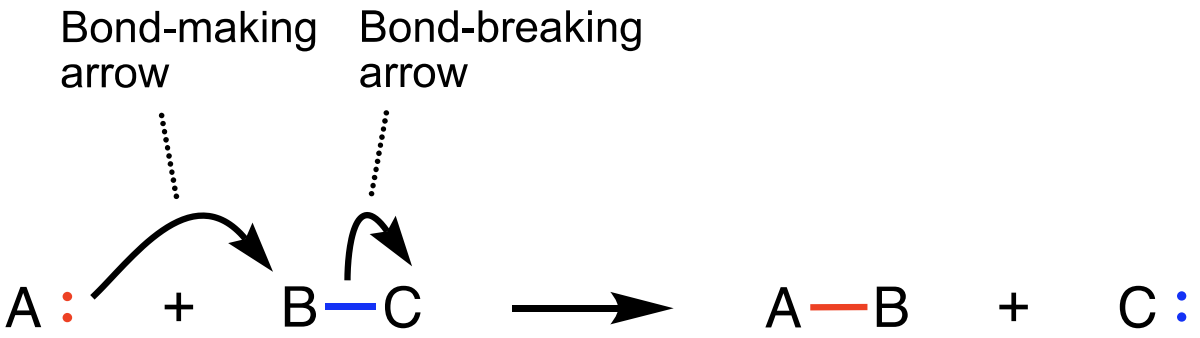


***Figure 2**. General overview of the electron-pushing formalism (EPF). In this reaction, the electrons of A (denoted by dots) collide with B to create a bond, while the bond between B and C breaks, leaving electrons on C. The bonds/electrons have been colour-coded, making them easier to distinguish.*

The limitations of rote memorization in learning organic chemistry

Despite the importance of the EPF to student understanding, many studies have shown that students struggle to properly use and interpret the EPF when working with reaction mechanisms. In a study by Bhattacharyya and Bodner (2005), 14 graduate students were given a familiar reaction they learned in their undergraduate courses. While approximately 50% of the participants drew the correct mechanism instantly, none of them could explain why any of the steps occurred. This observation indicates that some participants memorized entire

reaction mechanisms from start to finish without understanding the underlying chemical factors that guide the progression of the reaction. Another study found that many students could not provide a mechanism for a given reaction, and approximately one-fifth of participants drew the arrows only after being provided the products of the reaction (Grove, Cooper, and Rush, 2012). This finding demonstrates that some students work backwards and use the arrows as an afterthought (Grove, Cooper, and Rush, 2012). Students not considering the importance of a reaction mechanism is apparent, as several researchers have identified that students employ product-oriented thinking (i.e., thinking about the outcome of a reaction) rather than process-orientated thinking (i.e., thinking about the steps required to reach the outcome) (Grove, Cooper, and Cox, 2012; Dood and Watts, 2023).

One of the major issues of organic chemistry is students' reliance on memorization-oriented learning (Grove and Bretz, 2010, 2012; Graulich, 2015a). This strategy is frequently employed by beginner organic chemistry students (Bhattacharyya and Bodner, 2005; Flynn, 2011). In a recent study, 184 organic chemistry students were presented with the statement, "*Memorization was a large part of preparation for multi-step synthesis*." The students were asked to indicate their agreement on a scale of 1 to 5, resulting in a collective average response of 4.1. (Salame *et al.*, 2020). In addition, students were asked, "*Why do you think organic synthesis is challenging?*" to which more than half said the process was challenging because they had to "*remember/memorize all reactions, reagents and rules*" (Salame *et al.*, 2020). Students resort to rote memorization when a lot of information is presented, such as during an organic chemistry course (Newell and Shanks, 2003). One can assume that undergraduate students have several evaluations and tasks from other courses. As such, the students would be more inclined to employ a time-efficient study strategy, such as rote memorization. Study strategies requiring more critical thinking or analytical and abstract thought processes are much more time-consuming (Evans, 2003), and it may be likely that students do not have time to use these approaches.

Another possible explanation for students resorting to rote memorization could be the source itself—how the course is taught. The traditional way of teaching organic chemistry involves organizing reactions by functional group, which can often result in fragmentation of

knowledge (Lipton, 2020). An example of this knowledge fragmentation is the popular S_N2 mechanism being introduced in textbooks without mentioning the hybridization conditions required for the reaction (Loudon and Parise, 2016; Vollhardt and Schore, 2018; Lipton, 2020). The idea here is that an important factor influencing the feasibility of the reaction is not discussed when the reaction itself is introduced, making the learning process more difficult for students. This idea also relates to the use of heuristics, which are essentially "*rules of thumb for inference and choice*" (MacGillivray, 2014). Students are taught heuristics such as "*the more substituted the carbocation, the more stable*," and assume that they apply in all cases, not realizing that there are usually exceptions to every rule in science. Lastly, rote memorization limits the development of students' problem-solving abilities. The strategy has several consequences, including disengagement and seeing information as individual entities of fact rather than using them in conjunction with one another (Flynn, 2011; Grove, Cooper, and Cox, 2012).

Meaningfully learning organic chemistry allows students to predict reactivity better and develop a deeper understanding of the subject of interest (Ausubel, 1963; Kraft *et al.*, 2010). Not only does the learning strategy promote information transfer, but it also allows students to connect their prior knowledge with new information. Additionally, understanding builds on several cognitive skills, including interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining (Mayer, 2002). In contrast, rote memorization only relies on the cognitive skill of recall, focusing on remembering information stored in learners' long-term memories (Mayer, 2002). In addition, meaningful learning has been associated with higher self-efficacy levels, ultimately contributing to increased academic performance and retention (Richardson *et al.*, 2012; Hacıeminoğlu, 2021), whereas rote memorization does the opposite—it is associated with lower self-efficacy levels and decreased academic performance and retention. To fully understand how these learning approaches impact the ability to learn, the capacity to process information needs to be explored.

Cognitive load in chemistry: reduction strategies and common measures

In psychology, the capacity of our working memory dictates how much information we can hold and our ability to engage in cognitive tasks. The degree to which mental strain is placed on the working memory while completing a task is defined as "*cognitive load*" (Sweller, 1988). The construct is a dynamic internal process that has been shown to influence the parasympathetic and sympathetic nervous systems (Solhjoo *et al.*, 2019). Cognitive load is frequently used with cognitive load theory (CLT), a framework that assumes that we possess a limited storage capacity for our working memory (Kirschner, 2002). As such, the framework emphasizes "*learning by reducing working memory load*" (Mavilidi and Zhong, 2019). Unsurprisingly, chemistry experts possess the skills necessary to arrange their knowledge in a manner that conserves working memory capacity, allowing them to analyze information and make inferences more efficiently. (Kraft *et al.*, 2010). Experts achieve this feat by grouping concepts into meaningful "chunks," allowing them to process more information than novices (Thalmann *et al.*, 2019). Unfortunately, meaningfully understanding organic chemistry principles requires skills that elicit higher working memory loads, such as organizing knowledge into schematized representations (Richland *et al.*, 2012). Much research in several disciplines of education has looked at ways to reduce cognitive load for students. Researchers have studied how students' cognitive load is influenced by augmented reality (AR) (Buchner *et al.*, 2022), interactive videos (Liao *et al.*, 2019; Afify, 2020), mobile learning (Zhonggen *et al.*, 2019), and group collaboration (Liao *et al.*, 2019). These recent studies show that teaching by lowering cognitive load interests educational researchers and highlights the need for similar studies in chemistry.

Cognitive load-related studies have also been conducted in chemistry education settings. Keller *et al.* (2021) conducted a study using AR to help address students' difficulties in visualizing mental rotations of molecules and discerning spatial abilities. They found that using AR reduced the students' cognitive load (Keller *et al.*, 2021). Other researchers investigated the effectiveness of using heart rate to measure cognitive load while students worked through chemistry questions designed to elicit a greater cognitive load (Cranford *et al.*, 2014). Results from this study showed that the more difficult questions often increased participants' heart

rates, and students' heart rates were typically two to four times greater than faculty members who completed the same questions (Cranford *et al.*, 2014). The students' increased heart rates could be attributed to experiencing a higher cognitive load than professors, as they required more mental effort to complete the questions. Another research study used a seven-point Likert scale to elicit cognitive load responses from students (Milenković, Segedinac, Hrin, *et al.*, 2014). They were asked to rate the level of difficulty for tasks related to the three levels of Johnstone's Triangle, ranging from "extremely easy" to "extremely difficult," and results showed that the highest cognitive load corresponded to tasks related to the submicroscopic level (Milenković, Segedinac, Hrin, *et al.*, 2014).

Measuring cognitive load is challenging because it is an internal mental process. As such, measures that can track parasympathetic or sympathetic nervous system changes are considered effective techniques. These measures include physiological parameters such as heart rate variability, electroencephalography (EEG) and pupillometry (Urrestilla and St-Onge, 2020). Each measure has its advantages and disadvantages. Heart rate variability has proven effective in some contexts, but Urrestilla and St-Onge (2020) have shown the measure to be sensitive to changes in the electromagnetic field and breathing cycle. EEG is advantageous because it is affordable and allows us to measure brain activity directly; however, it is the most invasive of the three methods, as several electrodes must be placed on the scalp. Pupillometry was selected as the physiological measure of cognitive load in this study because it is non-invasive compared to EEG, where numerous electrodes are typically attached to the scalp, and not sensitive to changes in the electromagnetic field and the breathing cycle, both of which influence heart rate variability results (Urrestilla and St-Onge, 2020).

Pupillometry is the measurement of changes in the diameter of the pupil. This technique has proven useful in psychology, as higher pupil dilations have been shown to correlate with cognitive processing (Hess and Polt, 1964; Kahneman and Beatty, 1966). Specifically, pupillometry studies typically rely on the mechanism of task-evoked pupillary responses (TEPRs), which are the changes in pupil diameter in response to cognitive and emotional tasks (Beatty, 1982). Researchers have found links between pupillary response and task difficulty in visual searches (Porter *et al.*, 2007), multiplication (Hess and Polt, 1964) and digit recall

(Kahneman and Beatty, 1966; Piquado *et al.*, 2010). Unfortunately, pupillometry is not a perfect measure, as pupil diameter can be influenced by many external factors, including arousal (Hess and Polt, 1960), pain (Lowenstein, 1950), fatigue (Gilzenrat *et al.*, 2010), medication (Naicker *et al.*, 2016), stress (Qin *et al.*, 2012) and ambience.

Reasoning ability in organic chemistry education

Reasoning is a crucial skill in organic chemistry, yet there is limited use of the skill in organic chemistry education (Christian and Talanquer, 2012). Research has shown that it takes students many years of practice to reach a similar level of reasoning as organic chemists (Bodner and Domin, 2000). Additionally, students' arguments lack conceptual depth, even when they make a correct claim (Bodé *et al.*, 2019). Incorporating student reasoning allows educators to evaluate students' abilities to think like scientists rather than simply evaluating their ability to store content taught in a course (Berland and Reiser, 2011; Evagorou and Osborne, 2013). Drawing the product of a chemical reaction, drawing the mechanism of a chemical reaction and explaining why a reaction proceeds the way it does each require different levels of knowledge (Bodé *et al.*, 2019). However, students cannot answer the latter type of question effectively without fully developed reasoning skills.

Our group previously analyzed students' arguments in terms of four modes of reasoning: descriptive, relational, linear causal, and multi-component causal (Bodé *et al.*, 2019; Deng and Flynn, 2021; Deng *et al.*, 2022). The research findings demonstrated that many students exhibited proficiency in causal reasoning, which frequently coincided with correct chemical argumentation. These modes of reasoning were adapted from other researchers and increase in complexity in the order they are listed (Sevian and Talanquer, 2014). **Table 1** summarizes the differences between the modes of reasoning using a statement that argues why hydrochloric acid is more acidic than acetic acid.

***Table 1**. Modes of reasoning and example explanations.*

Mode of Reasoning	Definition	Example
Descriptive	No evidence provided or connected to claim	*"Hydrochloric acid is more acidic than acetic acid. It has a pK_a of –6, and acetic acid has a pK_a of 5."*
Relational	Evidence linked to claim with a correlation	*"Hydrochloric acid is more acidic than acetic acid due to their differing pK_a values (–6 for HCl and 5 for acetic acid). The lower the pKa value, the stronger the acid."*
Linear causal	Evidence linked to a claim with supporting reasons	*"HCl is more acidic than acetic acid because its conjugate base is more stable. This statement is supported by pK_a values (lower pK_a = more acidic). HCl has a lower pK_a value (–6) compared to acetic acid's pK_a value (5)."*
Multi-component causal	Multiple pieces of evidence linked to the claim with multiple reasons that explain why the evidence is relevant	*"Hydrochloric acid is more acidic than acetic acid because its pK_a value is –6, while acetic acid's pK_a value is 5. The difference is due to the relative stability of their conjugate bases being affected by multiple factors: the larger chlorine atom means chloride is more stable than acetate. However, the acetate is stabilized by the resonance effect, the carbonyl's inductive effect, and the oxygen atom's electronegativity bearing the negative charge. The pK_a values indicate chloride is more stable than acetate, so we can conclude that the larger atom has the greatest impact on relative base stability, making HCl the stronger acid."*

Descriptive reasoning is the most basic mode of reasoning. In this argument, a claim and evidence are present but are not linked. In the example shown in **Table 1**, the claim is "*hydrochloric acid is more acidic than acetic acid,*" and the evidence is the listed pK_a values that follow. However, these two elements are independent and unlinked, leaving readers questioning how pK_a influences acidity.

Relational reasoning is an intermediate mode of reasoning that links evidence and the claim with a correlation. In the example provided in **Table 1**, the evidence and claim are the same as the descriptive mode of reasoning example; however, there is an explanation that establishes a relationship between the two items.

Linear causal reasoning is an advanced mode of reasoning because it provides an explanation that establishes the relevance of the evidence to the claim. The example used in

Table 1 shows an argument like the relational one, but the main difference is a causal argument (conjugate base stability) establishes the link between the two elements in this example.

Multi-component causal reasoning is the most advanced mode of reasoning, requiring multiple explanations to support the relevance of each piece of evidence. In some cases, these pieces of evidence may work against each other and one needs to assess the strength of each factor. The multi-component causal example in **Table 1** shows all relevant pieces of evidence being listed and the factors being weighed against each other. In the end, a decision is made using the provided experimental data.

The ability of students to engage in causal reasoning is an important skill to develop as it can help prepare them for practical settings where the "correct answer" may not always be clear. By delving deeper into the "how" and "why" aspects, students can develop a scientific mindset that relies on evidence to justify inferences. This approach empowers students to question conventional "facts" and instead make well-informed decisions. Such critical thinking skills are valuable in navigating widely debated scientific issues like vaccine mandates and climate change, which are often plagued by factually incorrect evidence (McCright and Dunlap, 2011; Shelby and Ernst, 2013). Educators should therefore structure their courses and provide resources that can help students think critically and engage in deeper reasoning.

Curriculum redesign and learning modules for improving organic chemistry education

To address students' difficulties interpreting the EPF and also their lack of knowledge of important organic chemistry concepts (Bhattacharyya and Bodner, 2005; Ferguson and Bodner, 2008; Bhattacharyya, 2014; Anzovino and Bretz, 2015, 2016), many organic chemistry educators have modified their course curricula such that functional group type is no longer the main factor of organization (Bowman *et al.*, 2007; Grove *et al.*, 2008; Flynn and Ogilvie, 2015; Mooring *et al.*, 2016; Cooper *et al.*, 2019). These studies showed many promising results, including reduced attrition rates (Grove *et al.*, 2008), increased emotional satisfaction (Mooring *et al.*, 2016), mechanism competence, and long-term retention of the taught material (Bowman *et al.*, 2007).

The curriculum at the University of Ottawa was changed from a traditional functional group approach to one that focuses more on the different patterns and principles of reactivity (Flynn and Ogilvie, 2015). The efficacy of this curriculum redesign was tested via several card-sort studies that investigated how students organize reactions early into the course and toward the end. Findings from the study showed that students' categorizations became more organized (with a closer resemblance to that of organic chemistry experts) and more advanced (Galloway *et al.*, 2019; Lapierre and Flynn, 2020), and students that performed better in the card-sort task were more likely to perform better in their organic chemistry exams (Lapierre *et al.*, 2022). Additionally, students' EPF interpretations seemed less superficial, as they thought about the formalism less as arrows without any meaning and more about electron movement and bond breakage/formation (Galloway *et al.*, 2017). Lastly, students performed better on mechanism questions in tests, as seen by their higher scores (Webber and Flynn, 2018).

To further aid student success, the *OrgChem101* open education resource was created by Flynn and colleagues to help students learn key concepts in organic chemistry. The program is free for anyone to use and comprises three modules. *Organic Nomenclature: Name it + Draw it → Master It* (Flynn *et al.*, 2014; Bodé *et al.*, 2016) was the first module created. This program was designed to help students learn the names of organic chemistry structures by creating personalized quizzes and showing how some compounds are used in practical settings. The second module is *Acid–Base Reactions: Mastering the protons*, designed to address students' difficulties with acid–base chemistry (Stoyanovich *et al.*, 2015; Flynn and Amellal, 2016). This module incorporates several key elements for effective learning, including metacognitive support, educational videos, practice questions, and links to the real world. Such formatting allows students to monitor and evaluate their learning and have the opportunity for meaningful reflection on their knowledge. The *Organic Mechanisms: Mastering the arrows (OrgMech101)* module follows a similar format to the *Acid–Base Reactions* module, but the main focus is the teachings of the EPF (Visser and Flynn, 2018), such as drawing electron-pushing arrows and the products of a reaction. **Figure 3** details the features and learning outcomes of the module.

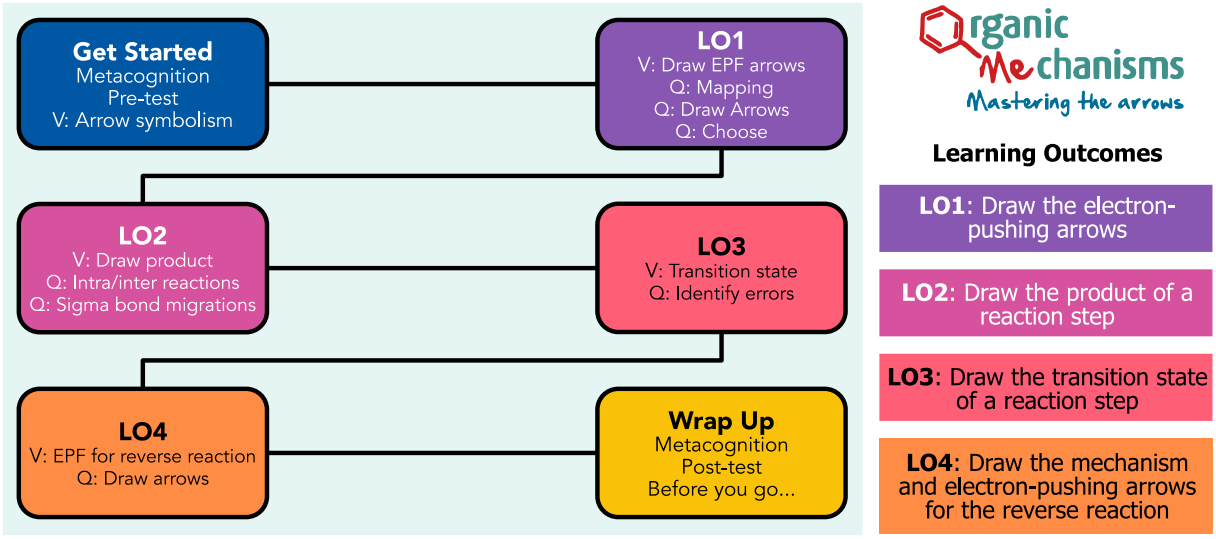


Figure 3*. Overview of the OrgMech101 module. There are four learning outcomes listed on the right-hand side. Each learning outcome contains videos (denoted by a V) and practice questions (denoted by a Q).*

Research conducted by Carle *et al.* (2020) shows that students can learn the EPF quickly using *OrgMech101*. The strategies taught in the module were employed by students post-exposure, and many stated the resource was useful for them and would help them in their organic chemistry course (Visser and Flynn, 2018; Carle *et al.*, 2020). While past research has identified students' growing abilities concerning patterns and principles of organic chemistry (Flynn and Ogilvie, 2015), we have not yet investigated the effect of EPF fluency on students' cognitive load or reasoning ability.

THEORETICAL FRAMEWORK

Information processing theory

The theoretical framework used to guide this study is Information processing theory (IPT) (Ausubel, 1968). IPT can be broken down into three main levels of memory storage: sensory, working, and long-term memory (**Figure 4**). First, a stimulus is processed in an individual's sensory memory through sensory events such as sight, taste, and sound. Information is stored differently depending on the stimulus and relevant senses, and storage at this stage is only for a short period (Tangen and Borders, 2017). At this stage, the information faces two possibilities: either it remains unprocessed, becomes overwritten by other incoming sensory inputs, and eventually becomes forgotten, or it undergoes processing in the working memory through attention (Tangen and Borders, 2017). The latter process can be achieved by promoting interest in the information or triggering pattern recognition (Huitt, 2003). Working memory is a location designated for the temporary storage of information needed for immediate tasks. Here, storage is limited, and information must be rehearsed and integrated actively to be retained (Bretz, 2001; Galloway and Bretz, 2015). The more the information is rehearsed, the more likely that the information is further processed (i.e., encoded) and transferred into long-term memory. At this stage, storage capacity is unlimited, and information can be retrieved later. If information cannot be retrieved from the long-term memory, it is considered forgotten (Tangen and Borders, 2017).

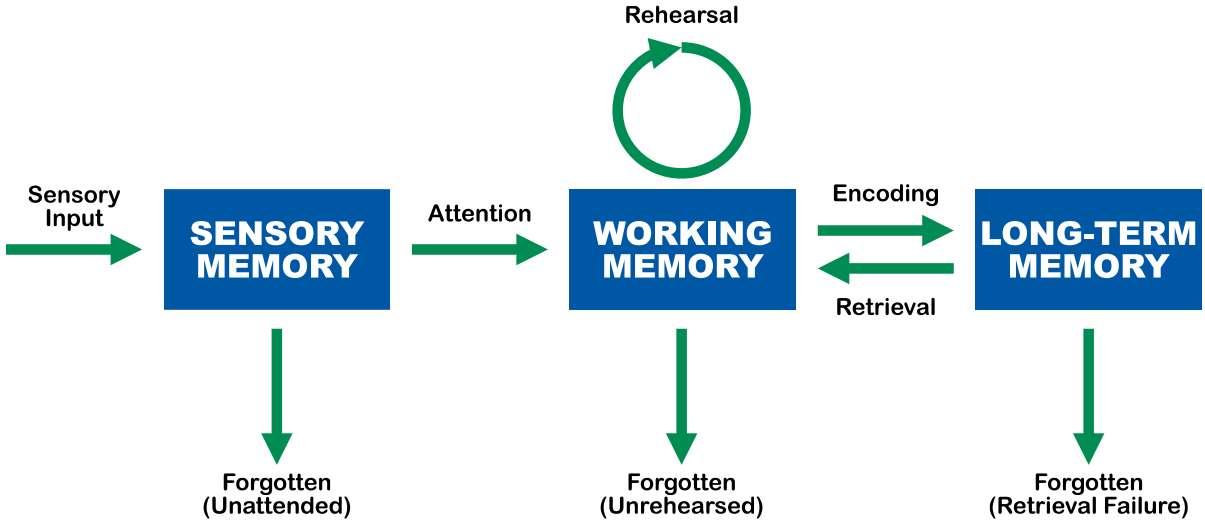


***Figure 4**. Overview of information processing theory.*

How IPT connects with the symbolism of organic chemistry goes back to the concept of meaningful learning. If the symbolism is taught meaningfully to students, they are more likely to attend to the information and hold it in their working memory (Galloway *et al.*, 2017). Additionally, connecting the new knowledge with prior knowledge in a non-random way is critical for long-term memory storage; therefore, not only do students need to have prior knowledge of the subject, but they need to see the incoming information as relevant for a connection to be made (Bretz, 2001; Novak, 2010). IPT has been modified recently to incorporate characteristics important to the learners which separate them from robots, including environment, emotions and culture (Mayer, 2012; Pazicni and Flynn, 2019).

Cognitive load theory

Cognitive load theory (CLT) is a second theoretical framework that acts as an extension of IPT. The learning theory suggests that cognitive load, or the mental effort required to process information, is essential to learning. The greater the cognitive load, the more difficult long-term information processing would be (Marcus *et al.*, 1996; Carlson *et al.*, 2003). CLT focuses on how information processing can be enhanced by managing the three types of cognitive load placed on learners (**Figure 5**).

Intrinsic load is the inherent complexity of the material being learned. This component of cognitive load depends on the number of elements and how they interact with each other. For example, if a student were looking to memorize the first twenty elements of the periodic table, the task itself would impose a higher intrinsic load compared to memorizing the first five elements, and the former would be more difficult to process in the working memory (Russell and Hannon, 2012). Although the intrinsic load depends on the task itself, a low cognitive load can be achieved by simplifying the intrinsic load. Rather than attempting to learn the first twenty elements, learning five elements in succession four times simplifies the intrinsic load, lowering the overall cognitive load in the process (Russell and Hannon, 2012).

Extraneous load is the irrelevant or unnecessary information that may interfere with learning. How information is presented is a factor that would influence extraneous load. For example, a laboratory manual with more pictures and diagrams embedded within blocks of text

would yield a smaller extraneous load than a laboratory manual with large paragraphs of text and without any images (Dechsri *et al.*, 1997). In other words, information is presented more effectively in the former, as students can visualize concepts and make connections, lowering their extraneous and overall cognitive load. The Cognitive Theory of Multimedia Learning also draws on a similar comparison, stating that we have separate channels for word and image processing, each with a limited capacity to process information (Mayer, 2021).

Germane load is the cognitive effort required to process new information (Sweller *et al.*, 1998). This component of cognitive load is generated when learners engage in deep cognitive processing, such as relating information to prior knowledge. The goal of instructional design and teaching is not to increase or decrease germane cognitive load per se, but to enhance it (Klepsch *et al.*, 2017). If the germane load is too high, learners may have difficulty processing and retaining information. On the other hand, if the germane load is too low, learners may not be engaged. Asking students to explain their answers is an excellent example of a task that optimizes germane load, as learners must access prior knowledge from their long-term memory (Kalyuga, 2011).

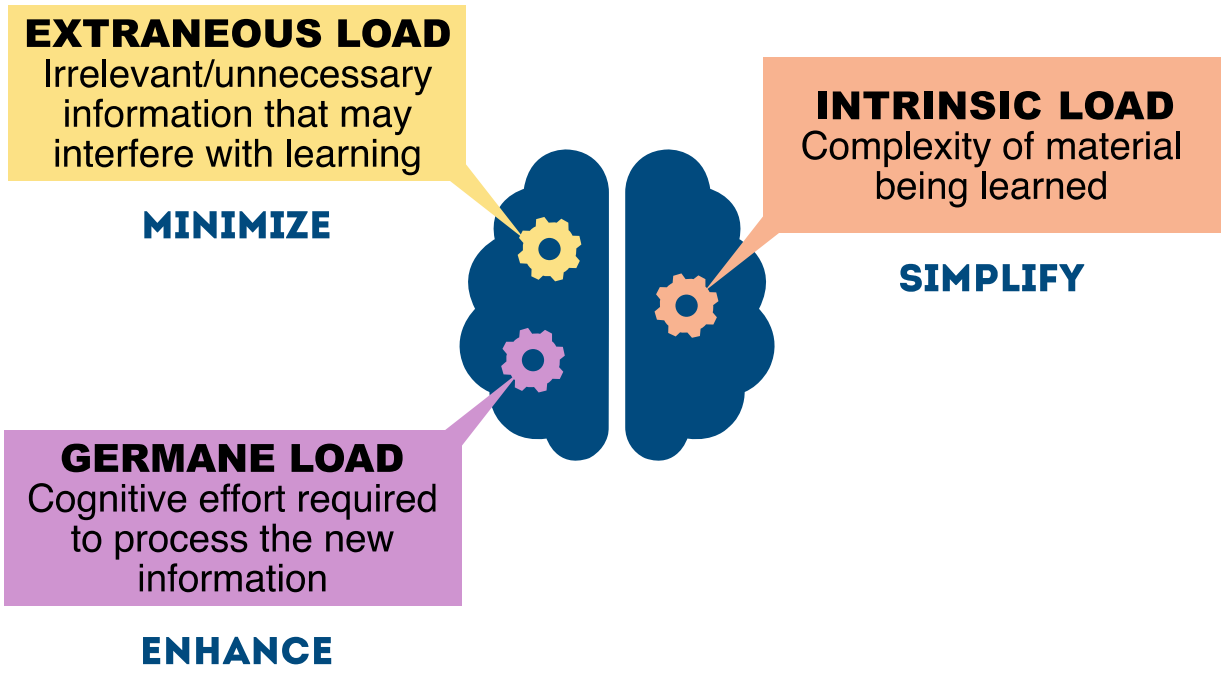


***Figure 5**. Overview of cognitive load theory, including the components that make up cognitive load.*

Examples can be made that incorporate the three concepts together. Instead of teaching acid–base chemistry in a single session with a text-heavy PowerPoint presentation, the material can be broken down into two sessions with a visually enhanced PowerPoint presentation and participation questions. By incorporating the questions into the presentation, students are encouraged to engage with the material actively, leading to a simultaneous reduction in extraneous load and an enhancement of germane load. The session being broken up in two helps simplify the intrinsic load.

The three types of cognitive loads are each important processes in managing cognitive load and information processing. Cognitive load management aims to *minimize* extraneous load, *enhance* germane load, and *simplify* intrinsic load to maximize learning (Cook, 2006). By reducing extraneous load, learners have a greater cognitive capacity available to process the new information (i.e., germane load) and to deal with the inherent complexity of the subject matter (i.e., intrinsic load).

RESEARCH QUESTIONS

This project aims to determine the importance of EPF fluency to student success by assessing its impact on students' cognitive load and reasoning ability. We investigated three research questions (RQs):

- **RQ1**: What is the relationship between students' fluency and cognitive load in organic chemistry?
- **RQ2**: How do students' abilities to reason through organic chemistry reactions vary with increased EPF fluency?
- **RQ3**: How do students' organic chemistry arguments change with increased EPF fluency?

Additionally, we will explore the following RQ unrelated to EPF fluency to provide insight into students' learning processes further:

- **RQ4**: What are students' common misinterpretations after learning about chemical factors affecting stability/reactivity?

For RQ1 and RQ2, we hypothesized that participants more fluent in the EPF would exhibit lower pupil measures due to decreased cognitive load. We also hypothesized that participants more fluent in the EPF would demonstrate greater reasoning ability since they would have a greater cognitive capacity to engage in chemical reasoning. For RQ3, we expected to see more structured arguments for participants in both groups to represent learning gains from both interventions. Lastly, for RQ4, we expected that participants would have more difficulty with case comparison questions involving multiple chemical factors than single chemical factor questions. We also anticipated that participants' errors would be due to reasons already identified in the literature (e.g., superficial use of heuristics and product-oriented thinking).

METHODS

Participant recruitment and screening

Thirty-six participants were recruited from three Organic Chemistry II courses at the University of Ottawa (two English and one French), taught by two different instructors. The Research Ethics Board (REB) at the University of Ottawa approved the study described in this work (REB File #H-01-22-7661). The researcher (AY) contacted the professors teaching the course at the start of the term and asked them to share an overview of the study and a one-minute video with the students to further encourage recruitment. The researcher (AY) also made an announcement at the start of class advertising the study and encouraging participation. Full recruitment documents are available in *Appendix A: Recruitment text and videos*. Students interested in participating completed a *SurveyMonkey* intake form, which provided essential information and allowed them to electronically consent to participate. The consent form included sections detailing the study's purpose, participation requirements, risks, benefits, confidentiality and anonymity, data conservation, compensation, voluntary participation, use of eye-tracking technology, and consent to participate. Participants were also told that they may withdraw from the study at any time.

The last section of the intake form contained eye-tracking exclusion criteria questions. If participants did not pass the exclusion criteria questions, they could still participate in the study; however, they would not use an eye tracker. Exclusion criteria questions filtered participants who wore glasses at close distances, had eye movement abnormalities, or recently had eye surgery. At the end of the intake form, participants were redirected to a booking page where they could schedule an in-person session up to a month in advance. The full list of items discussed with participants before the pre-test is shown in *Appendix B: Study introduction information*.

Study design and data collection

This study followed a pre-post design separated by a learning period with a control and treatment group. Participants were randomly assigned to the *OrgMech101* (treatment) group or the *Acid–Base Reactions* (control) group for the intervention. An automatic email was sent to

participants one week after completing the post-test. In this email, participants were asked to complete an attached test identical to the first two. This test could be completed electronically and did not require participants' in-person presence. In addition, participants were also asked to complete a demographic questionnaire via *SurveyMonkey.* Data collection spanned four months from September to December 2021. The pre-test was incorporated to provide a reference point of participants' EPF fluency level and to help account for learning gains from students as the semester progressed. In addition, the combination of the pre- and post-tests were used to measure learning gains for each participant before and after the intervention, making the increase in organic chemistry knowledge as the semester progressed an unimportant factor. Participants were compensated with a $40 Amazon gift card, with $10 portions being secured at different stages of the study to encourage full participation. These stages were the completion of the following tasks: the pre-test, post-test, delayed post-test, and demographic questionnaire. **Figure 6** summarizes the overall workflow of the study described above.

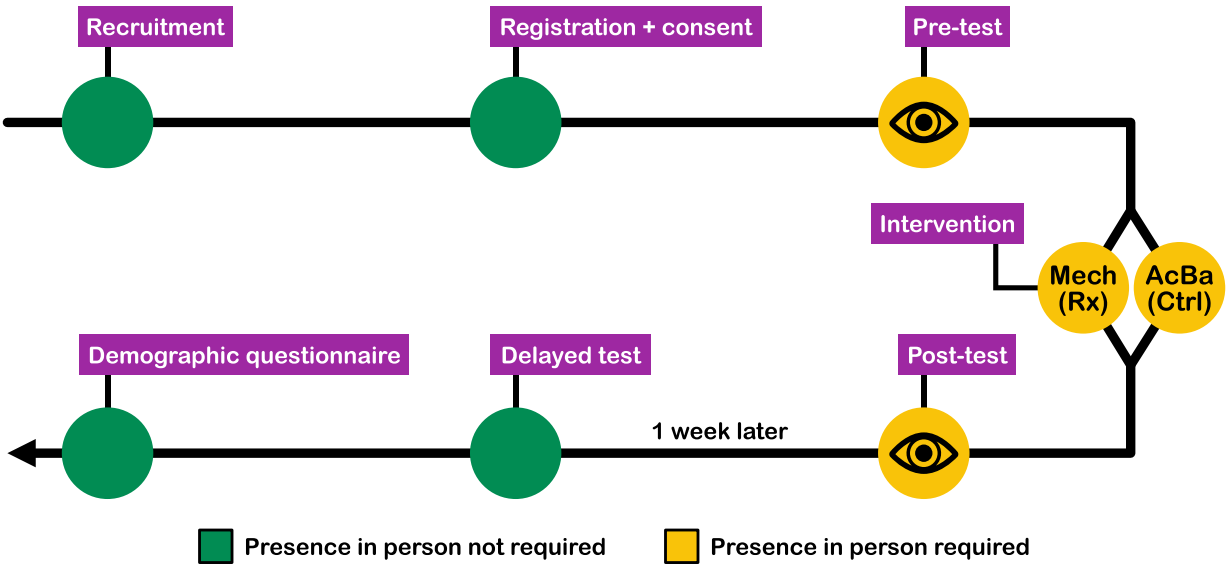


Figure 6*. The overall workflow of the study. The eye symbols in the pre- and post-test indicate that eye tracking was used at these steps. For the intervention, Mech represents the treatment module (OrgMech101), and AcBa represents the control module (Acid–Base Reactions).*

Data collection was broken up into two sessions. The first session took place in a 3.5 x 3-metre office. In this session, participants completed a pre-test, intervention, and post-test. The tests were formatted as a PDF file and conducted via *Notability*, a note-taking application for macOS, using a *Wacom Intuos Art Pen Tablet* and an *iMac* (21.5-inch, Late 2015). On the Wacom tablet, we programmed four buttons to undo/redo pen strokes and scroll up/down to simplify the process. Before starting the pre-test, participants went through a tutorial page where they could practice drawing an organic structure as often as they wanted before the pre-test to become more comfortable with the drawing tablet and the application (**Figure 7**).

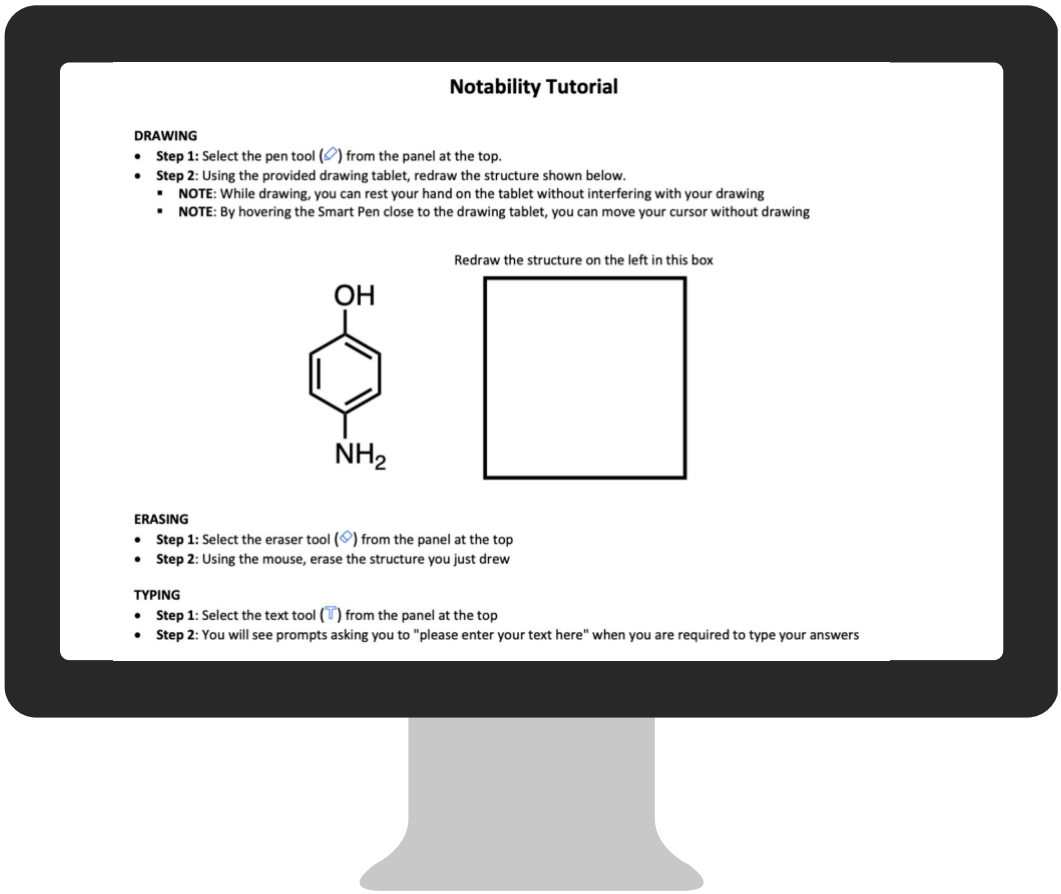


Figure 7*. The tutorial page that preceded the pre-test. Participants could redraw the shown structure as many times as they wanted until they became comfortable with the application and drawing tablet. When they felt comfortable, they proceeded to the EPF questions (found below the tutorial page).*

The pre-/post-test

The tests used to assess EPF fluency and reasoning ability initially contained ten questions (six EPF questions and four case comparison questions) that were reviewed, modified, and validated by two organic chemistry experts. We removed two EPF questions as we felt the tests were too long for participants and may introduce cognitive fatigue as a confounding variable. Q1 and Q2 focused on drawing electron-pushing arrows, where participants were given a sequence of mechanistic steps, and they had to draw where the EPF arrows would go for the reaction to proceed correctly. Q3 and Q4 focused on drawing the outcome of a mechanistic step, given the arrows (**Figure 8**). We opted to replace Q4 midway through the data collection process because it generated a ceiling effect. This effect occurred because participants achieved high scores for that question on the pre-test, hindering our ability to observe learning gains in the post-test. To limit confounding results due to the changed question, we excluded Q4 from our analysis. The questions on the test were a mix of familiar and unfamiliar questions, but the general idea was that participants should be able to answer all questions if they properly interpreted the symbolism. These questions were adapted from a previous study conducted by Carle, Visser and Flynn (2020) and Flynn and Featherstone (2017) and vary in difficulty to elicit various levels of cognitive load.

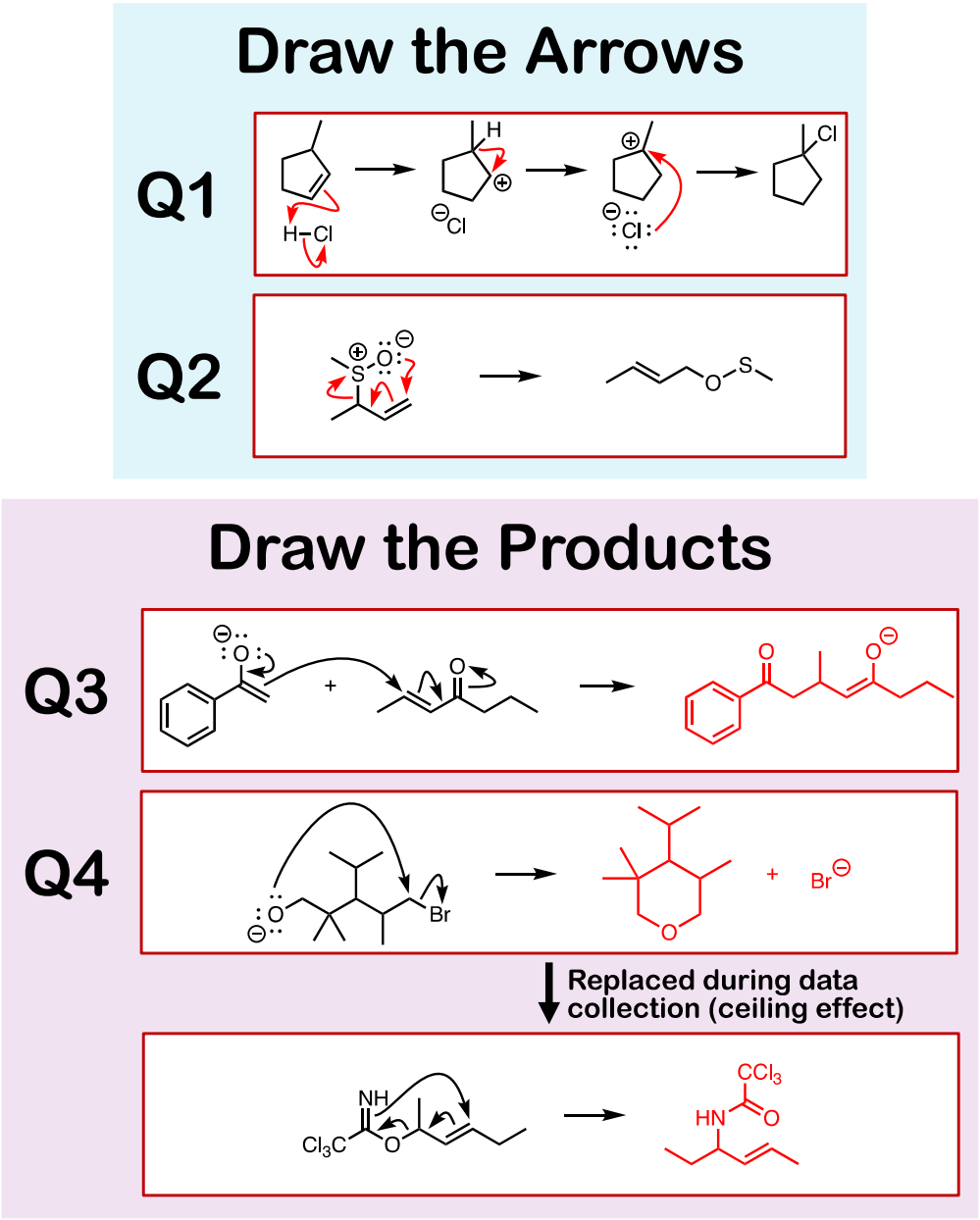


Figure 8*. EPF Questions in the pre-/post-test. Arrows or structures in red represent solutions to the problems and were what students had to draw. Q4 was replaced midway through data collection due to a ceiling effect.*

For Q5–Q8, participants were given a mechanistic step of a reaction that was drawn proceeding in two ways and were asked to predict which option was more favourable (**Figure 9**). We specifically chose to incorporate case comparison questions, as they allow students to engage in deeper reasoning by considering the various factors that affect the reactivity of organic molecules (Alfieri *et al.*, 2013). In several questions, we incorporated multiple factors that influence the stability/reactivity of each reaction to challenge participants' thinking and

have them consider synergistic or antagonistic effects they may have (Kuhn and Dean Jr., 2004; Hmelo-Silver *et al.*, 2007; Kuhn *et al.*, 2008). When explaining changes in reactions or properties, studies have demonstrated that students frequently rely on a single factor, so designing questions with multiple factors to consider can help challenge students and elicit deeper thinking (Talanquer, 2006).

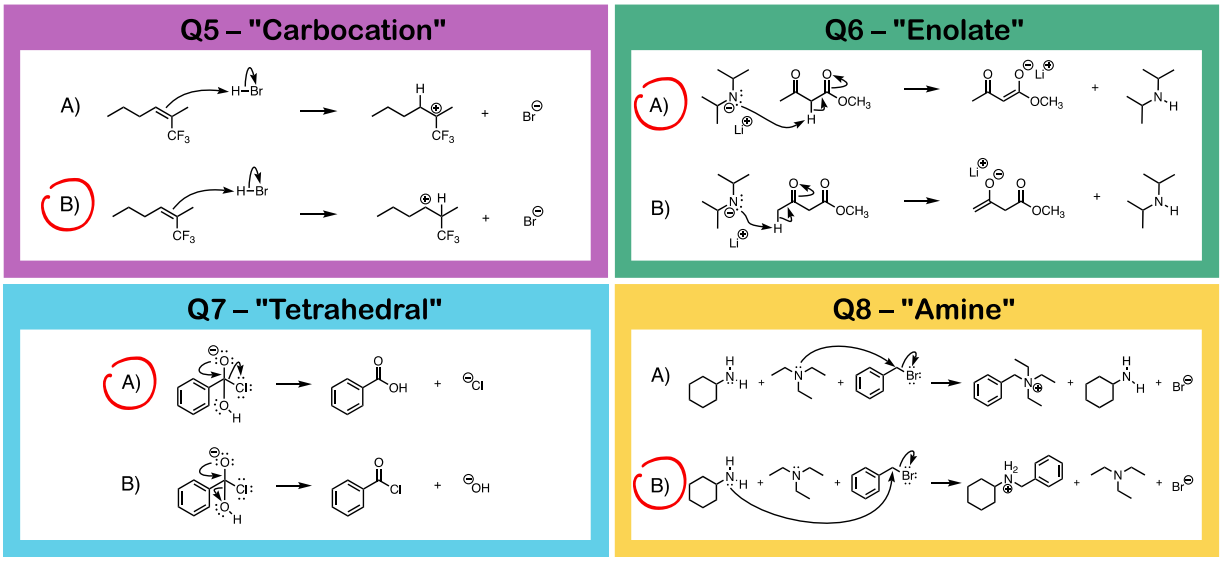


Figure 9*. Case comparison questions. Participants were asked to circle the more favourable option (circled in the figure) and explain their reasoning in detail in a textbox below each question.*

Q5 involves a hydrohalogenation reaction, where a hydrogen halide (in this case, HBr) is added to an alkene. Under normal conditions, the positive charge on the reaction intermediate is more stable on the most substituted carbon. To increase the difficulty of the question, we introduced an electron-withdrawing group (CF_3) to the alkene. Participants needed to determine which carbon in the alkene would connect with the hydrogen and which carbon would acquire an unstable positive charge. Participants would have to identify the effect of the degree of substitution (hyperconjugation) and the impact of the electron-withdrawing group (inductive effect) and make a decision based on the factor they think would predominate. In reality, these effects are impossible to weigh without experimental data or years of experience.

The reaction in Q6 (enolate) features a strong base called lithium diisopropylamide (LDA) and a molecule with two ketones (1,3-dicarbonyl). Unlike the previous question, either option can be favourable over the other. Option A produces what is known as the thermodynamic enolate, while option B yields the kinetic enolate. The kinetic enolate always

forms first and may also be the thermodynamic (i.e., more stable) enolate. When the kinetic enolate is not the most stable form of the product and the conditions allow for reaction reversibility, the thermodynamic enolate will eventually be formed. Factors that dictate the reversibility of the reaction include temperature, pressure, and solvent (Xie *et al.*, 2003). We specifically asked participants to predict which reaction would lead to the major equilibrium (i.e., more stable) product to differentiate between these outcomes. Participants needed to weigh out accessibility versus stability and demonstrate a proper understanding of the major equilibrium product.

The reaction in Q7 involves a phenomenon called a tetrahedral intermediate collapse. The starting material in this reaction becomes more stable when a leaving group is ejected. The two leaving group options are chloride (Cl^-) and hydroxide (OH^-). Participants needed to identify which ion was more likely to leave the molecule, resulting in the desired stabilization of the starting material. In this case, the chloride ion is the better leaving group due to its larger size, which can help distribute electron density better, effectively stabilizing the negative charge it carries.

The reaction in Q8 features an S_N2 substitution reaction with a secondary or tertiary amine (with the number of non-H atoms attached to the nitrogen dictating the degree). This reaction features a collision between a nucleophile (amine) and an electrophile (benzyl bromide). Nucleophiles are electron-rich molecules or atoms, while electrophiles are electron-deficient molecules or atoms. When a nucleophile and an electrophile collide, they interact in a way that allows a favourable chemical reaction to occur. The critical factor determining the reactivity of the two amines is their steric hindrance or overall bulkiness. More bulky amines are generally less likely to engage in this reaction since the neighbouring atoms (in this case, the three alkyl chains) near the nucleophilic site make it difficult for the amine to approach the electrophilic site of interest. As a result, the reaction pathway becomes unfavourable due to the higher activation energy required for the reaction. As such, option B (less bulky nucleophile) is more likely to proceed. Option A is designed to mislead students by presenting a product with characteristics like the carbocation from Q1 that might appear stabilizing. This feature can

potentially lead participants to select option A as the correct answer. In this situation, participants will have not considered the feasibility of the reaction mechanism itself.

Participants were given a textbox to explain their reasoning for each question and could either write out their answers using the drawing tablet or type out their answers using a keyboard. For the latter option, a digital keyboard was shown on the computer screen to limit the number of times participants would look down (which may have influenced pupil data). Participants were urged by the researcher (AY) to provide sufficient detail in their responses at the start of the study. A pK_a table was provided if a participant requested one. The researcher (AY) kept track of each participant's time to complete the pre-test and post-test.

Eye-tracking methodology and equipment

Out of the 36 participants, 34 met the eligibility criteria for using the eye tracker and willingly provided their consent for its use. The remaining two students participated in the study without using an eye tracker. The eye-tracker was incorporated to measure both the intrinsic and germane cognitive load of participants. The intrinsic load is affected by the questions themselves and the germane cognitive load is related to how effectively the content is processed (Zagermann *et al.*, 2016). We used the *Pupil Core*, a highly versatile and precise video-based eye tracker. The *Pupil Core* is an eye-tracking instrument that is comprised of a wearable eye-tracking headset that includes a camera for each pupil and a mounted world camera to capture participants' fields of view. This setup allows the eye-tracking software to track which regions of the tests participants focus on while pupil diameter changes can simultaneously be tracked with the eye cameras. The headset itself connects to a computer via a USB connection. Data were captured via *Pupil Capture*, the open-source software designed by the same manufacturer. Before data collection, we tested the validity of using pupil diameter to measure cognitive load by performing a task developed by Klinger (2010). In qualitative research, validity is the accuracy with which an instrument measures what is intended to be measured (Golafshani, 2015). The validity test was conducted to assess how accurate our eye tracker would be at measuring cognitive load. Details about this test can be found in *Appendix C: Pupil Core validity test*.

We set luminescence in the room to 25% to minimize ambience, thereby reducing pupil constriction and allowing us to more easily measure changes in pupil diameter. In addition, the room the study took place in contained no windows, so ambience levels stayed consistent throughout the study. To reduce contrast, a white wall was positioned behind the monitor used by the participants during the test (**Figure 10**). The researcher (AY) sat approximately 2.5 metres behind each participant so they would not be a distraction. At the beginning of each test, the cameras were manually adjusted so that each pupil was centred, and a 5-point calibration was performed to ensure maximal data precision and accuracy. Calibration was restarted if a participant recorded more than three degrees of accuracy error. The eye-tracker recorded data for the entirety of the pre- and post-tests.

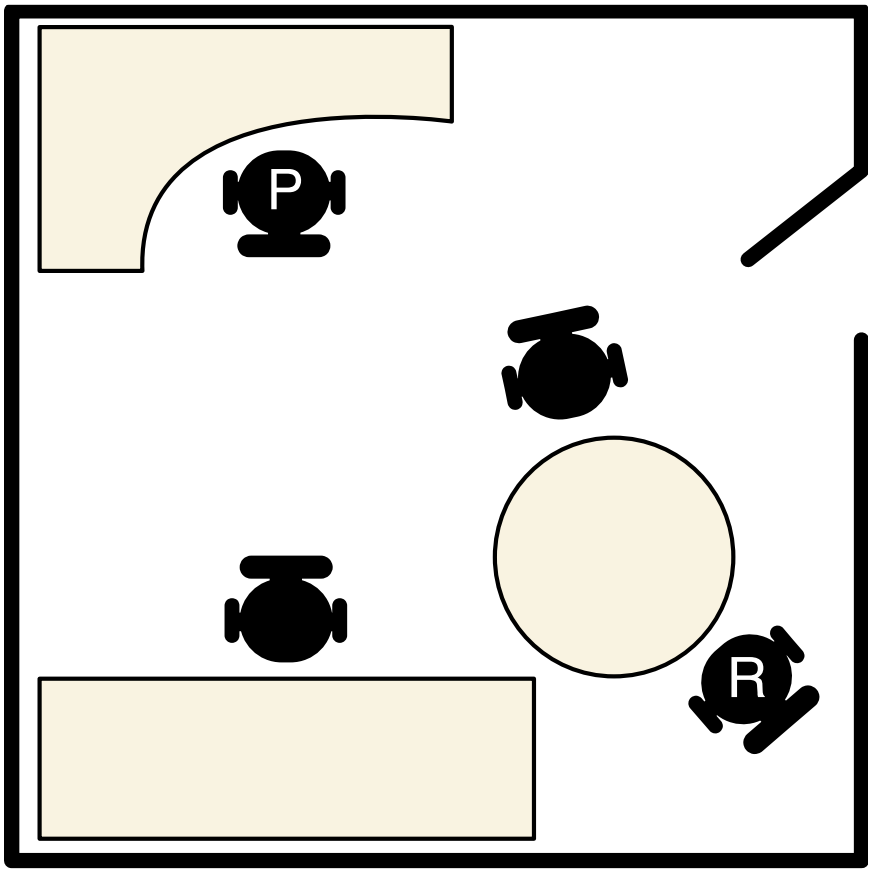


Figure 10*. The layout of the room in which data collection took place. The area where participants sat is denoted by the letter "P," and the area where the researcher (AY) sat is indicated with an "R."*

The *OrgMech101* and *Acid–Base Reaction* modules

After completing the pre-test, participants were randomly assigned into one of two groups: *Acid–Base* (control) and *OrgMech101* (treatment). **Figure 11** previews the overall composition of each module. We selected the *Acid–Base Reactions* module for the control group because it contains minimal elements of the EPF. However, the module is still likely to teach and reinforce essential concepts that participants would be tested on in their courses, making their participation in the study meaningful to them.

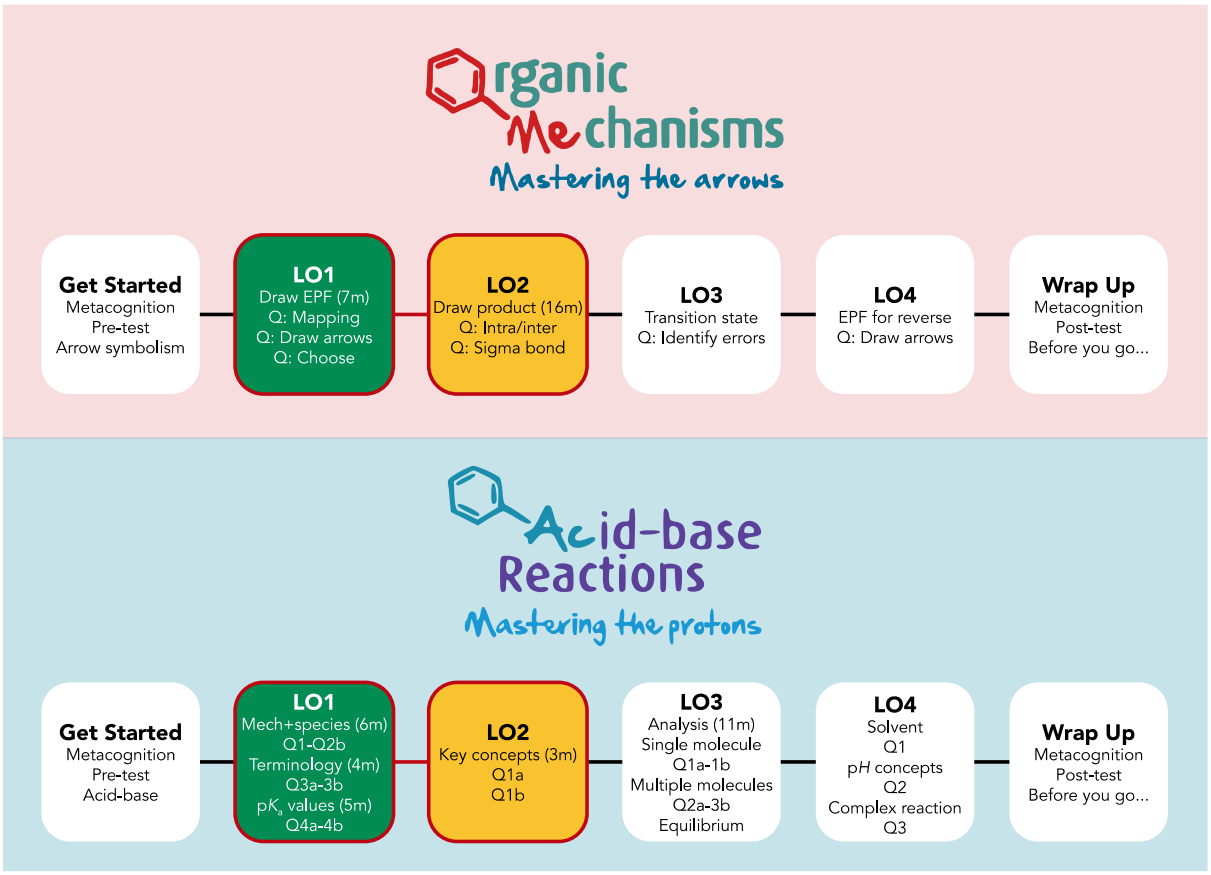


Figure 11*. Overview of the module that students in each group completed. The coloured boxes represent the portions of the module that participants were required to complete. Practice question sections are denoted by a "Q." Participants from both groups watched the "Analysis Methods" video in LO3 of the Acid–Base module after they completed their respective modules.*

The treatment group completed the first two learning outcomes (LOs) of the *OrgMech101* module, which emphasized learning how to correctly draw electron-pushing arrows and the products of a reaction given the arrows, as well as interpreting reaction symbolism. The section with the first LO consisted of a six-minute video about drawing the EPF arrows of a reaction mechanism, followed by three sections of problems where participants could practice mapping atoms, drawing arrows, and selecting reactions with correctly drawn arrows. The second LO consisted of a 16-minute video about drawing the product of a reaction given the arrows, followed by practice questions of the same subject. For the first two LOs, the intention was to use unfamiliar reactions, so participants could focus on interpreting the symbolism rather than recalling memorized reactions. After completing the two LOs, participants watched a ten-minute video from the *Acid–Base* module, which taught students how stability is influenced by the following chemical factors: electronegativity, atom size, resonance, hybridization, induction, charge, and solvent.

Participants in the control group completed the first two LOs of the *Acid–Base Reactions* module, which included identifying acids and bases, applying acid–base terminology (e.g., protonating or deprotonating a compound and drawing a conjugate acid/base), and associating pK_a with acid/base strength and stability. The *Acid–Base* module had more sections for practice questions than the *OrgMech101* module; however, the videos were much shorter, offsetting the effect. Participants in the *Acid–Base* module watched a three-minute video about EPF arrows and mechanisms towards the beginning of the module. Including this brief EPF practice provides a more authentic representation of students' classroom experience with the EPF. Yet, it is not explicit EPF practice like that which the *OrgMech101* group goes through—we were interested in EPF fluency in this study, and exposure to EPF for three minutes was unlikely to provide sufficient fluency to participants.

After completing the first two LOs, the *Acid–Base* group participants watched the same video as the treatment group about chemical factors influencing stability. The video was incorporated in both groups because we believed it might influence students' answers to the case comparison questions, and we did not want to provide one group with an advantage. Though we would have liked to incorporate LO3 and LO4 of each module, we did not want to

prolong the intervention for students, as studies have shown that task performance decreases over time (Guo *et al.*, 2016; Reteig *et al.*, 2019). Participants were given as much time as they liked to go through the module and practice questions, and they were told to notify the researcher (AY) when they completed LO2.

Each participant took the module individually (no other participants were present). The eye tracker was not used during the intervention; however, participants' screens were shared and recorded via Zoom. This way, the researcher (AY) could reference how long participants spent on the intervention and how many practice questions they went through. the interventions were designed such that they would take approximately 35 minutes.

Delayed post-test and demographic questionnaire

We emailed students a PDF version of the test completed prior one week later. Participants either printed a hard copy of the PDF and handwrote their answers or completed it on their tablets. Participants were asked to time themselves while completing the test and report it to the researcher (AY) when submitting. The participants submitted their delayed test via email by scanning the hard copies or emailing the completed PDF if they used a tablet.

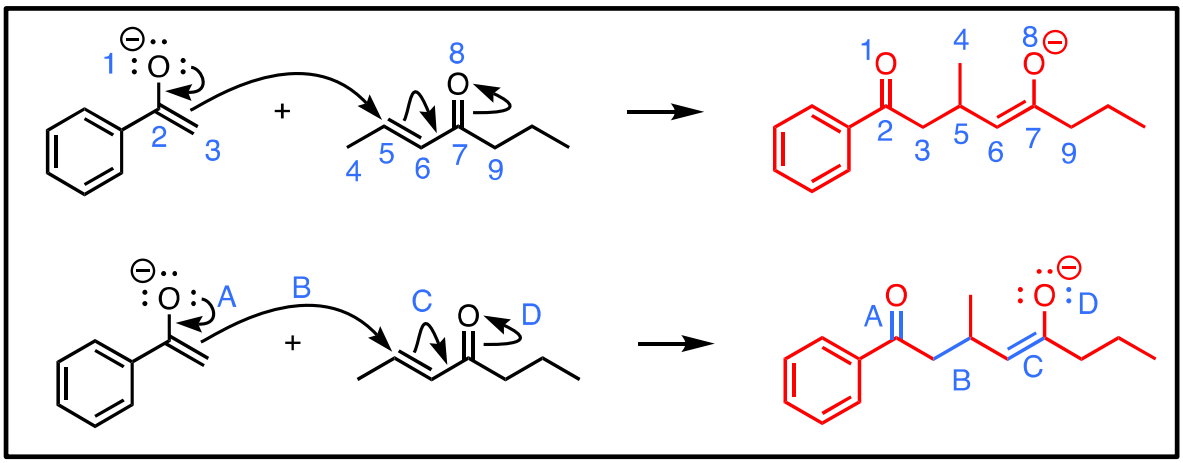
We decided to administer the delayed post-test one week later considering the results of Ebbinghaus's Forgetting Curve, which showed that newly learned information is forgotten over time (Ebbinghaus, 1885; Murre & Dros, 2015). A time interval was selected that allowed for a long enough gap between the post-test and the delayed post-test but not too long that most information is forgotten. We did not want to extend the time between the post-test and the delayed post-test for too long, as that would increase the risk of confounding variables. We originally wanted to analyze the delayed post-test data to measure knowledge retention; however, data collection spanned the entire semester, so although 35/36 participants completed a delayed post-test, we did not analyze the data.

After completing the delayed post-test, participants were asked to complete a demographic questionnaire. We collected this information because it is critical to determine if a particular group benefits from this treatment and if one is marginalized. Participants completed the delayed post-test electronically and at their leisure. We did not use eye-tracking technology

for the delayed post-test, as its primary purpose was to measure students' knowledge retention.

EPF question and eye tracking data analysis

Responses to the EPF questions (Q1–Q3) were evaluated similarly to our previous work (Carle *et al.*, 2020). One point was awarded for each correct arrow for the draw-the-arrow questions (Q1 and Q2). A correctly drawn arrow is defined as one that starts at the proper atom/bond and ends at the correct atom/bond. Looking at **Figure 8**, Q1 has four arrows and Q2 has three arrows, so a maximum of four and three marks can be awarded, respectively. Although we considered arrows drawn from atoms to be incorrect, as seen by the work of Carle *et al.* (2020), we elected to award full points if we encountered such a case, as there is an inconsistency in the arrow-drawing notation in organic chemistry, as seen by recent publications where authors drew arrows from atoms in their mechanisms (Hurtado-Rodríguez *et al.*, 2022; Liu *et al.*, 2022). For the draw-the-product question (Q3), one point was awarded for correctly making/breaking a bond and one extra point was awarded for correctly drawing the final product (**Figure 12**), as there were some instances where participants would formulate the correct bonds but miscount the number of carbons in the product. The weighting of each question was adjusted so each of the three questions would be worth four marks. This adjustment was made so that one question would not impact the final score more than the others.



Action	Bond involved	Arrows	Type of bond
Make	O1—C2	A	π
Break	C2—C3	B	π
Make	C3—C5	B	σ
Break	C5—C6	C	π
Make	C6—C7	C	π
Break	C7—C8	D	π

Figure 12*. Marking scheme for the Q3 draw-the-product question. One point was awarded for correctly performing each action. One extra point was awarded for drawing the correct product.*

Participants' pre- and post-test scores were calculated along with their normalized learning gains (NLG), highlighted in **Equation 1**. If participants obtained a lower score on the post-test than the pre-test, their NLG was automatically assigned a score of zero to indicate no learning gains. We did this because the NLG equation traditionally would yield a negative value if the post-test score was lower than the pre-test score; however, we felt it was more appropriate to eliminate any possibility of negative learning gains as they skew the data (Coletta and Steinert, 2020). Additionally, the idea of "negative learning gains" is misleading as it implies a loss of learning, which is unlikely in a learning environment.

$$\text{Normalized learning gain} = \frac{\text{post} - \text{pre}}{100 - \text{pre}}$$

Equation 1*. The equation for normalized learning gain. If a participant's post-test score was lower than their pre-test score, their normalized learning gain was automatically assigned a value of zero to indicate no learning gain.*

Eye-tracking data were exported using *Pupil Player*. The program contains a built-in blink detector that generated a list of all recognized blink intervals in a separate CSV file. The blink detector threshold is adjustable, which we modified to the most sensitive setting to eliminate any data that may be inaccurate (filter length = 0.10 secs, onset confidence threshold = 0.05, offset confidence threshold = 0.05). We were not concerned with eliminating too much data because the sampling rate of the eye-tracker was 250 Hz (i.e., the eye-tracker generated approximately 250 readings per second). After these settings were applied, the eye-tracking data were exported as a CSV file.

The data were further cleaned using a custom-written Python script (*Appendix D: Bach Filter code used to clean and analyze pupil data*). We used several packages, including *pandas*, *matplotlib*, *numpy* and *csv*. Any data points with confidence readings below 0.99 were removed. In addition, any data points generated by the blink detector were omitted. As left and right eye data were collected separately, we combined the two datasets using an interpolation function of Python. After that, a baseline value was subtracted from all data points to normalize them, making it more appropriate to compare pupil data across participants. The baseline pupil diameter was calculated as the average of the first 20 readings (approximately 0.1 seconds) after participants navigated from the dark *Pupil Capture* recording screen to the light *Notability* screen containing the pre-/post-test. The change in contrast produced a large decrease in participants' pupil diameter, allowing us to identify a consistent point to use as a baseline measure. After considering several other options, this method was the most successful and accurate approach for normalizing the data.

We cycled through each participant's recordings and determined the frame number at which they began Q1 and completed Q3. Using this information, an average pupil diameter in that region was generated and corrected using the obtained baseline pupil diameter. A maximum pupil diameter value was also generated; however, we cross-referenced the timestamp at which the maximum diameter occurred with the world frame recording to confirm that the maximum value was obtained while participants were looking at the question/molecules or drawing and not due to an extraneous factor (e.g., looking away from the test). From there, we were able to generate plots to assess the relationship between raw

pre- and post-test scores and baseline-corrected average and maximum pupil diameters to evaluate the correlation between EPF fluency and cognitive load. We also plotted the normalized learning gain versus a pre-test-corrected post-test score (post-test score – pre-test score) to measure the correlation between the change in EPF fluency and the change in cognitive load.

Case comparison questions data analysis

We individually reviewed all responses for the case comparison questions and assigned specific chemical concepts to reflect participants' arguments. We then looked at how these concepts were connected to determine what mode of reasoning participants used. These concepts were either pieces of evidence or causal links. Once responses were analyzed, a codebook was created inductively with all concepts included. After this, a second round of analysis/coding was done to ensure arguments were consistently coded. Three other researchers then reviewed the codebook and further eliminated and combined codes to refine the codebook, which is shown in **Table 2**. *Appendix E: Modes of reasoning coding instructions* highlights the coding process in more detail.

Inter-rater reliability analysis was conducted for the modes of reasoning between two researchers with 60 random responses from the combined 288 responses available. Inter-rater reliability analysis is often performed in qualitative research to ensure the data are not dependent on the subjective judgement of one researcher. Assessing the reliability of data analysis methods is important in qualitative research, as it provides insight into the reproducibility of the obtained data under the same conditions (Golafshani, 2015). By involving additional raters and measuring the degree of agreement among them, the method provides a more robust analysis of the data. Yet inter-rater reliability analysis should not be performed when the data is sensitive, such as when unique perspectives or experiences are provided. For example, it would be inappropriate for a researcher with limited knowledge of Indigenous perspectives to perform inter-rater reliability analysis for data from an Indigenous study. In the first trial, researchers achieved 76.7% agreement and a Cohen's kappa value of $\kappa = 0.58$, constituting moderate agreement (McHugh, 2012). The two researchers discussed the

discrepancies until agreement was reached and independently performed another round of inter-rater reliability with an additional 60 responses and obtained 90% agreement and κ = 0.83, which constitutes an almost perfect agreement (McHugh, 2012).

To provide insight into how participants' arguments differ in the *OrgMech101* group compared to the *Acid–Base* group, we analyzed the frequency of chemical concepts used and the number of times two concepts were used in the same argument. This analysis would allow us to generate a *Gephi* network to visualize how participants' arguments changed from the pre-test and post-test. A *Gephi* network is a diagram created using the *Gephi* software. In this case, networks consist of nodes representing individual chemical factors and edges representing the relationships between those factors.

***Table 2**. Codebook for arguments used in the tests. Examples used represent actual responses from participants that were coded for that particular concept by the researcher (AY).*

Code	Definition	Example
Electronegativity	Discuss the electronegativity of various atoms and functional groups	*"Option A is between two oxygens which are electronegative."*
Leaving group ability	Discuss the leaving group ability of multiple atoms and functional groups	*"Chlorine is more likely to be the leaving group (take electrons with it)."*
Atom size	Discuss the size of atoms and how it affects various chemical properties	*"Chlorine is below oxygen in the periodic table and is therefore a larger atom."*
Induction	Discuss the inductive effect of nearby atoms or groups on the stability of a charged species	*"CF_3 group has a lot of electron density and pulls electron density from the carbon."*
Resonance	Discuss how stability is affected by resonance or the number of resonance structures	*"There are more resonance structures that can be drawn for this compound."*
Delocalization	Discuss the movement of electrons across atoms in a molecule	*"The negative charge can be distributed among the two O atoms."*
π bond conditions	Discuss substitution effects on π bonds or preferred location of a π bond in a molecule	*"The base is bulky that means the Hofmann Product is favoured."*
Sterics	Discuss molecule accessibility or reaction hindrance caused by steric interactions.	*"Given the branching of the molecule, there would be more steric clash."*
Charge stability	Discuss a positive or negative charge being stabilized	*"F will not stabilize the positive charge/carbocation effectively."*
Stability (thermodynamic)	Discuss the stability of reactants, intermediates, or products	*"Since the products of B are more stable, it is the more likely outcome."*
Stability (kinetic)	Discuss how activation energy changes or references transition state stabilization	*"The E_a to form this carbocation is higher than what we saw in A."*

Reactivity	Discuss the reactivity of different molecules or overall reaction speed	*"A is more likely as the tertiary carbon with the positive charge is more reactive."*
Nucleophile strength	Discuss the relative strength of different nucleophiles in a reaction	*"The Benzene-N group is more likely to act as a nucleophile."*
Acid strength	Discuss the strength of acids	*"H_2O, as its conjugate acid, is a weak acid."*
Base strength	Discuss the strength of bases	*"The conjugate base of the acid in A is the weaker base."*
pK_a	Discuss the acidity or basicity of various compounds based on their pK_a values	*"Conjugate acid pK_as are –8 for Cl^- and 15 for OH^-"*
Other	Miscellaneous arguments (e.g., hybridization, bond strength, orientation, octet violation)	

We conducted several analyses using the statistics panel of *Gephi* for each network. The *average weighted degree* measures the average strength of links for each node, considering the weight of each connection. The measure is calculated by adding the weights of all the edges connected to a node and dividing by the total number of edges. The average node size follows a similar principle but looks at the node's size rather than the connection's weight. *Graph density* measures how many links exist in a network relative to the maximum number of possible connections. The maximum value of the graph density measure is 1, which indicates that all possible edges exist in the network. A density of 0 means that there are no edges in the network. *Modularity* measures the degree to which a network can be divided into non-overlapping groups (Blondel *et al.*, 2008). This measure is typically used to identify clusters of nodes in a network that are densely connected to each other but have few connections to nodes in other communities.

Lastly, we analyzed the correctness levels of participants' responses according to **Table 3**. One of three codes were assigned to each participant's response: correct, partially correct, or incorrect. More detail and examples of how some responses were coded can be found in *Appendix F: Correctness of response coding guidelines*. We performed inter-rater reliability analysis with 60 questions to ensure the responses were correctly coded. We obtained a percent agreement of 53/66 (88%) and $\kappa = 0.82$, indicating almost perfect agreement.

***Table 3**. Codebook for the correctness of an argument. An example is provided using participant responses for Q5 (carbocation).*

Code	Definition	Example
Correct	Arguments used in response are scientifically accurate	*"Fluorine atoms on carbon withdraw electron density, causing a partial positive charge. This polarization destabilizes reaction A due to the proximity of two positive charges."*
Partially Correct	Some (or not all) arguments used in response are scientifically accurate	*"F atoms stabilize the hydrogen via induction. B also is better because the positive charge is further away from the electronegative atom."*
Incorrect	Incorrect claim made or arguments are scientifically inaccurate	*"I think A is correct because CF_3 is very electronegative and can better stabilize the positive charge."*

RESULTS AND DISCUSSION

Increased EPF fluency is associated with a lower pupil diameter

Eye-tracking results showed that participants in the *OrgMech101* group exhibited significant decreases in pupil diameter while participants in the *Acid–Base* group did not. **Figure 13** summarizes the overall distribution of pre- and post-test scores for participants in both groups. Participants in each group had similar levels of chemistry knowledge going into the study since they obtained similar averages in the pre-test. There was no apparent relationship between study participation date and exercise scores when visually inspected. September participants (*N* = 11) scored 68% and 74% on the EPF portion of the pre- and post-test. The scores increased for October participants (*N* = 10; 76% and 85%) before decreasing in November (*N* = 13; 67% and 75%). Only two participants completed the study in December, both of whom performed well on the test (98% pre- and post-test scores). Both groups spent roughly an equal amount of time completing the pre and post-test, while participants in the *OrgMech101* spent approximately ten additional minutes completing the intervention. While differences in learning between each group may have been simply due to more time engaged in studying the treatment module, we did not want to restrict the time at which all participants complete the intervention, as they all work at different paces. Additionally, although the treatment group spent more time in the intervention, they were at greater risk of cognitive fatigue. On average, participants in the *OrgMech101* group completed the pre- and post-test in 34.5 (± 8.68) and 19.9 (± 6.34) minutes. In comparison, the *Acid–Base* group finished the pre- and post-test in 37.4 (± 18.9) and 19.8 (± 6.53) minutes, on average, respectively. The average intervention time for participants in the *OrgMech101* group was 49.2 (± 9.70) min and 40.5 (± 8.95) min for the *Acid–Base* group. Overall, the total time spent between the pre-test, intervention, and post-test for the *OrgMech101* and *Acid–Base* groups did not differ by much (84 vs 78 minutes respectively). Participants in both groups scored higher on the post-test, which is not uncommon for pre/post designs in education research (Morgulis *et al.*, 2012; Stensaker *et al.*, 2017) and may be attributed to participants gaining more confidence as they repeated the test.

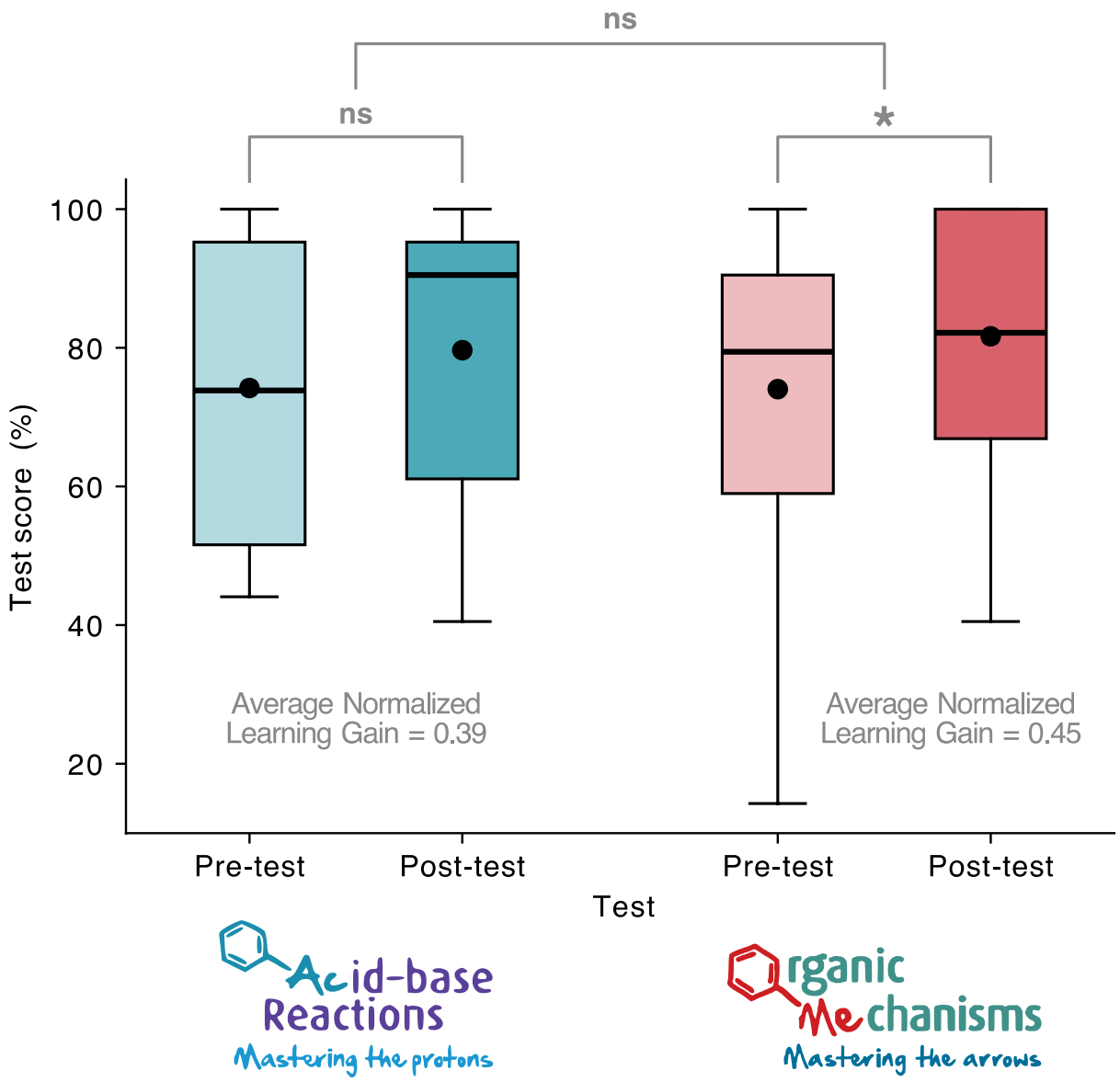


Figure 13*. EPF pre- and post-test scores for Acid–Base and OrgMech101 participants. The median for each sample is indicated with a horizontal black line, while the mean is marked with a black circle. An asterisk denotes a significant difference, while ns denotes no significant difference. Average normalized learning gains are also provided for each group. N = 17 for the Acid–Base group and N = 19 for the OrgMech101 group.*

To determine if there was a significant difference in scores pre-test to post-test for each group, we conducted a Wilcoxon signed-rank test, the non-parametric equivalent of a paired-samples *t*-test (since the data were not normally distributed). Learning gains (i.e., post-test score vs pre-test score) for participants in the *Acid–Base* group were not found to be statistically significant at $\alpha = 0.05$ ($Z = 1.531$, $p = 0.126$), while learning gains for the

OrgMech101 group were found to be statistically significant ($Z = 2.098$, $p = 0.036$). These results are expected because participants in the *OrgMech101* group primarily learned about the EPF, while participants in the *Acid–Base* group did not. The NLG of the *OrgMech101* group was slightly larger than the *Acid–Base* group; however, the difference was statistically insignificant according to a Mann-Whitney U Test ($U = 108$, $p = 0.625$). While this difference was insignificant, previous studies assessing the effectiveness of the *OrgMech101* module in teaching the EPF produced significant results with a much larger sample size and lower variability (Carle *et al.*, 2020).

On average, participants in the *OrgMech101* group experienced a 56% reduction in pupil diameter change during the post-test (**Figure 14**). In contrast, participants in the *Acid–Base* group had their average pupil diameter stay relatively consistent (4% decrease in the post-test). A paired-samples *t*-test shows that the drop in diameter is significant for the *OrgMech101* group ($t(15) = 2.242$, $p = 0.020$). The effect size for the difference between the tests was calculated using Hedges' correction, resulting in a value of 0.532, which is considered a medium effect. A paired-samples *t*-test conducted for the *Acid–Base* group found no significant differences between the pre and post-test scores ($t(13) = 0.051$, $p = 0.480$). These significant differences are not due to differences in the average post-test pupil diameter changes (0.30 mm for *OrgMech101* and 0.24 mm for *Acid–Base* group), but they are present because of the major differences in the pre-test average pupil diameter changes (0.67 mm for *OrgMech101* and 0.24 for *Acid–Base* group). In other words, the average pre-test diameter of participants in the *OrgMech101* group was much higher in the pre-test and, therefore, was more likely to provide decreased results compared to the *Acid–Base* group. An ANCOVA with pre-test diameter changes as a covariate was conducted to address these findings. As expected, the statistical test determined that the difference in diameters when controlling for the pre-test was not statistically significant ($F = 0.187$, $p = 0.669$, $N = 30$). Given the limited sample size for the statistical tests, we recommend any interpretations be exercised with caution. In any case, these statistical results provide an overview of the data trends.

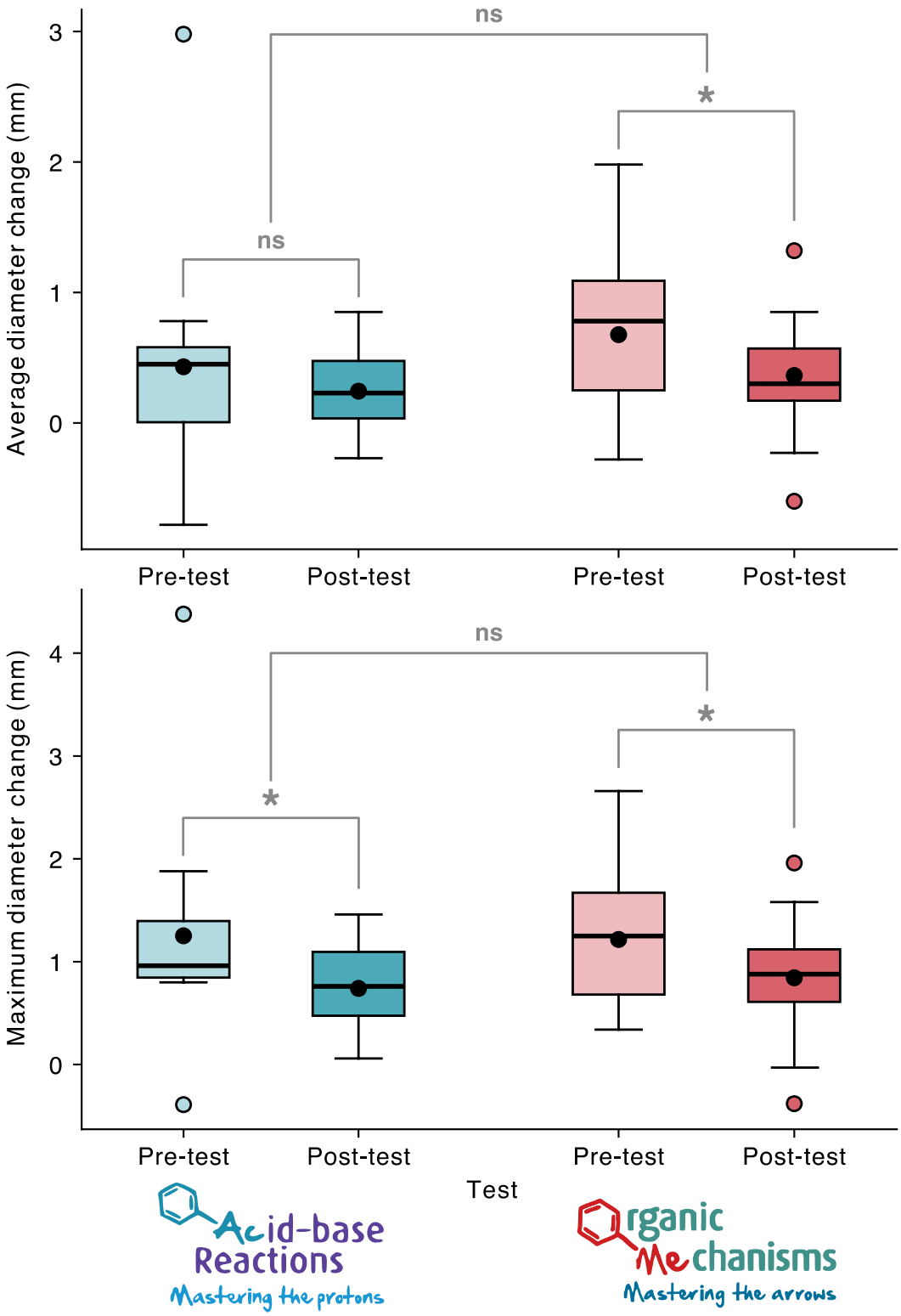


Figure 14. *Boxplots of average (top) and maximum pupil diameters (bottom) for the Acid–Base and OrgMech101 groups. Medians are represented with a horizontal black line, and means are indicated with a black circle. An asterisk denotes a significant difference, and ns denotes no significant difference. N = 17 for the Acid–Base group and N = 19 for the OrgMech101 group.*

The mean maximum diameter change decreased from 1.21 mm to 0.84 mm for the *OrgMech101* group, while the mean maximum pupil diameter change decreased from 1.03 mm to 0.74 mm for the *Acid–Base* group (**Figure 14**). These changes correspond to decreases of 31% and 28%, respectively. For context, Kahneman and Beatty (1966) observed pupil dilation changes of 0.5 mm in a digit span task when participants were asked to remember and recite a 7-digit sequence compared to a 3-digit sequence. The difference in maximum pupil diameter changes was found to be significant with moderate effect sizes for the *OrgMech101* group ($t(16) = 2.624$, $p = 0.009$, Hedges' $g = 0.606$) and the *Acid–Base* group as well ($t(13) = 1.991$, $p = 0.034$, Hedges' $g = 0.501$). Based on an ANCOVA test conducted, the differences in *Acid–Base* and *OrgMech101* maximum diameter changes from pre-test to post-test were not significant with pre-test scores as a covariate ($F = 0.026$, $p = 0.872$, $N = 31$). The decreases in maximum pupil diameter change may indicate that participants in both groups exhibited lower cognitive load in the post-test compared to the pre-test. The reduction may also be due to an external factor, such as decreased stress.

We considered which types of cognitive load (intrinsic, extraneous, germane) may have been influenced during the study. Intrinsic cognitive load is unlikely to have been affected, as it is a factor that cannot be modified by the learner (Raker *et al.*, 2013). Participants' cognitive load may have decreased between the pre-test and post-test due to their increased familiarity with the questions and task (Luten, 2002). However, the differences in average pupil diameter decreases between the *Acid–Base* group and *OrgMech101* group are large enough to rule out this possibility as a significant contributing factor. Perhaps the *OrgMech101* module minimized extraneous cognitive load, as participants would no longer view the symbols they were learning as complex. Extraneous cognitive load is affected by a phenomenon known as element interactivity, where an element is "*anything that needs to be or has been learned, such as a concept or procedure*" (Sweller, 2010). A task with low element interactivity allows individuals to process individual elements independently. For example, learning the chemical symbol of copper does not require knowledge of the chemical symbol of iron. Therefore, learning chemical symbols is a task of low element interactivity, which may lessen extraneous cognitive load (Sweller, 2010). Upon completing the *OrgMech101* module, participants may have been

able to shift their attention from viewing the EPF arrows as a whole collection to viewing each arrow as a separate entity. This attention redirection can reduce element interactivity and extraneous cognitive load (Sweller, 2010). We can also consider the possibility that the enhancement in germane cognitive load resulted in the significant pupil diameter drop, indicating that participants became more efficient in processing the relevant information. The EPF usage strategies shown in the module (such as mapping) likely resulted in optimization of germane cognitive load, thereby reducing the overall cognitive load (as seen by the pupil diameter decreases).

A scatterplot of all participants' pre-and post-test scores versus their average pupil diameter change produced a negative correlation in **Figure 15**. We combined the data from both groups because the assumption linking cognition with pupil diameter means that participants who scored low on the pre- or post-test should exhibit the same magnitude of pupil diameter change, regardless of their assigned group. Spearman's rank correlation was computed to assess the relationship between pre-test EPF score and average pupil diameter change. The two variables had a negative, significant correlation, $\rho(31) = 0.322$, $p = 0.039$, which is considered a weak relationship (Dancey and Reidy, 2007). The correlation indicates that participants who scored higher on the pre-test were more likely to have a lower average pupil diameter change. Since pupil dilations are associated with cognitive processing, we can associate the smaller average diameter with less mental expenditure (Kahneman and Beatty, 1966).

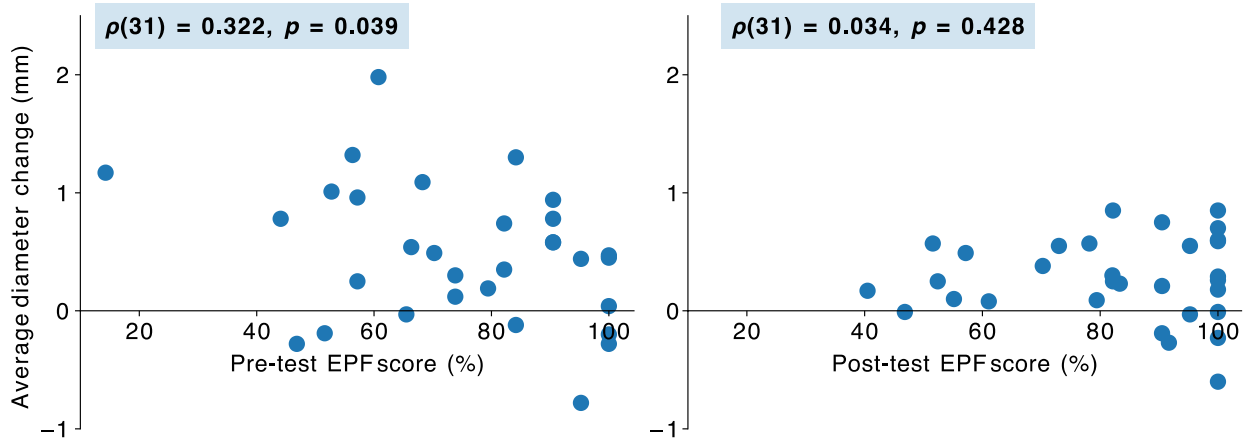


Figure 15*. Scatterplot of raw test scores versus average diameter change for the pre- and post-tests (baseline corrected).*

The relationship between post-test EPF score and average pupil diameter change (**Figure 15**) was found to be weaker, as indicated by Spearman's rank correlation ($\rho(31) = 0.034$, $p = 0.428$). Since there is lower variability in average pupil diameter changes in the post-test, there is a higher threshold to reach significance (Hinde *et al.*, 1993). Although the minimum average pupil diameter change threshold remained relatively stable from the pre-test to the post-test, the maximum threshold dramatically differed. In the post-test, none of the participants' average pupil diameter change exceeded 1 mm, whereas the data of six participants in the pre-test exceeded that value. On average, participants also scored higher on the post-test, and nine participants obtained a perfect score on the post-test compared to five on the pre-test. This ceiling effect limits the ability of the instrument to measure variability in post-test scores.

Similar to the average pupil diameter change, the scatterplots in **Figure 16** show a negative relationship between participants' pre- and post-test scores versus their maximum pupil diameter change. Spearman's rank correlation was computed to assess the relationship between pre-test EPF score and maximum pupil diameter change. The two variables had a moderate negative correlation, $\rho(31) = 0.419$, $p = 0.009$. The post-test, once again, showed no significant relationship ($\rho(32) = 0.109$, $p = 0.276$). Similarly to the post-test results in **Figure 15**, the combination of participants scoring higher and increased perfect scores lowered the overall variability and made assessing a significant relationship difficult (Wang *et al.*, 2008).

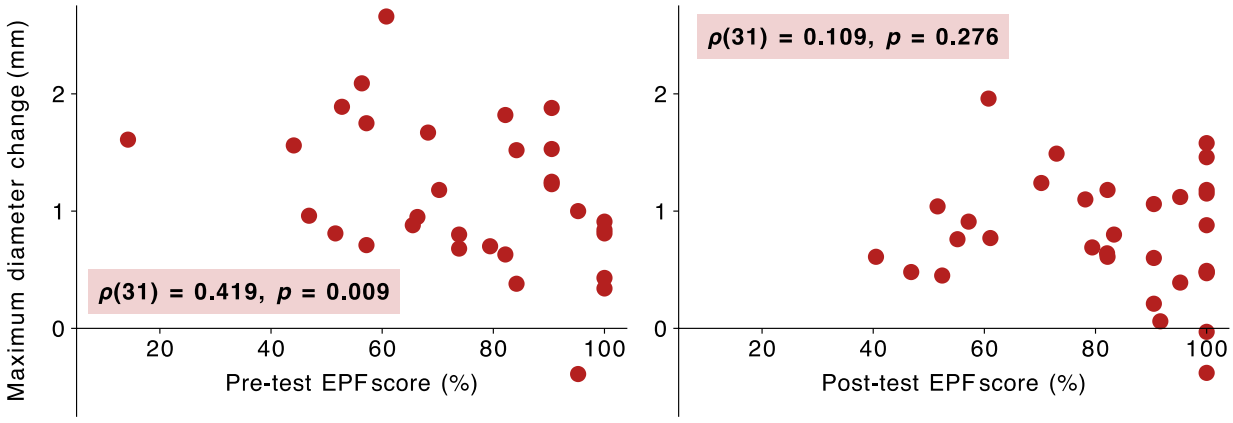


***Figure 16**. Raw test scores pre- and post and maximum pupil diameter change (baseline corrected).*

For a more comprehensive analysis and to reduce the contributions of variability and the ceiling effect in the data, we also measured the relationship between the change in test scores (via normalized learning gain) and the differences in average and maximum pupil diameter changes (average/maximum post diameter – average/maximum pre diameter). **Figure 17** highlights these results. A large pre-test-corrected average or maximum pupil diameter change indicates that participants' pupil diameter was higher in the post-test than the pre-test, which may mean an increased cognitive load. A Spearman's rank correlation showed a moderate negative relationship between the two variables, $\rho(25) = 0.500$, $p = 0.005$. The relationship indicates that those with higher EPF learning gains exhibited a greater decrease in their average pupil diameter in the post-test compared to the pre-test. In the post-test, participants with low learning gains exhibited a smaller reduction in their average pupil diameter. In other words, participants who did not show much improvement in EPF comprehension were more likely to have a larger pupil diameter in the post-test than in the pre-test.

Spearman's rank correlation was computed to assess the relationship between NLG and the pre-test-corrected maximum pupil diameter change (**Figure 17**). Although the results showed a stronger relationship than that of **Figure 15** and **Figure 16**, the relationship was still insignificant at $\alpha = 0.05$, $\rho(24) = 0.139$, $p = 0.258$. Unlike the average pupil diameter change, which considers tens of thousands of data points, a maximum pupil diameter change value is based on a single data point. Therefore, the measure is sensitive and more likely to be influenced by external factors like stress and emotion. As such, these external factors may have been the reason for the pupil diameter decrease.

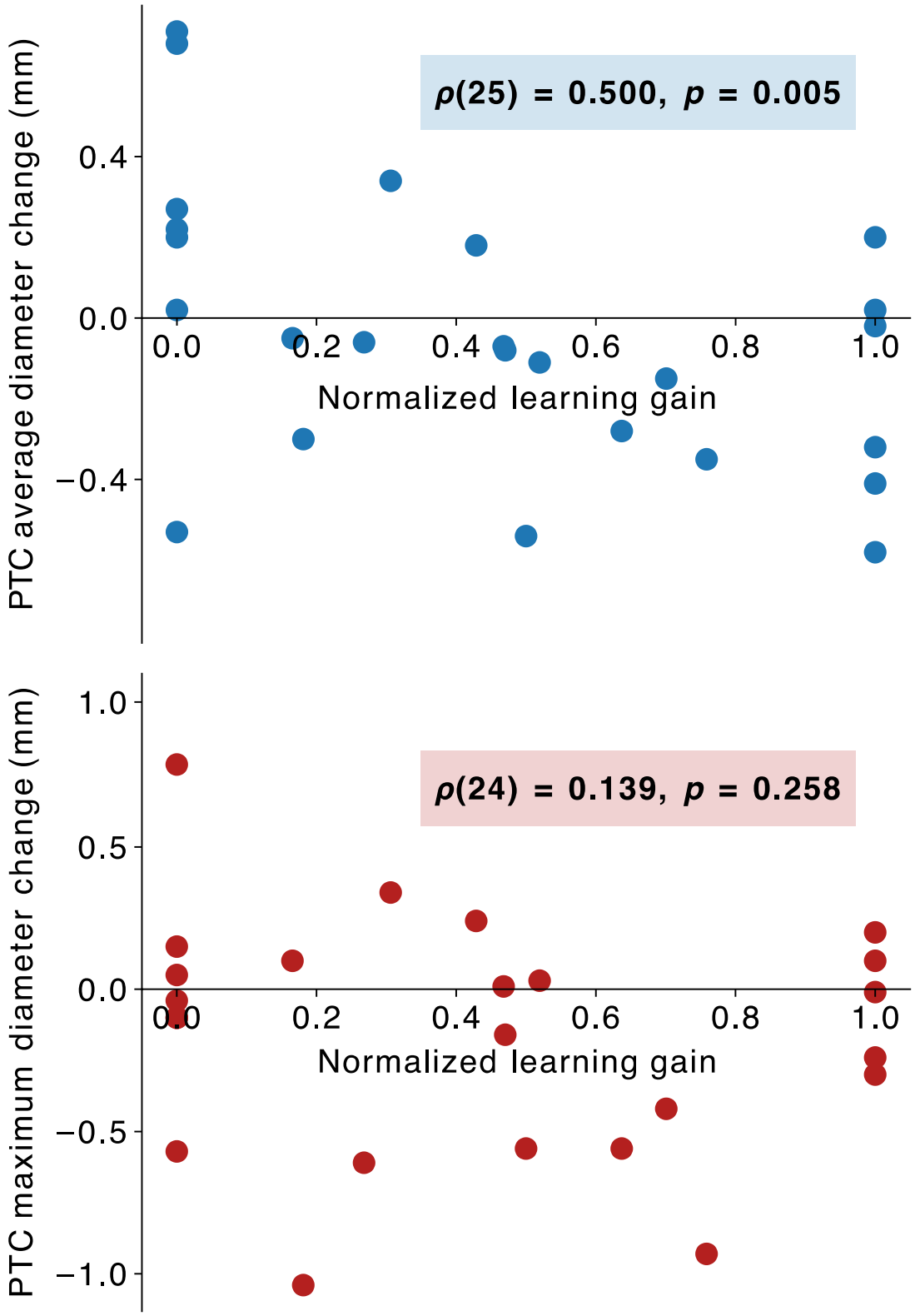


Figure 17*. Normalized learning gain and pre-test corrected (PTC) average and maximum pupil diameter changes.*

Based on the trend seen in **Figure 17**, as students become more fluent/knowledgeable in the EPF, their ability to process the information and complete the test improves, decreasing their cognitive load. The decrease in cognitive load effectively reduces the mental effort required to interpret the information. This change may allow students to retrieve information from their long-term memory and more effectively process information in their working memory.

EPF fluency may allow participants to think less about symbolism and more about reactivity

We analyzed participants' responses to the case comparison questions and determined their mode of reasoning for each question that was answered. **Figure 18** summarizes these results for the *OrgMech101* and *Acid–Base* groups. In the pre-test, both groups' causal (linear and multi-component) modes of reasoning accounted for 73% of their overall arguments. However, in the post-test, the percentage of participants' causal modes of reasoning increased to 82% for the *OrgMech101* group and decreased to 64% for the *Acid–Base* group. A chi-square test determined that the differences in these frequencies were not statistically significant (χ^2 = 0.688, p = 0.203). Looking at each individual case comparison question, we see either an increase or no change in causal modes of reasoning for the *OrgMech101* group and either no change or a decrease in causal modes of reasoning for the *Acid–Base* group. In the *Acid–Base* group, seven out of 68 responses (10%) were identified as having a more advanced mode of reasoning from pre- to post. In the *OrgMech101* group, 20 out of 76 responses (26%) were coded as having a more advanced mode of reasoning from pre- to post. We conducted an additional chi-square test to evaluate these differences, but the results were not statistically significant (χ^2 = 3.095, p = 0.079, φ = 0.147). Though insignificant, the effect size (φ) suggests a small tendency for the *OrgMech101* group to have a slightly higher proportion of participants showing an increase in reasoning ability compared to the *Acid–Base* group.

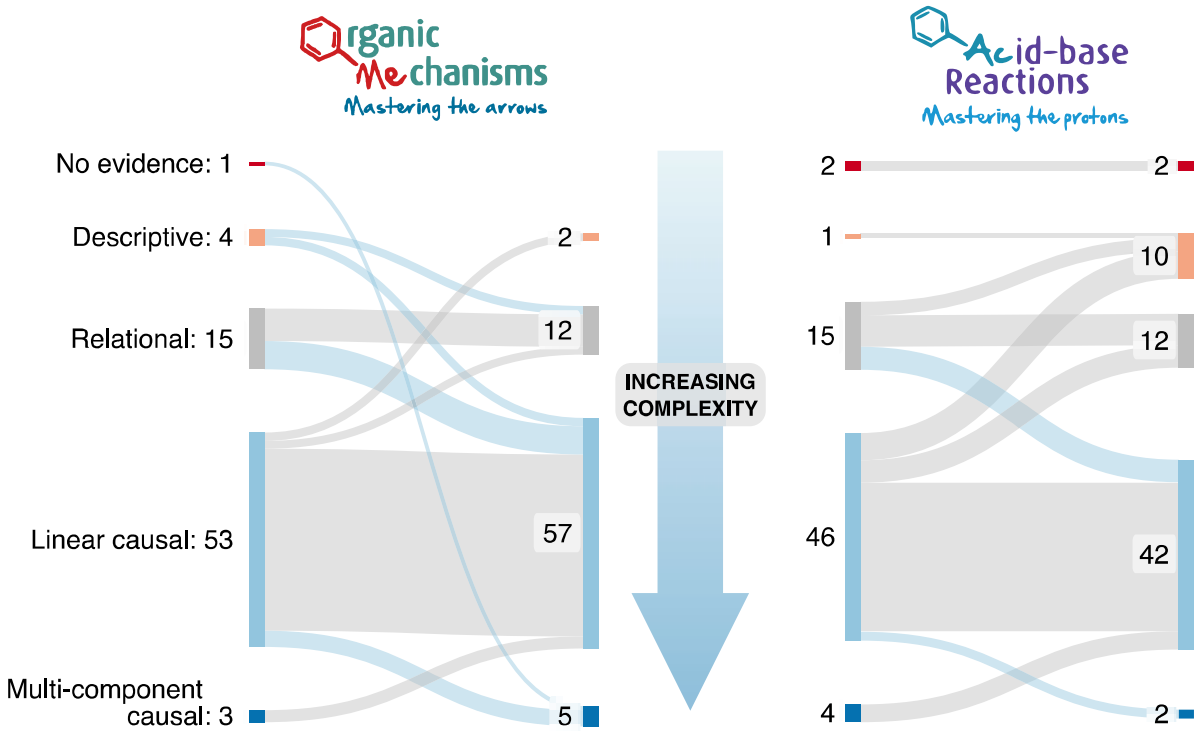


***Figure 18**. Modes of reasoning Sankey Diagram for OrgMech101 (left) and Acid–Base group participants (right). The modes of reasoning have been organized from least complex (top) to most complex (bottom). Increases in modes of reasoning from pre-test to post-test have been shown with a blue edge (connection). No change and decreases in modes of reasoning from pre-test to post-test have been coloured grey.*

The decreases in casual modes of reasoning for the *Acid–Base* group may be attributed to cognitive fatigue, as a frequent sign of this phenomenon is a decrease in commitment to a particular task (Robert J. Hockey, 1997; Van Der Linden, 2011). In other words, after spending a fair amount of time doing the pre-test and intervention, participants may have spent less effort answering the case comparison questions in the post-test because they already did so in the pre-test. Since we offered compensation to participants as a reward for participating in the study, participants may have completed the post-test quickly, as there was no consequence for providing superficial answers.

One explanation for the increase in causal arguments in the *OrgMech101* group could be that participants no longer needed to dedicate as much of a cognitive load to interpret the EPF language of the reaction. Instead, participants could focus on the features of the chemical reactions and the role that various chemical factors play. The results from **Figure 15** support this argument, showing much lower variability and a lower average pupil diameter in the post-

test. We also see evidence for this interpretation from the participants' responses to the questions. For example, **Figure 19** shows the pre-test and post-test responses of P30, a participant in the *OrgMech101* group. In their pre-test response for Q5 (carbocation), P30 does not use any arguments or pieces of evidence. They instead use the provided textbox to describe the features of the reaction they are looking at. Such descriptors include "*the double bond takes the hydrogen on the carbon next to the CF_3*" and "*the bond between HBr donates its electrons to Br, giving it a positive charge*." We expect participants' written arguments to reflect their thoughts during that particular moment since a close relationship was established between cognitive processes and written language production (Kellogg, 2008). Think-aloud protocols follow a similar assumption, but rather than using written words to indicate thoughts, spoken words represent their underlying cognitive processes (Rose and McKinley, 2020).

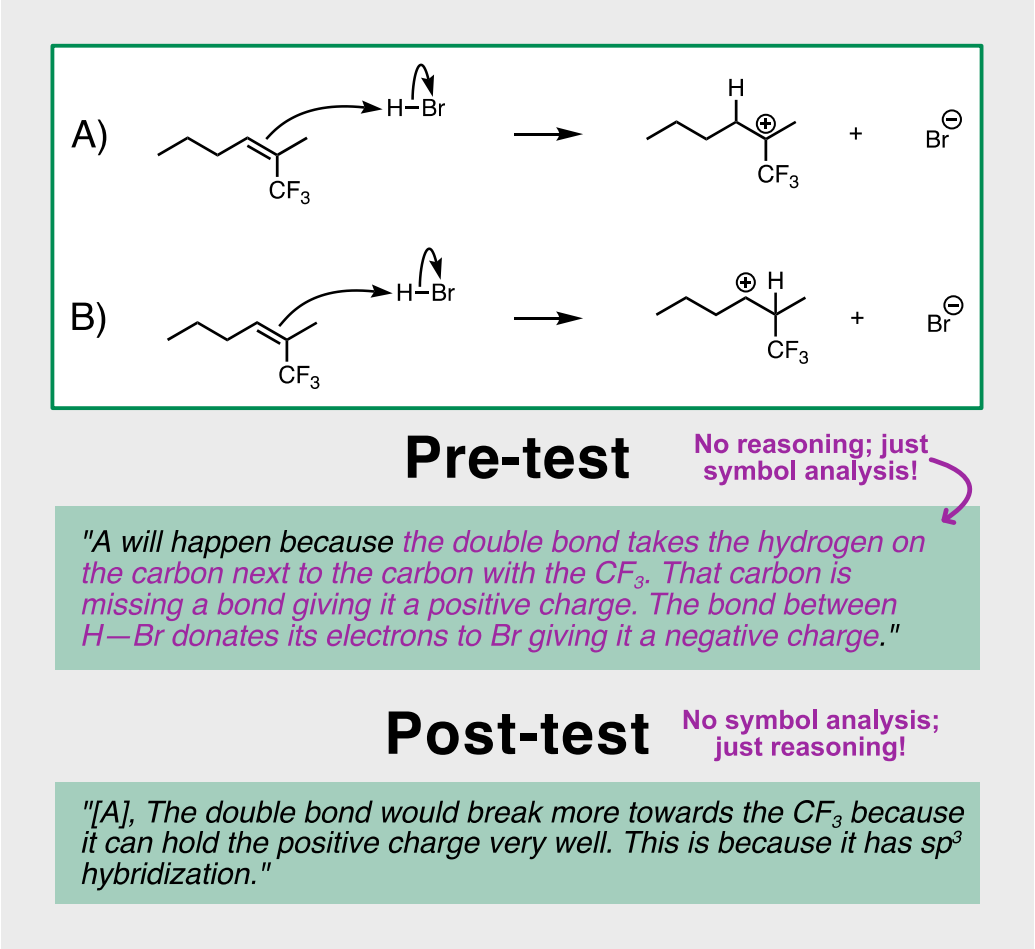


***Figure 19**. P30's pre- and post-test responses to Q5 (carbocation). P30 was a participant in the OrgMech101 group.*

From the words used in their response, we can assume that P30 was actively dedicating cognitive resources to interpret the reaction. Although the incorrect option was selected and a wrong argument was made in the post-test, the participant no longer thought about the symbols and instead focused more on reaction features. P30 mentioned charge stability and referred to the hybridization of the product, which were concepts addressed in the intervention's video on chemical factors affecting stability. This change could be due to the intervention itself, and factors like EPF fluency or cognitive load may have had no impact. However, P30 did not provide any evidence or reasoning in the pre-test, making that justification difficult. Additionally, two participants in the *Acid–Base* group also did not provide evidence in the pre-test, but unlike P30, they did not demonstrate improved reasoning despite watching the same video. Although the pre-test responses of those participants did not indicate a focus on interpreting the symbolism like P30's did, this observation was not seen in any of the other 17 participants in the *Acid–Base* group. In the *OrgMech101* group, we see a comparable result with P12, who provided an argument for Q6 (enolate) in **Figure 20**.

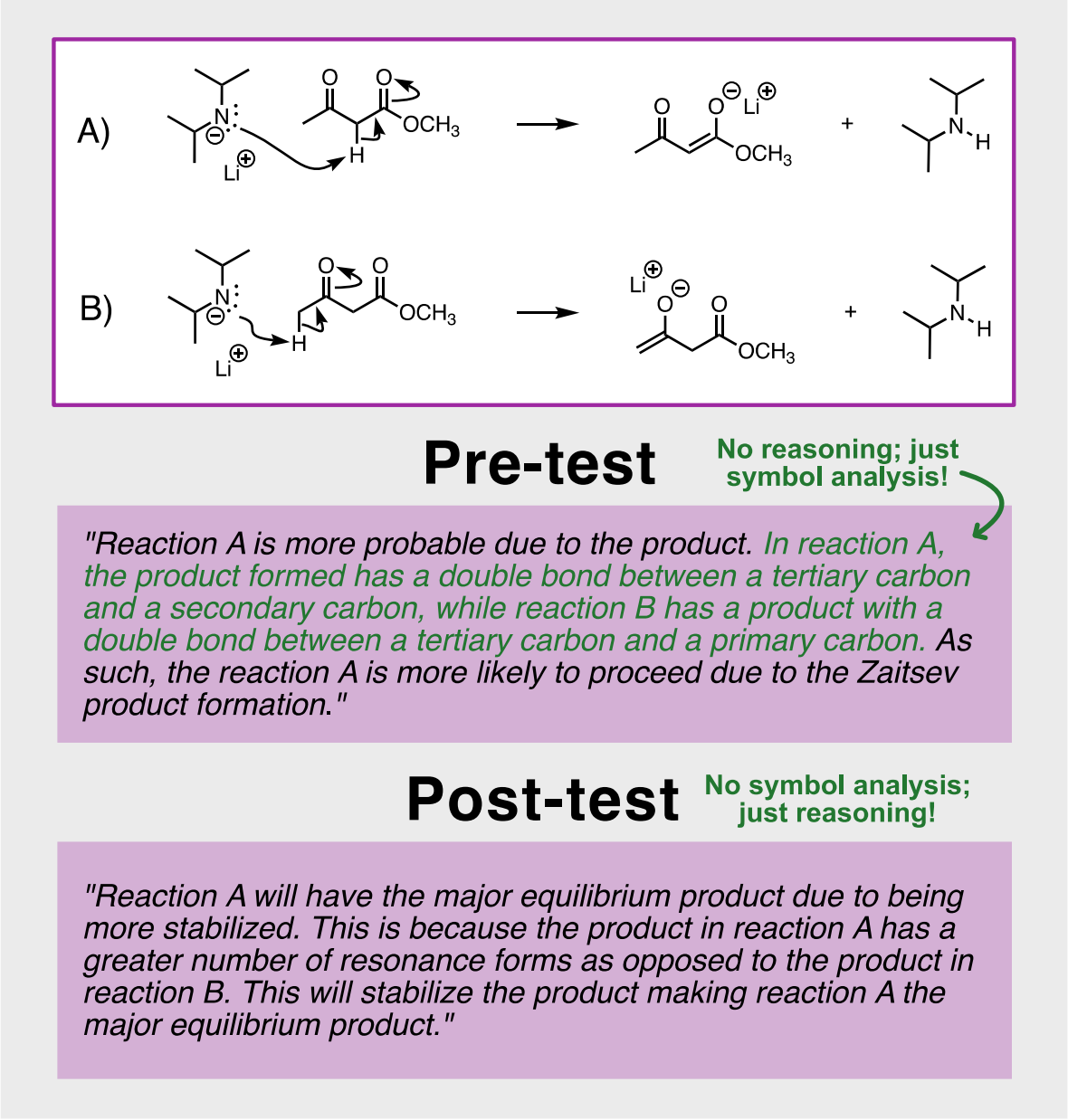


***Figure 20**. P12's pre- and post-test responses to Q6 (enolate). P12 was a participant in the OrgMech101 group.*

In the pre-test, P12 describes the difference between the two products in terms of the bonds and connections and then proceeds to list a concept (i.e., Zaitsev product formation) as the reason why the reaction would proceed without linking the two points. The participant appears to use Zaitsev product formation as a heuristic because they did not justify how the rule is relevant in this reaction. The participant seems to have made a connection in their mind between "*Zaitsev product formation*" and "*degree of substitution for an alkene*," without understanding how or why this relationship exists: alkenes with more alkyl groups exhibit

greater hyperconjugation effects, which contribute to the stabilization of the overall molecule. Without understanding the latter concept, the heuristic may fail students, as it is a general guideline and does not hold true for all elimination reactions. Students using heuristics without understanding its scientific basis is common in organic chemistry, as several studies have found that students resort to this strategy (Ferguson and Bodner, 2008; Taber, 2009; Maeyer and Talanquer, 2010). This learning strategy has benefits, such as reducing the cognitive load associated with information processing, allowing for more efficient learning. However, when heavily dependent on this strategy, it can introduce biases in students' thinking (Talanquer, 2014). In P12's post-test response, they discuss the stability of the product and support their argument by referencing the number of possible resonance structures that can be made. Rather than primarily discussing the structure of the products in terms of the nature of the π bond, the participant immediately recognizes that the favourable product is that which is more stabilized because more resonance structures can be drawn.

The remaining responses that increased in complexity may be attributed to the effect of the interventions. As will be seen in the next section, there were much more references to chemical factors that were discussed in the module, so naturally, participants' responses should become more sophisticated. However, the nature in which participants' responses changed greatly differed from that of P12 and P30. For example, P19 provided an argument that considered factors of electronegativity, atom size and ability to stabilize charge in Q6 (enolate) in the pre-test (**Figure 21**). In the post-test, however, they still refer to the factors above, but they also discuss base strength (i.e., "*therefore its conjugate base is a lot weaker*"), list the pK_a values of the conjugate acids, and reference the stability of products (i.e., "*reactions proceed to favour most stable [products]*").

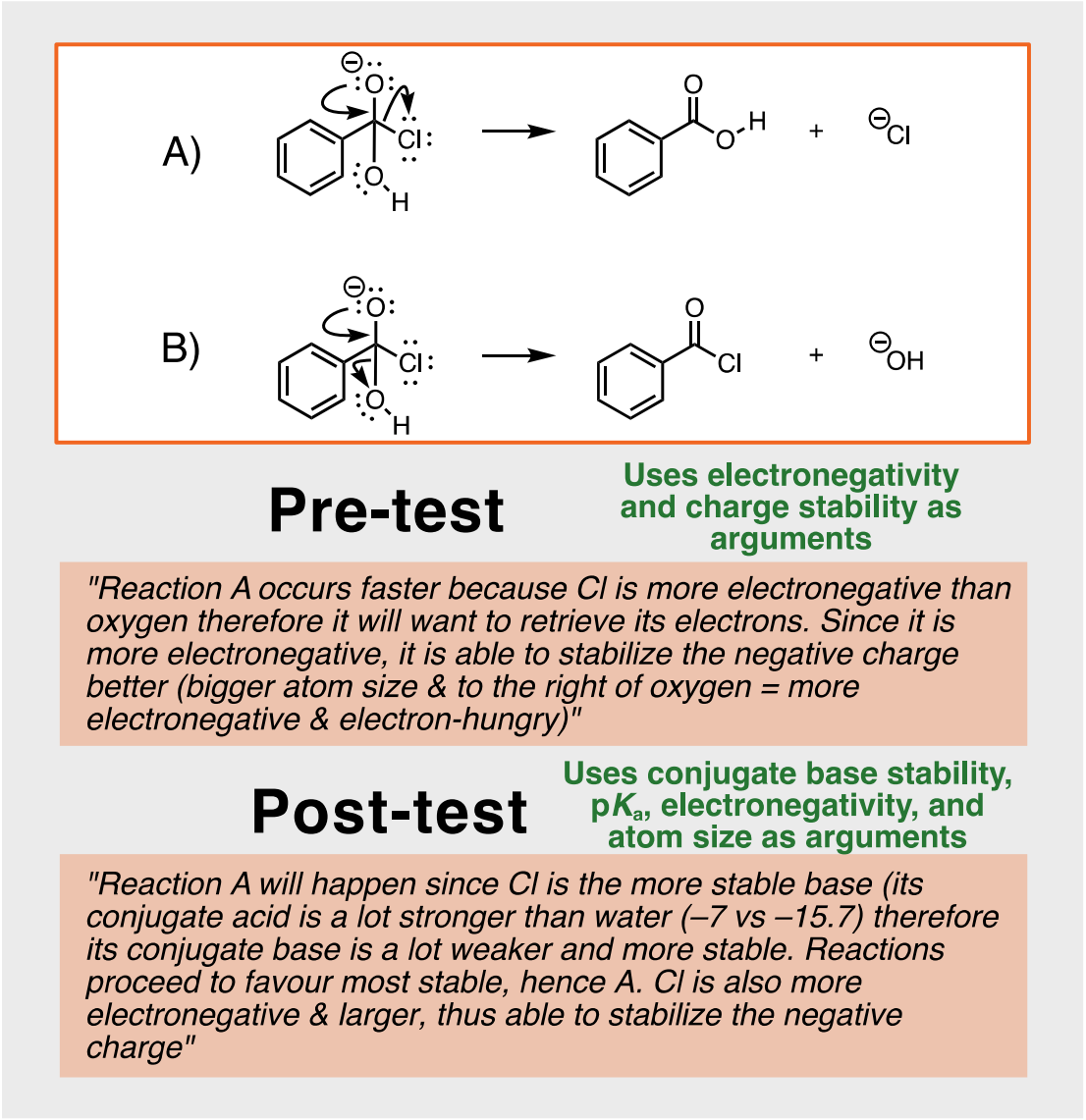


Figure 21. *P19's pre- and post-test response to Q7 (tetrahedral). P19 was a participant in the OrgMech101 group.*

This change may have been brought about solely due to the intervention, but we cannot rule out the possibility that P19, who obtained a perfect score on the post-test after scoring 84% on the pre-test, did not benefit from increased EPF fluency (and potentially a lower cognitive load). There is always some uncertainty when studying cognitive load, and we must rely on indirect measures to make inferences about what is happening in the mind (Sweller *et al.*, 2011).

EPF fluency may allow students to make more connections between concepts

Gephi networks of participants' pre- and post-test arguments for the *Acid–Base* and *OrgMech101* group were created. The pre-test networks (in grey) were overlayed over the post-test networks (in purple) to highlight differences in the arguments used (**Figure 22**). In the *Acid–Base* group, leaving group ability, thermodynamic stability, charge stability and induction were the most used arguments for the pre-test, while leaving group ability, charge stability, induction, and resonance were the most used arguments for the post-test. In the *OrgMech101* group, thermodynamic stability, leaving group ability, resonance, and induction were the most used arguments in the pre-test. In contrast, the former three and charge stability were the most used arguments in the post-test. Charge stability and thermodynamic stability were the concepts most frequently used with other concepts in the pre-test and post-test for both groups.

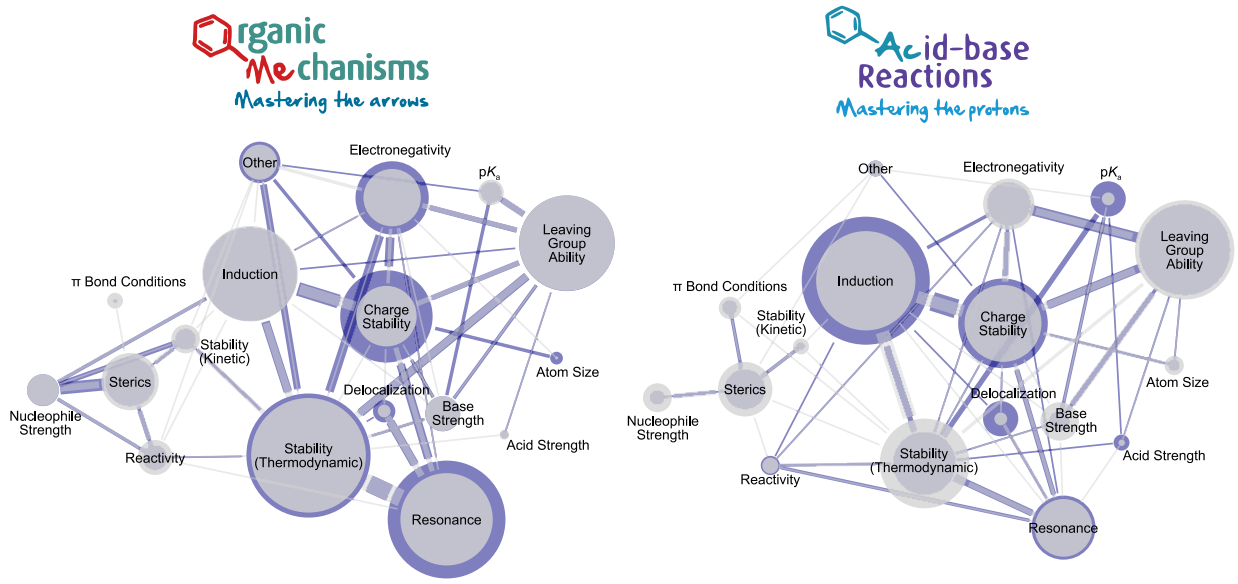


***Figure 22**. Gephi diagram of arguments used in the case comparison questions. The left figure shows arguments made by participants in the OrgMech101 group, and the right figure shows arguments made by participants in the Acid–Base group. Arguments made in the pre-test are shown in light grey, and arguments made in the post-test are shown in blue. The size of each concept represents the number of times the specific concept was referenced, and the size of the connection between two concepts represents the number of times they were used together.*

For the *Acid–Base* group, the concepts with the greatest increase in the post-test compared to the pre-test are delocalization, pK_a, charge stability, acid strength and resonance. The increase in these arguments is not surprising since they were all essential elements in the *Acid–Base* module. Acid strength and pK_a are central concepts to the module, while the remaining concepts are frequently referenced in LO3. The *OrgMech101* group sees similar increases, with the addition of electronegativity and thermodynamic stability. In the *Acid–Base* group, there is a notable increase in the use of pK_a with another argument, while charge stability is the argument with the greatest increase pre-test to post-test for the *OrgMech101* group. Once again, the former should not be surprising because pK_a was a central element in the *Acid–Base* module.

Charge stability was often referenced for both groups, mainly in conjunction with electronegativity, induction, resonance and leaving group ability. This observation makes sense because participants typically referred to charge stability only when there was an explicit positive or negative charge. When analyzing reactions, numerous studies have consistently demonstrated that students prioritize explicit features (such as the presence of a positive or negative charge) over implicit ones (Anzovino and Bretz, 2016; Talanquer, 2017).

Additionally, we conducted several analyses using the statistics panel of *Gephi* for each network. We summarize these statistics in **Table 4**.

Table 4*. Summary statistics from the Gephi application.*

Statistic	*OrgMech101*		*Acid–Base*	
	Pre-test	**Post-test**	**Pre-test**	**Post-test**
Average weighted degree	10.12	11.41	8.47	7.77
Average node size	14.88	15.82	13.53	12.77
Graph density	0.29	0.24	0.27	0.23
Modularity	0.29	0.33	0.28	0.32

Based on the four statistics, the networks for both groups appear to change similarly. In both groups, the graph density decreased (by 0.05 in the *OrgMech101* group and 0.04 for the *Acid–Base* group), indicating that the networks became less dense after participants went through their respective modules. This result suggests that the participants' arguments in both

groups became more ordered in the post-test. The modularity statistics corroborate this claim as they increase by the same amount in both groups. The only discrepancy between the two groups was the average weighted degree, which is the average number of connections for each concept. This parameter increased in the *OrgMech101* group and decreased in the *Acid–Base* group, indicating that participants in the *OrgMech101* group displayed more/stronger connections between concepts while participants in the *Acid–Base* group displayed less/weaker connections. Additionally, while the network was less dense overall, the edges were thicker, on average, and concepts used together in the pre-test were more frequently used in the post-test.

We performed a paired-samples *t*-test to determine if the differences in average weighted degrees were significant. Results from the test showed that the differences in average weighted degree for the *OrgMech101* group pre-test to post-test narrowly missed statistical significance ($t(16) = 1.584$, $p = 0.066$). The *Acid–Base* group also failed to reach the significance threshold ($t(16) = 1.112$, $p = 0.141$). The effect size (Hedges' g) for these statistical tests were 0.366 and 0.208, respectively, indicating small to medium effect sizes (Hedges, 1981). The test output suggests that the overall strength of the connections between nodes in the network may have increased for the *OrgMech101* group, but we could not rule out the possibility that this increase was due to chance. Similar statistical tests were performed for the average node size for each group, but no significant differences for the *OrgMech101* group ($t(16) = 0.662$, $p = 0.259$, $g = 0.153$) or *Acid–Base* group ($t(16) = 0.961$, $p = 0.175$, $g = 0.222$).

The decrease in overall density and increase in modularity may be an indicator of better knowledge, as previous studies showed that experts demonstrate a greater degree of organization of their concepts and thoughts compared to novices (Chi *et al.*, 1981; Ericsson and Charness, 1994; Charney *et al.*, 2007; Galloway *et al.*, 2018). The underlying principle here is that expertise is built on a foundation of knowledge and experience, which allows the organization and categorization of information more efficiently. In the context of the results from **Figure 22** and **Table 4**, participants in both groups seemed to limit their connections such that they were more organized rather than making connections that were not necessarily appropriate.

Lastly, the lack of use of hyperconjugation as an argument, in general, was concerning. A tertiary carbocation is stabilized via hyperconjugation, where surrounding C—H σ bonds can overlap with the empty p-orbital and delocalize the electrons, partially neutralizing the positive charge. Hyperconjugation requires knowledge of molecular orbitals and their relative energetic levels, and students have been shown to not only possess a limited understanding of molecular orbital diagrams but also do not often consider electronic factors when analyzing mechanisms (Jenkins *et al.*, 2019; Deng and Flynn, 2021). Instead, participants frequently attributed the inductive effect as the sole cause of carbocation stability with increasing degree of substitution.

Despite learning stability factors, participants made common organic chemistry errors

When exploring RQ4, rather than looking at participants' responses in each group, we elected to look at their responses as a collective since EPF fluency is not relevant for RQ4. **Figure 23** summarizes the correctness of participants' responses in the post-test. We only focused on the post-test to see participants' responses after briefly learning about chemical factors affecting stability. The most incorrect arguments were seen in Q5 (carbocation), while the least incorrect ones were seen in Q7 (tetrahedral). Q5 (carbocation) had multiple competing factors that influenced the reaction, and the conclusion went against what participants had learned as a heuristic (tertiary carbocations being more stable than secondary carbocations). Therefore, participants needed evidence to be able to make a conclusion. Given that Q7 (tetrahedral) is the only question that doesn't require participants to consider multiple factors, the low number of incorrect arguments is not surprising and confirms what was said by Talanquer (2006) about students only considering one aspect when analyzing chemical reactions. Although Q7 (tetrahedral) produced the least incorrect arguments, the question also had the most partially correct ones.

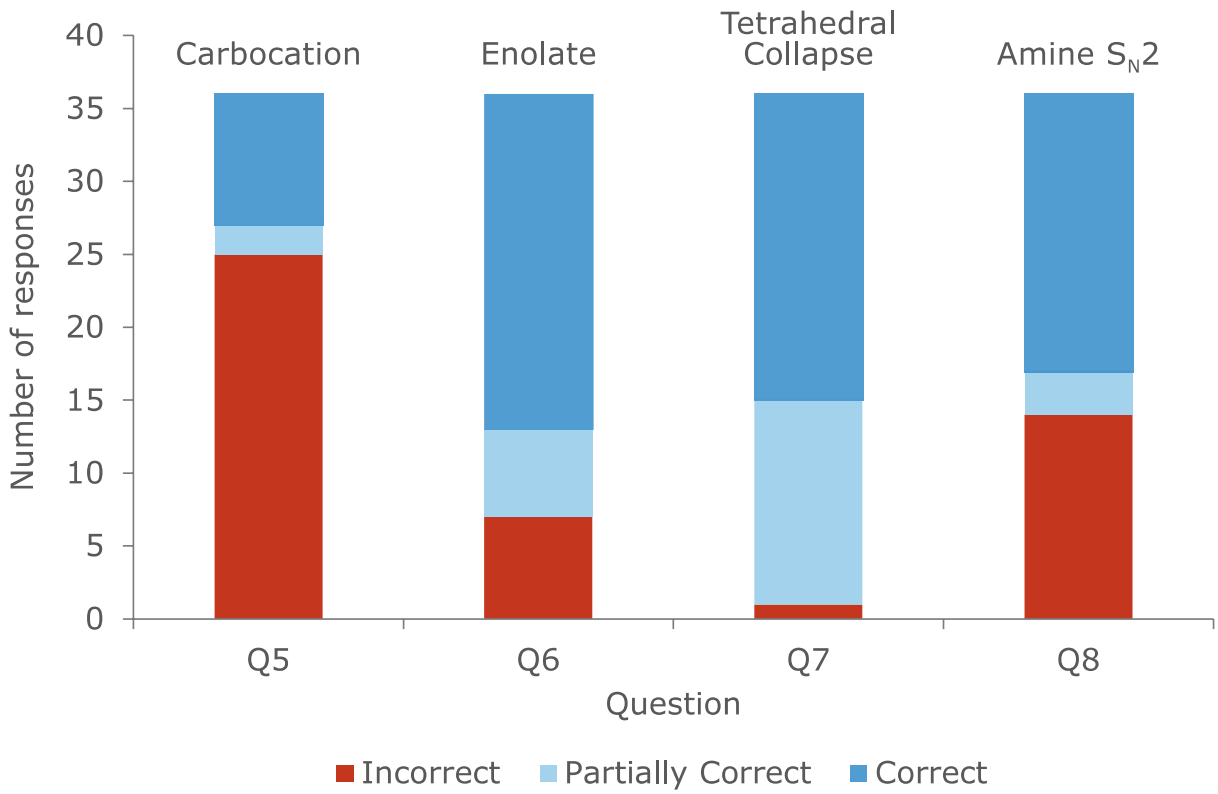


Figure 23*. Correctness of post-test responses broken down by question for participants in both Acid–Base and OrgMech101 groups.*

We analyzed participants' partially correct and incorrect responses in detail and found five common errors (**Figure 24**). Thirteen out of 47 incorrect arguments (27%) from post-test responses were incorrect because participants only considered the carbocation's degree of substitution in Q5 (carbocation). Many participants did not mention the electron-withdrawing CF_3 or its effect, despite the group being present in the carbocation. This observation shows a disconnect in students' abilities to identify multiple chemical factors in a reaction. Additionally, it likely reinforces the over-emphasis on rote memorization for our teaching. If participants can identify a stable carbocation without considering the destabilizing effect an electron-withdrawing group can have on the stability, it may demonstrate that they do not really comprehend the concept of carbocation stability and instead rely on the heuristic that "*a more substituted carbocation is the more stable carbocation*." Dood *et al.* (2020) also came across this finding, stating that this superficial knowledge hinders students' development of deeper reasoning ability. If educators want students to understand the chemical basis for activity and

reactivity, then we need to find a way to teach this concept to students better, so they do not rely purely on the heuristic. Students prioritizing the degree of substitution over the effect of induction is likely because the degree of substitution is a surface feature; a carbocation's degree of substitution is either primary, secondary, or tertiary, making it much easier to interpret than induction, a factor that primarily relies on electronic properties and analysis of the molecule itself. This observation is consistent with several research studies showing that students prioritize focusing on surface features rather than implicit chemical properties (Anderson and Bodner, 2008; Graulich *et al.*, 2019).

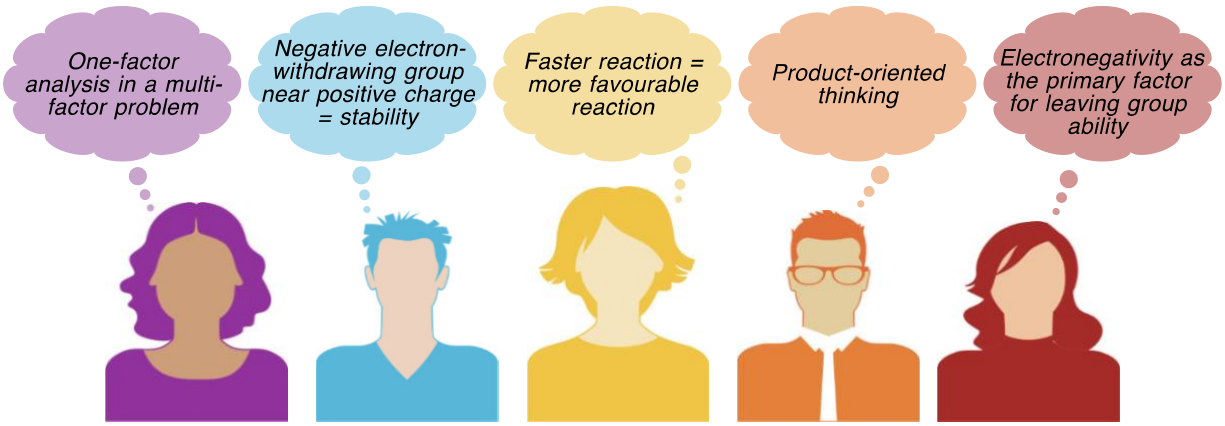
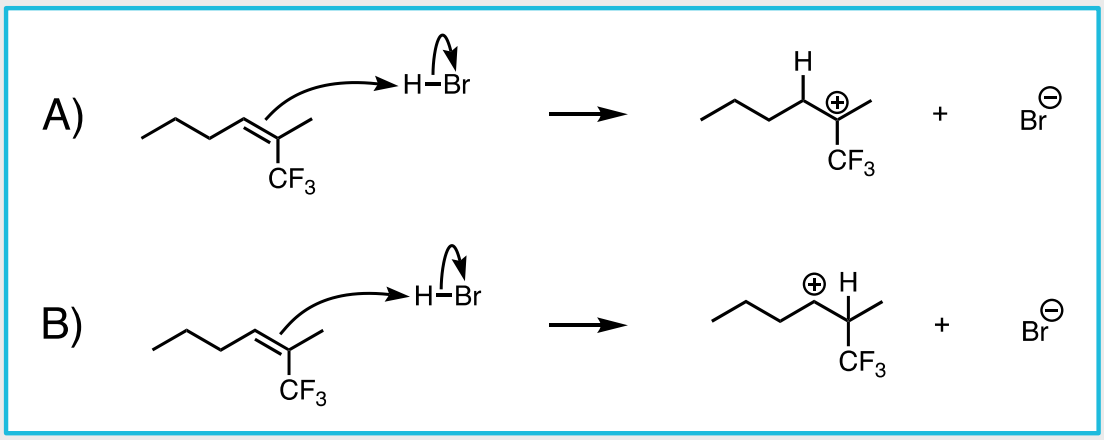


Figure 24*. The five most common misinterpretations made by students in the pre-test. The text in the thought bubbles is meant to summarize each misinterpretation and are not actual responses from participants.*

The second-most apparent error made by participants was misinterpreting the direction of an inductive effect, with 9/47 (19%) of incorrect arguments related to this misinterpretation. The IUPAC definition of induction is "*an experimentally observable effect (on rates of reaction, etc.) of the transmission of charge through a chain of atoms by electrostatic induction*" (Gold, 2019). In other words, induction is a polar effect that relies on differences in the electronegativity of adjacent atoms. Students seem to interpret induction strictly as an electron-withdrawing effect, and some participants even confused induction with hyperconjugation. For example, on Q5 (carbocation) of their pre-test, P32 referenced charge stability like many other participants, however when trying to justify how the charge gets stabilized, they stated it was "*through induction or hyperconjugation*" and added that they forgot which factor contributed to the stabilization in parentheses. In the context of Q5 (carbocation), several students recognized that CF_3 is an electron-withdrawing group (EWG);

however, they believed this had a stabilizing effect rather than a destabilizing one. **Figure 25** presents some of these responses.



"Product A is more unstable. CF_3 is an EWG, which pulls electron density away from the positive charge, further destabilizing it." — **correct response**

"[A], CF_3 is an EWG, meaning it can neutralize the positive charge (because opposite charges attract)." — **P3**

"In A, the CF_3 group is closer to the charge, stabilizing it by delocalization." — **P7**

"Product A is more stable… The positive charge close to CF_3 (highly electronegative) is stabilizing compared to B." — **P31**

***Figure 25**. Sample of student responses that incorrectly discussed induction in Q5 (carbocation).*

Five out of 47 (10%) incorrect responses were due to the third misinterpretation, where participants selected the kinetic product rather than the thermodynamic product for Q6 (enolate) (**Figure 26**). The question specifically asked participants to choose the reaction that would produce the major equilibrium product; yet despite this prompt, participants selected the reaction with the most accessible proton. The argument used was that because the proton was more accessible, the reaction was more likely to occur. This argument is true; however, the

participants failed to notice that although the proton is more accessible, the resulting product is more unstable since fewer resonance, hyperconjugation and inductive effects are present. This observation shows students may have difficulty interpreting that many reactions are reversible, and the first product produced may not be the most stable one. The reaction they selected would indeed occur more quickly due to accessibility of the proton; however, chemical reactions favour that which produces the more stable product. In other words, the reaction will equilibrate (if conditions permit) and produce the more stable product. This idea heavily reinforces that the fastest reaction does not always constitute the more favourable product.

Chemistry education studies on students' understanding of kinetic vs thermodynamic products are lacking, but from the results seen here, some students seem to have difficulty distinguishing the two from each other. Kinetic and thermodynamic enolates are taught to students in the latter half of their organic chemistry course, so it is possible that some participants took part in this study before learning the concept. However, participants throughout data collection made this misinterpretation, and there was no apparent trend or observation where this misinterpretation ceased.

P15 participated in this study about one month into the course and likely did not know anything about kinetic vs thermodynamic enolates. Their post-test response for Q6 (enolate) indicated that "*it is easier for [strong, bulky bases] to gain access to hydrogens on the ends of molecules*" and referenced another heuristic that "*strong, bulky bases favour the Hofmann, or the less-substituted products*." In contrast, P29 participated in this study one week before the end of the semester, where they would have learned about kinetic vs thermodynamic enolates. They provided this response for Q6 (enolate): "*the hydrogen on the outside is easier to reach, or rather less sterically hindered.*" Therefore, although participants who made the error before learning about kinetic vs. thermodynamic products may have avoided the same mistake had they taken part in the study after learning the concept, participants who did learn about the concept still struggled to apply it effectively. This misinterpretation draws on similar ideas from what was previously discussed; that is, the accessibility of a proton is an easy feature to analyze —a proton is either easy or difficult to access due to neighbouring functional groups.

Recognizing the effects of resonance, induction, and hyperconjugation require a more comprehensive analysis and understanding of the electronics involved in the reaction.

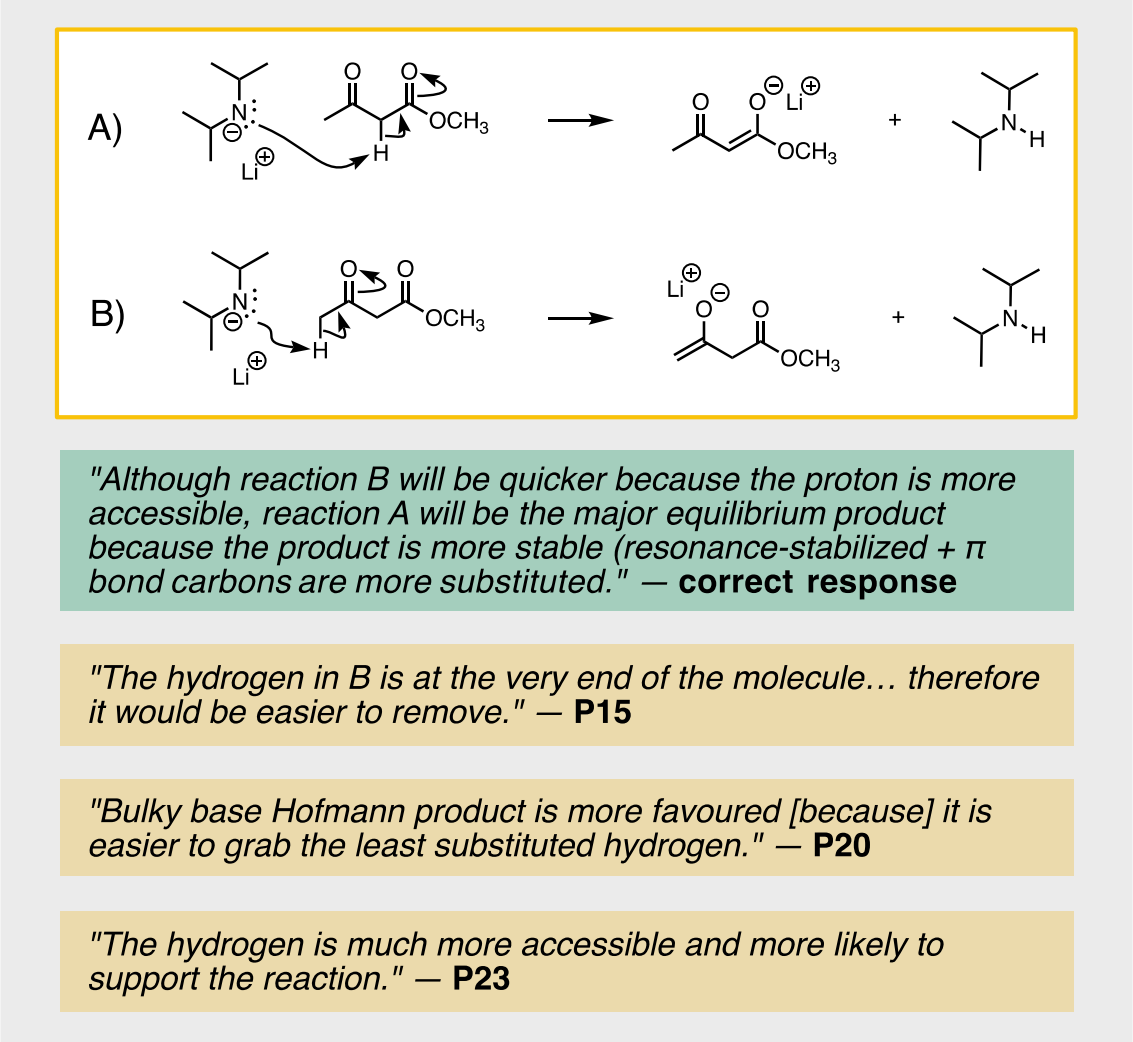
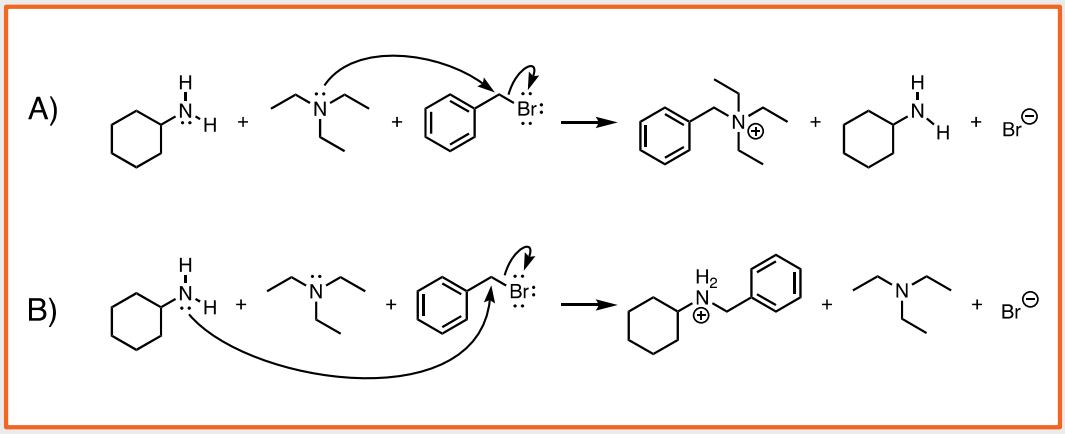


Figure 26*. Sample of student responses that predicted the kinetic product rather than the thermodynamic product in Q6 (enolate).*

The fourth common misinterpretation that students made was looking at elements of stability without considering the mechanism. In Q8 (amine), participants claimed that the favourable reaction would involve a tertiary amine acting as a nucleophile (**Figure 27**). Participants frequently discussed the increased charge stability in product A due to the adjacent alkyl electron-donating groups (EDG). While this observation would theoretically stabilize the positive charge in the nitrogen, it does not consider the steric requirements needed for the reaction to proceed. A tertiary amine is much bulkier than a secondary amine, and the

reaction's nature requires the amine's electrons to access its region of interest directly. As a result, the reaction requires a significant amount of energy to proceed, making it an unfavourable reaction. The idea that students do not analyze the mechanism's effect and instead look at the products suggests that they prioritize product-oriented thinking rather than process-orientated thinking, which other researchers have also discovered (Grove, Cooper, and Cox, 2012; Dood and Watts, 2023). In organic chemistry, reactions must be evaluated not only based on product stability, but also considering factors such as the activation energy and the overall feasibility of the reaction. (Graulich, 2015b).



"Although the product in A has the ability to be stabilized by induction, a tertiary amine acting as a nucleophile is improbable (large activation energy) for steric reasons. B is correct." — **correct response**

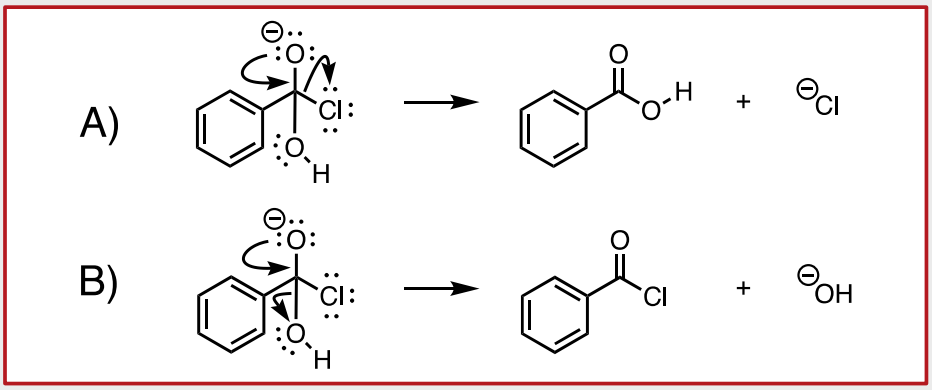
"A is favoured b/c alkanes are EDG and A has four… electron density is added into the positive charge, lessening it." — **P4**

"[In A,] the product has the ability to stabilize the positive charge among the carbons." — **P12**

"Alkyl groups are EDG which help stabilize the positive charge. The product will be more stable than the one in A due to this." — **P21**

Figure 27*. Sample of student responses that selected option A in Q8 (amine) without considering the mechanism.*

The last common misinterpretation worth noting occurred in Q7 (tetrahedral). Although 35/36 participants correctly selected the option that chloride would act as the leaving group, many referenced the difference in electronegativity between the two leaving groups when asked to explain their reasoning (**Figure 28**). Leaving group ability is a complex idea that relies on many factors, such as the size and stability of the leaving group. Typically, the larger the leaving group, the more stable it is and the better it can accommodate a negative charge. Using electronegativity to determine leaving group ability does not always work. For example, fluorine is the most electronegative atom in the periodic table, with an electronegativity of 3.98. Yet, it is a bad leaving group because of the reluctance of the fluorine atom to act as a free ion due to the significant difference in electronegativity in a C—F bond (O'Hagan, 2008).



"Reaction A is more favoured because chlorine is a larger atom, therefore, it is capable of spreading out its charge more, making it more stable with a negative charge." — **correct response**

"A because chlorine is more electronegative than oxygen. It wants the negative charge more." — **P2**

"A is more stable since chlorine is more electronegative than oxygen." — **P14**

"[A], since chlorine is more electronegative than oxygen, it can stabilize the charge much more effectively." — **P35**

***Figure 28**. Sample of student responses that discussed electronegativity of chlorine vs oxygen/hydroxide in Q7 (tetrahedral).*

Participants frequently used electronegativity as an argument for leaving group ability, despite the factor not always being appropriate to use. Moreover, they stated that chlorine was more electronegative than oxygen, which is untrue. Oxygen possesses an electronegativity of 3.44, while the electronegativity of chlorine is 3.16. This misinterpretation seems to stem beyond students' interpretations, as several publications from peer-reviewed journals have incorrectly stated that chlorine is more electronegative than oxygen (Van Der Heijden *et al.*, 2005; van der Heijden *et al.*, 2007; Subbiah *et al.*, 2018).

Students' tendency to use electronegativity as a means for determining leaving group ability draws on what was seen in prior research, where it was shown that students used electronegativity as the primary factor when assessing leaving group ability of halides (Popova and Bretz, 2018). This comparison is only ever appropriate when comparing atoms of a similar size, and halogens significantly differ in size from each other. Similarly, oxygen and chloride are found on different rows of the periodic table, and their size difference makes comparing electronegativity values inappropriate. Additionally, some participants compared the electronegativity of chlorine (an atom/element) with hydroxide (a compound). Electronegativity is typically used when comparing two atoms or elements with each other; hydroxide is a compound made up of multiple elements, making the comparison of electronegativities between the two more difficult. Rather than using electronegativity as the primary factor when assessing leaving group ability, they should be able to identify it as a factor and see if other factors may predominate over electronegativity.

The findings of Q7 (tetrahedral) highlight the importance of instructors asking causal mechanistic reasoning questions to probe students' understanding. At first glance, it seems that participants grasp the concept of leaving group ability well, as 35/36 participants correctly stated that chloride was a better leaving group than hydroxide; however, they had a misguided justification of why the leaving group was better. In fact, of the 36 participants, 19 used electronegativity as part of their argument for why chlorine was a better leaving group than hydroxide. Such mistakes can only be brought to light by probing students' reasoning abilities and addressing these misinterpretations early in the learning process.

CONCLUSIONS

This study provides valuable insights into the effects of increased EPF fluency on cognitive load and reasoning ability. The study consisted of a pre-post format separated by a learning phase with a control and treatment group, and an eye tracker was used to measure cognitive load. Participants in the treatment group (*OrgMech101*) received focused training on the electron-pushing formalism (EPF) and exhibited a significant decrease in their average pupil diameter while completing the post-test compared to the control group (*Acid–Base*). The reduction in pupil diameter suggests less cognitive load being used, which may be due to participants no longer dedicating as great of a cognitive capacity to processing the symbols they were looking at after becoming more fluent with the EPF. A significant negative relationship was observed between students' EPF fluency and their average and maximum pupil diameter changes during the pre-test to support these results further. This relationship showed that participants who scored higher on the EPF portions of pre- and post-tests were more likely to exhibit a lower pupil diameter change (and potentially cognitive load).

Although participants in the *OrgMech101* group demonstrated increased reasoning ability compared to the *Acid–Base* group, the difference was not statistically significant at $\alpha = 0.05$ ($p = 0.07$). However, disregarding a p-value of 0.07 solely due to its failure to meet the conventional significance threshold does not seem to be a thoughtful approach. The apparent increased reasoning ability in the treatment group may be due to participants' improved interpretation of the EPF and associated symbols, reducing their cognitive load and allowing for more effective chemical reasoning. Regarding case comparison questions, participants performed better on questions with one chemical factor to consider than two. Case-comparison questions with two factors highlighted students' over-reliance on product-oriented thinking, one-factor decision making, and heuristics.

Potential limitations

While the present study offers valuable insights into the relationship between EPF fluency, cognitive load, and reasoning ability, the research's limitations may impact the interpretation and generalizability of the results. First, participants' EPF fluency level depends only on their performance in two draw-the-arrow and one draw-the-product questions. This sample of questions may not be large enough to accurately represent students' EPF comprehension.

The eye tracker was placed on participants' heads, and although they were instructed to limit movement as much as possible, they were allowed to move freely. Head movements or facial muscle contraction could have caused slippage in the device and altered readings. No additional data collection method was paired with eye-tracking (i.e., data triangulation), so changes in pupil diameter may have been due to reasons beyond changes in cognitive load, such as stress. Future studies should incorporate additional measures such as cognitive load rating scales or retrospective interviews to further corroborate findings from other measures and strengthen the validity of the pupil data (**Figure 29**).

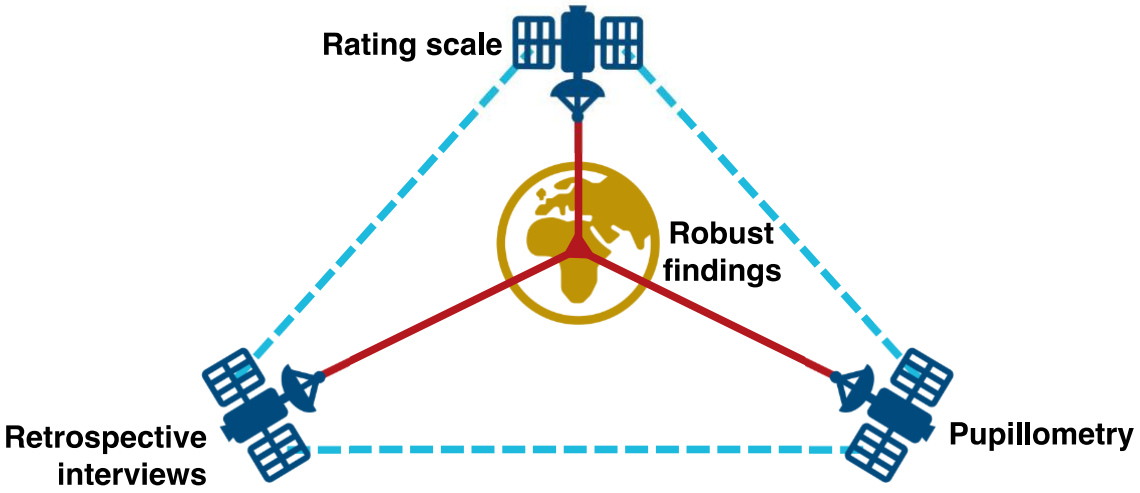


Figure 29*. Example of data triangulation using pupillometry, rating scale measures and retrospective interviews.*

While the sampling rate of the eye-tracker produced 250 readings per second, a few increases in pupil diameter were likely due to pen stroke colour changes, though this effect is likely minimal. Although the sample was large enough to perform qualitative analyses, many quantitative analyses conducted to analyze eye-tracking data were done using a relatively small

sample size, especially when separating the two groups. Additionally, in the intake form, there were no eye-tracking exclusion criteria questions about medication; therefore, participants taking medication may have had their pupil diameter altered.

For the case comparison questions, some participants did not provide detailed answers, despite the researcher (AY) instructing participants to be as detailed as possible in their responses. Therefore, their responses may not be representative of their knowledge. Additionally, even with the shortening of the pre/post-test, participants still took longer than expected to complete the pre-test, intervention, and post-test, which may have introduced cognitive fatigue as a confounding variable. Lastly, participants were only recruited from one institution for organic chemistry. Each institution has its teaching practices, resources, and emphasis on topics that potentially influence students' knowledge and performance. Therefore, generalizing findings to students in other institutions with different curricula may be inappropriate.

Implications for teaching

The most important teaching implication from this study is that organic chemistry should be taught in a manner that optimizes germane cognitive load as expertise is developed, thereby lessening the overall mental effort and allowing students to process and encode information into their long-term memory effectively. One possible way of achieving germane cognitive load optimization is to prioritize the teachings of the EPF. This notation is the language of organic chemistry and is central to understanding and analyzing chemical reactions taught later in the subject. By emphasizing the significance of the EPF, students can develop a solid foundation that would put them in a better position to succeed as they progress through the subject.

When assessing students' understanding, designing assessments that encourage reasoning and critical thinking is essential. One practical approach to achieving this is by including causal mechanistic reasoning questions in assessments. These questions prompt students to consider not only the "what" of a chemical reaction, but also the "how" and "why," the latter of which is crucial for the development of reasoning ability. Students can expand their

knowledge of organic chemistry by analyzing the mechanisms and factors that dictate how/why the reaction proceeds.

Additionally, steps need to be taken to build the metacognitive skill of knowing when heuristics are appropriate to use. While heuristics are helpful for student learning in some contexts, they can also limit students' understanding of organic chemistry due to the oversimplification that comes with the strategy. A proposed method is to introduce a heuristic and provide several examples that contradict the rule. This approach would allow educators to guide students toward more effective learning of the subject matter. Steps should also be taken to encourage students to have a more "process-oriented" way of thinking. Adopting this teaching style would allow students to examine the characteristics of reaction products and analyze the reaction's mechanism.

Because students seem to grasp reactions with a single factor to consider quite well, additional practice should be dedicated to interpreting reactions with multiple factors. By providing exposure to scenarios that require the consideration of numerous factors, students can develop a more complex understanding of reactivity and avoid isolated thinking where only one factor is considered. If these factors are competing, providing experimental data (e.g., pK_a values, equilibrium constants, thermodynamic parameters) would offer valuable insights into the dominant factor, preventing students from making unnecessary guesses and mimicking the work of scientists, who use experimental evidence to test hypotheses. For example, when considering the acid–base reaction between hydroxide and 1-pentyne, students may face challenges in determining its directionality due to competing factors that stabilize each conjugate base. On one hand, the electronegativity of oxygen increases the stability of the hydroxide, while the hybridization state of 1-pentyne allows for greater electron donation, thereby stabilizing its conjugate base. When looking at the pK_a values of the acids (water = ~16 vs 1-pentyne = ~25), only then does it become clear that the effect of electronegativity predominates. If students have access to these experimental data, instead of thinking about which direction the reaction proceeds, they are instead able to think about what chemical factors are prevalent and which one predominates.

Implications for research

To gain a more comprehensive understanding of the relationship between EPF fluency, cognitive load, and reasoning ability, further research is needed with larger sample sizes and more challenging EPF questions to maximize the ability to view participant learning gains. Moreover, future studies should aim to expand on the existing findings and validate the obtained pupil data through various means, such as think-aloud protocols and response process validity. These methods allow researchers to gain deeper insights into how students reason through different reactions, and they can also help ensure the validity and reliability of the collected data.

Another area of exploration could involve investigating the long-term effects of EPF fluency on students' academic performance in their organic chemistry courses. A study that tracks students' academic performance, germane cognitive load, and experience over an extended period would provide insight into whether a strong foundation in EPF fluency positively impacts students' overall understanding and success in organic chemistry.

REFERENCES

Afify M. K., (2020), Effect of interactive video length within E-learning environments on cognitive load, cognitive achievement and retention of learning. *Turk. Online J. Distance Educ.*, 68–89, DOI: 10.17718/tojde.803360.

Alfieri L., Nokes-Malach T. J., and Schunn C. D., (2013), Learning through case comparisons: A meta-analytic review. *Educ. Psychol.*, **48**(2), 87–113, DOI: 10.1080/00461520.2013.775712.

Anderson T. L. and Bodner G. M., (2008), What can we do about 'Parker'? A case study of a good student who didn't 'get' organic chemistry. *Chem. Educ. Res. Pract.*, **9**(2), 93–101, DOI: 10.1039/B806223B.

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 Bretz S. L., (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.

Ausubel D. P., (1968), *Educational psychology: a cognitive view*, 2nd ed. Holt, Rinehart and Winston.

Ausubel D. P., (1963), *The psychology of meaningful verbal learning*, Grune & Stratton.

Beatty J., (1982), Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.*, **91**(2), 276–292, DOI: 10.1037/0033-2909.91.2.276.

Becker N., Stanford C., Towns M., and Cole R., (2015), Translating across macroscopic, submicroscopic, and symbolic levels: the role of instructor facilitation in an inquiry-oriented physical chemistry class. *Chem. Educ. Res. Pract.*, **16**(4), 769–785, DOI: 10.1039/C5RP00064E.

Berland L. K. and Reiser B. J., (2011), Classroom communities' adaptations of the practice of scientific argumentation. *Sci. Educ.*, **95**(2), 191–216, DOI: 10.1002/sce.20420.

Bhattacharyya G., (2014), Trials and tribulations: student approaches and difficulties with proposing mechanisms using the electron-pushing formalism. *Chem. Educ. Res. Pract.*, **15**(4), 594–609, DOI: 10.1039/C3RP00127J.

Bhattacharyya G. and Bodner G. M., (2005), "It gets me to the product": How students propose organic mechanisms. *J. Chem. Educ.*, **82**(9), 1402, DOI: 10.1021/ed082p1402.

Blondel V. D., Guillaume J.-L., Lambiotte R., and Lefebvre E., (2008), Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **2008**(10), P10008, DOI: 10.1088/1742-5468/2008/10/P10008.

Bodé N. E., Caron J., and Flynn A. B., (2016), Evaluating students' learning gains and experiences from using nomenclature101.com. *Chem. Educ. Res. Pract.*, **17**(4), 1156–1173, DOI: 10.1039/C6RP00132G.

Bodé 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. *J. Chem. Educ.*, **96**(6), 1068–1082, DOI: 10.1021/acs.jchemed.8b00719.

Bodner G. M. and Domin D. S., (2000), Mental models: the role of representations in problem solving in chemistry. *Univ. Chem. Educ.*

Bowman B. G., Karty J. M., and Gooch G., (2007), Teaching a modified Hendrickson, Cram, and Hammond curriculum in organic chemistry. *J. Chem. Educ.*, **84**(7), 1209, DOI: 10.1021/ed084p1209.

Bretz S. L., (2001), Novak's Theory of Education: human constructivism and meaningful learning. *J. Chem. Educ.*, **78**(8), 1107, DOI: 10.1021/ed078p1107.6.

Buchner J., Buntins K., and Kerres M., (2022), The impact of augmented reality on cognitive load and performance: a systematic review. *J. Comput. Assist. Learn.*, **38**(1), 285–303, DOI: 10.1111/jcal.12617.

Carle M. S., Visser R., and Flynn A. B., (2020), Evaluating students' learning gains, strategies, and errors using OrgChem101's module: organic mechanisms—mastering the arrows. *Chem. Educ. Res. Pract.*, **21**(2), 582–596, DOI: 10.1039/C9RP00274J.

Carlson R., Chandler P., and Sweller J., (2003), Learning and understanding science instructional material. *J. Educ. Psychol.*, **95**(3), 629–640, DOI: 10.1037/0022-0663.95.3.629.

Charney J., Hmelo-Silver C. E., Sofer W., Neigeborn L., Coletta S., and Nemeroff M., (2007), Cognitive apprenticeship in science through immersion in laboratory practices. *Int. J. Sci. Educ.*, **29**(2), 195–213, DOI: 10.1080/09500690600560985.

Chi M. T. H., Feltovich P. J., and Glaser R., (1981), Categorization and representation of physics problems by experts and novices. *Cogn. Sci.*, **5**(2), 121–152, DOI: 10.1207/s15516709cog0502_2.

Christian K. and Talanquer V., (2012), Content-related interactions in self-initiated study groups. *Int. J. Sci. Educ.*, **34**(14), 2231–2255, DOI: 10.1080/09500693.2012.708064.

Coletta V. P. and Steinert J. J., (2020), Why normalized gain should continue to be used in analyzing preinstruction and postinstruction scores on concept inventories. *Phys. Rev. Phys. Educ. Res.*, **16**(1), 010108, DOI: 10.1103/PhysRevPhysEducRes.16.010108.

Cook M. P., (2006), Visual representations in science education: The influence of prior knowledge and cognitive load theory on instructional design principles. *Sci. Educ.*, **90**(6), 1073–1091, DOI: 10.1002/sce.20164.

Cooper M. M., Corley L. M., and Underwood S. M., (2013), An investigation of college chemistry students' understanding of structure-property relationships. *J. Res. Sci. Teach.*, **50**(6), 699–721, DOI: 10.1002/tea.21093.

Cooper M. M., Stowe R. L., Crandell O. M., and Klymkowsky M. W., (2019), Organic Chemistry, Life, the Universe and Everything (OCLUE): a transformed organic chemistry curriculum. *J. Chem. Educ.*, **96**(9), 1858–1872, DOI: 10.1021/acs.jchemed.9b00401.

Cooper M. M., Underwood S. M., and Hilley C. Z., (2012), Development and validation of the implicit information from Lewis structures instrument (IILSI): do students connect

structures with properties? *Chem. Educ. Res. Pract.*, **13**(3), 195–200, DOI: 10.1039/C2RP00010E.

Cranford K. N., Tiettmeyer J. M., Chuprinko B. C., Jordan S., and Grove N. P., (2014), Measuring load on working memory: the use of heart rate as a means of measuring chemistry students' cognitive load. *J. Chem. Educ.*, **91**(5), 641–647, DOI: 10.1021/ed400576n.

Dancey C. P. and Reidy J., (2007), *Statistics without maths for psychology: using SPSS for Windows*, 4th ed. Pearson/Prentice Hall.

Dechsri P., Jones L. L., and Heikkinen H. W., (1997), Effect of a labratory manual design incorporating visual information-processing aids on student learning and attitudes. *J. Res. Sci. Teach.*, **34**(9), 891–904, DOI: 10.1002/(SICI)1098-2736(199711)34:9<891::AID-TEA4>3.0.CO;2-P.

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.

Deng J. M., Rahmani M., and Flynn A. B., (2022), The role of language in students' justifications of chemical phenomena. *Int. J. Sci. Educ.*, **44**(13), 2131–2151, DOI: 10.1080/09500693.2022.2114299.

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. 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.

Erduran S., (2019), *Argumentation in chemistry education: research, policy and practice*, Royal Society of Chemistry.

Ericsson K. A. and Charness N., (1994), Expert performance: Its structure and acquisition. *Am. Psychol.*, **49**(8), 725–747, DOI: 10.1037/0003-066X.49.8.725.

Evagorou M. and Osborne J., (2013), Exploring young students' collaborative argumentation within a socioscientific issue. *J. Res. Sci. Teach.*, **50**(2), 209–237, DOI: 10.1002/tea.21076.

Evans J. St. B. T., (2003), In two minds: dual-process accounts of reasoning. *Trends Cogn. Sci.*, **7**(10), 454–459, DOI: 10.1016/j.tics.2003.08.012.

Ferguson R. and Bodner G. M., (2008), Making sense of the arrow-pushing formalism among chemistry majors enrolled in organic chemistry. *Chem. Educ. Res. Pract.*, **9**(2), 102–113, DOI: 10.1039/B806225K.

Flynn A. B., (2011), Developing problem-solving skills through retrosynthetic analysis and clickers in organic chemistry. *J. Chem. Educ.*, **88**(11), 1496–1500, DOI: 10.1021/ed200143k.

Flynn A. B. and Amellal D. G., (2016), Chemical information literacy: pK_a values—where do students go wrong? *J. Chem. Educ.*, **93**(1), 39–45, DOI: 10.1021/acs.jchemed.5b00420.

Flynn A. B., Caron J., Laroche J., Daviau-Duguay M., Marcoux C., and Richard G., (2014), Nomenclature101.com: a free, student-driven organic chemistry nomenclature learning tool. *J. Chem. Educ.*, **91**(11), 1855–1859, DOI: 10.1021/ed500353a.

Flynn A. B. and Featherstone R. B., (2017), Language of mechanisms: exam analysis reveals students' strengths, strategies, and errors when using the electron-pushing formalism (curved arrows) in new reactions. *Chem. Educ. Res. Pract.*, **18**(1), 64–77, DOI: 10.1039/C6RP00126B.

Flynn A. B. and Ogilvie W. W., (2015), Mechanisms before reactions: a mechanistic approach to the organic chemistry curriculum based on patterns of electron flow. *J. Chem. Educ.*, **92**(5), 803–810, DOI: 10.1021/ed500284d.

Galloway K. R. and Bretz S. L., (2015), Measuring meaningful learning in the undergraduate general chemistry and organic chemistry laboratories: a longitudinal study. *J. Chem. Educ.*, **92**(12), 2019–2030, DOI: 10.1021/acs.jchemed.5b00754.

Galloway K. R., Leung M. W., and Flynn A. B., (2018), A comparison of how undergraduates, graduate students, and professors organize organic chemistry reactions. *J. Chem. Educ.*, **95**(3), 355–365, DOI: 10.1021/acs.jchemed.7b00743.

Galloway K. R., Leung M. W., and Flynn A. B., (2019), Patterns of reactions: a card sort task to investigate students' organization of organic chemistry reactions. *Chem. Educ. Res. Pract.*, **20**(1), 30–52, DOI: 10.1039/C8RP00120K.

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.

Gilzenrat M. S., Nieuwenhuis S., Jepma M., and Cohen J. D., (2010), Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cogn. Affect. Behav. Neurosci.*, **10**(2), 252–269, DOI: 10.3758/CABN.10.2.252.

Gkitzia V., Salta K., and Tzougraki C., (2011), Development and application of suitable criteria for the evaluation of chemical representations in school textbooks. *Chem. Educ. Res. Pract.*, **12**(1), 5–14, DOI: 10.1039/C1RP90003J.

Gkitzia V., Salta K., and Tzougraki C., (2020), Students' competence in translating between different types of chemical representations. *Chem. Educ. Res. Pract.*, **21**(1), 307–330, DOI: 10.1039/C8RP00301G.

Golafshani N., (2015), Understanding reliability and validity in qualitative research. *Qual. Rep.*, **8**(4), 597–606, DOI: 10.46743/2160-3715/2003.1870.

Gold V. ed., (2019), *The IUPAC Compendium of Chemical Terminology: The Gold Book*, 4th ed. International Union of Pure and Applied Chemistry (IUPAC), DOI: 10.1351/goldbook.

Goodwin W. M., (2008), Structural formulas and explanation in organic chemistry. *Found. Chem.*, **10**(2), 117–127, DOI: 10.1007/s10698-007-9033-2.

Graulich N., (2015a), Intuitive judgments govern students' answering patterns in multiple-choice exercises in organic chemistry. *J. Chem. Educ.*, **92**(2), 205–211, DOI: 10.1021/ed500641n.

Graulich N., (2015b), The tip of the iceberg in organic chemistry classes: how do students deal with the invisible? *Chem. Educ. Res. Pract.*, **16**(1), 9–21, DOI: 10.1039/C4RP00165F.

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.

Grove N. P. and Bretz S. L., (2012), A continuum of learning: from rote memorization to meaningful learning in organic chemistry. *Chem. Educ. Res. Pract.*, **13**(3), 201–208, DOI: 10.1039/C1RP90069B.

Grove N. P. and Bretz S. L., (2010), Perry's Scheme of Intellectual and Epistemological Development as a framework for describing student difficulties in learning organic chemistry. *Chem. Educ. Res. Pract.*, **11**(3), 207–211, DOI: 10.1039/C005469K.

Grove N. P., Cooper M. M., and Cox E. L., (2012), Does mechanistic thinking improve student success in organic chemistry? *J. Chem. Educ.*, **89**(7), 850–853, DOI: 10.1021/ed200394d.

Grove N. P., Cooper M. M., and Rush K. M., (2012), Decorating with arrows: toward the development of representational competence in organic chemistry. *J. Chem. Educ.*, **89**(7), 844–849, DOI: 10.1021/ed2003934.

Grove N. P., Hershberger J. W., and Bretz S. L., (2008), Impact of a spiral organic curriculum on student attrition and learning. *Chem. Educ. Res. Pract.*, **9**(2), 157–162, DOI: 10.1039/B806232N.

Guo Z., Chen R., Zhang K., Pan Y., and Wu J., (2016), The impairing effect of mental fatigue on visual sustained attention under monotonous multi-object visual attention task in long durations: an event-related potential based study. *PLoS One*, **11**(9), e0163360, DOI: 10.1371/journal.pone.0163360.

Hacıeminoğlu E., (2021), Factors predicting middle school pupils' learning orientations: a multilevel analysis. *Educ. Q. Rev.*, **4**(3), 409–423, DOI: 10.31014/aior.1993.04.03.349.

Hedges L. V., (1981), Distribution Theory for Glass's estimator of effect size and related estimators. *J. Educ. Stat.*, **6**(2), 107, DOI: 10.2307/1164588.

Hess E. H. and Polt J. M., (1960), Pupil size as related to interest value of visual stimuli. *Science*, **132**(3423), 349–350, DOI: 10.1126/science.132.3423.349.

Hess E. H. and Polt J. M., (1964), Pupil size in relation to mental activity during simple problem-solving. *Science*, **143**(3611), 1190–1192, DOI: 10.1126/science.143.3611.1190.

Hinde R. A., Tamplin A., and Barrett J., (1993), A comparative study of relationship structure. *Br. J. Soc. Psychol.*, **32**(3), 191–207, DOI: 10.1111/j.2044-8309.1993.tb00995.x.

Hmelo-Silver C. E., Marathe S., and Liu L., (2007), Fish swim, rocks sit, and lungs breathe: expert-novice understanding of complex systems. *J. Learn. Sci.*, **16**(3), 307–331, DOI: 10.1080/10508400701413401.

Huitt W., (2003), The information processing approach to cognition. *Educ. Psychol. Interact.*

Hurtado-Rodríguez D., Salinas-Torres A., Rojas H., Becerra D., and Castillo J.-C., (2022), Bioactive 2-pyridone-containing heterocycle syntheses using multicomponent reactions. *RSC Adv.*, **12**(54), 34965–34983, DOI: 10.1039/D2RA07056A.

Jenkins J. L., Shoopman B. T., and Department of Chemistry, Eastern Kentucky University, Richmond, Kentucky, United States of America, (2019), Identifying misconceptions that limit student understanding of molecular orbital diagrams. *Sci. Educ. Int.*, **30**(3), 152–157, DOI: 10.33828/sei.v30.i3.1.

Johnstone A., (1982), Macro- and micro-chemistry. *Sch. Sci. Rev.*, **64**, 377–379.

Johnstone A. H., (2000), Teaching of chemistry - logical or psychological? *Chem. Educ. Res. Pract.*, **1**(1), 9–15, DOI: 10.1039/A9RP90001B.

Kahneman D. and Beatty J., (1966), Pupil diameter and load on memory. *Science*, **154**(3756), 1583–1585, DOI: 10.1126/science.154.3756.1583.

Kalyuga S., (2011), Cognitive load theory: how many types of load does It really need? *Educ. Psychol. Rev.*, **23**(1), 1–19, DOI: 10.1007/s10648-010-9150-7.

Keller S., Rumann S., and Habig S., (2021), Cognitive load implications for augmented reality supported chemistry learning. *Information*, **12**(3), 96, DOI: 10.3390/info12030096.

Kellogg R. T., (2008), Training writing skills: A cognitive developmental perspective. *J. Writ. Res.*, **1**(1), 1–26, DOI: 10.17239/jowr-2008.01.01.1.

Kirschner P. A., (2002), Cognitive load theory: implications of cognitive load theory on the design of learning. *Learn. Instr.*, **12**(1), 1–10, DOI: 10.1016/S0959-4752(01)00014-7.

Klepsch M., Schmitz F., and Seufert T., (2017), Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Front. Psychol.*, **8**, 1997, DOI: 10.3389/fpsyg.2017.01997.

Klinger J., (2010), Measuring cognitive load during visual tasks by combining pupillometry and eye tracking.

Kraft A., Strickland A. M., and Bhattacharyya G., (2010), Reasonable reasoning: multi-variate problem-solving in organic chemistry. *Chem. Educ. Res. Pract.*, **11**(4), 281–292, DOI: 10.1039/C0RP90003F.

Kuhn D. and Dean Jr. D., (2004), Connecting scientific reasoning and causal inference. *J. Cogn. Dev.*, **5**(2), 261–288, DOI: 10.1207/s15327647jcd0502_5.

Kuhn D., Iordanou K., Pease M., and Wirkala C., (2008), Beyond control of variables: what needs to develop to achieve skilled scientific thinking? *Cogn. Dev.*, **23**(4), 435–451, DOI: 10.1016/j.cogdev.2008.09.006.

Lapierre K. R. and Flynn A. B., (2020), An online categorization task to investigate changes in students' interpretations of organic chemistry reactions. *J. Res. Sci. Teach.*, **57**(1), 87–111, DOI: 10.1002/tea.21586.

Lapierre K. R., Streja N., and Flynn A. B., (2022), Investigating the role of multiple categorization tasks in a curriculum designed around mechanistic patterns and principles. *Chem. Educ. Res. Pract.*, **23**(3), 545–559, DOI: 10.1039/D1RP00267H.

Liao C.-W., Chen C.-H., and Shih S.-J., (2019), The interactivity of video and collaboration for learning achievement, intrinsic motivation, cognitive load, and behavior patterns in a digital game-based learning environment. *Comput. Educ.*, **133**, 43–55, DOI: 10.1016/j.compedu.2019.01.013.

Lipton M. A., (2020), Reorganization of the organic chemistry curriculum to improve student outcomes. *J. Chem. Educ.*, **97**(4), 960–964, DOI: 10.1021/acs.jchemed.9b00606.

Liu Y., Zeng Z., Huang W., Shreeve J. M., and Tang Y., (2022), From nitro- to heterocycle-functionalized 1,2,4-triazol-3-one derivatives: achieving high-performance insensitive energetic compounds. *J. Org. Chem.*, **87**(6), 4226–4231, DOI: 10.1021/acs.joc.1c03065.

Loudon G. M. and Parise J., (2016), *Organic Chemistry*, sixth edition. Macmillan Education.

Lowenstein O., (1950), Role of sympathetic and parasympathetic systems in reflex dilatation of the pupil: pupillographic studies. *Arch. Neurol. Psychiatry*, **64**(3), 313, DOI: 10.1001/archneurpsyc.1950.02310270002001.

Luten R., (2002), Managing the unique size-related Issues of pediatric resuscitation: reducing cognitive load with resuscitation aids. *Acad. Emerg. Med.*, **9**(8), 840–847, DOI: 10.1197/aemj.9.8.840.

MacGillivray B. H., (2014), Heuristics structure and pervade formal risk assessment: heuristics structure and pervade formal risk assessment. *Risk Anal.*, **34**(4), 771–787, DOI: 10.1111/risa.12136.

Maeyer J. and Talanquer V., (2010), The role of intuitive heuristics in students' thinking: Ranking chemical substances. *Sci. Educ.*, **94**(6), 963–984, DOI: 10.1002/sce.20397.

Marcus N., Cooper M., and Sweller J., (1996), Understanding instructions. *J. Educ. Psychol.*, **88**(1), 49–63, DOI: 10.1037/0022-0663.88.1.49.

Mathewson J. H., (2005), The visual core of science: definition and applications to education. *Int. J. Sci. Educ.*, **27**(5), 529–548, DOI: 10.1080/09500690500060417.

Mavilidi M. F. and Zhong L., (2019), Exploring the development and research focus of cognitive load theory, as described by its founders: interviewing John Sweller, Fred Paas, and Jeroen van Merriënboer. *Educ. Psychol. Rev.*, **31**(2), 499–508, DOI: 10.1007/s10648-019-09463-7.

Mayer R. E., (2021), Cognitive theory of multimedia learning, in *The Cambridge Handbook of Multimedia Learning*, Mayer R. E. and Fiorella L. (eds.)., Cambridge University Press, pp. 57–72, DOI: 10.1017/9781108894333.008.

Mayer R. E., (2012), Information processing., in *APA educational psychology handbook, Vol 1: Theories, constructs, and critical issues.*, Harris K. R., Graham S., Urdan T., McCormick C. B., Sinatra G. M., and Sweller J. (eds.)., American Psychological Association, pp. 85–99, DOI: 10.1037/13273-004.

Mayer R. E., (2002), Rote versus meaningful learning. *Theory Pract.*, **41**(4), 226–232, DOI: 10.1207/s15430421tip4104_4.

McCright A. M. and Dunlap R. E., (2011), The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *Sociol. Q.*, **52**(2), 155–194, DOI: 10.1111/j.1533-8525.2011.01198.x.

McHugh M. L., (2012), Interrater reliability: the kappa statistic. *Biochem. Medica*, **22**(3), 276–282.

Milenković D. D., Segedinac M. D., and Hrin T. N., (2014), Increasing high school students' chemistry performance and reducing cognitive load through an instructional strategy

based on the interaction of multiple levels of knowledge representation. *J. Chem. Educ.*, **91**(9), 1409–1416, DOI: 10.1021/ed400805p.

Milenković D. D., Segedinac M. D., Hrin T. N., and Cvjetićanin S., (2014), Cognitive load at different levels of chemistry representations. *Croat. J. Educ.*, **16**(3), 699–722.

Mooring S. R., Mitchell C. E., and Burrows N. L., (2016), Evaluation of a flipped, large-enrollment organic chemistry course on student attitude and achievement. *J. Chem. Educ.*, **93**(12), 1972–1983, DOI: 10.1021/acs.jchemed.6b00367.

Morewedge C. K. and Kahneman D., (2010), Associative processes in intuitive judgment. *Trends Cogn. Sci.*, **14**(10), 435–440, DOI: 10.1016/j.tics.2010.07.004.

Morgulis Y., Kumar R. K., Lindeman R., and Velan G. M., (2012), Impact on learning of an e-learning module on leukaemia: a randomised controlled trial. *BMC Med. Educ.*, **12**(1), 36, DOI: 10.1186/1472-6920-12-36.

Naicker P., Anoopkumar-Dukie S., Grant G. D., Neumann D. L., and Kavanagh J. J., (2016), Central cholinergic pathway involvement in the regulation of pupil diameter, blink rate and cognitive function. *Neuroscience*, **334**, 180–190, DOI: 10.1016/j.neuroscience.2016.08.009.

Newell B. R. and Shanks D. R., (2003), Take the best or look at the rest? Factors influencing "one-reason" decision making. *J. Exp. Psychol. Learn. Mem. Cogn.*, **29**(1), 53–65, DOI: 10.1037/0278-7393.29.1.53.

Novak J. D., (2010), *Learning, creating, and using knowledge: concept maps as facilitative tools in schools and corporations*, 2nd ed. Routledge.

Nyachwaya J. M. and Wood N. B., (2014), Evaluation of chemical representations in physical chemistry textbooks. *Chem. Educ. Res. Pract.*, **15**(4), 720–728, DOI: 10.1039/C4RP00113C.

O'Hagan D., (2008), Understanding organofluorine chemistry. An introduction to the C–F bond. *Chem Soc Rev*, **37**(2), 308–319, DOI: 10.1039/B711844A.

Pazicni S. and Flynn A. B., (2019), Systems thinking in chemistry education: theoretical challenges and opportunities. *J. Chem. Educ.*, **96**(12), 2752–2763, DOI: 10.1021/acs.jchemed.9b00416.

Piquado T., Isaacowitz D., and Wingfield A., (2010), Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, **47**(3), 560–569, DOI: 10.1111/j.1469-8986.2009.00947.x.

Popova M. and Bretz S. L., (2018), Organic Chemistry Students' Understandings of what makes a good leaving group. *J. Chem. Educ.*, **95**(7), 1094–1101, DOI: 10.1021/acs.jchemed.8b00198.

Porter G., Troscianko T., and Gilchrist I. D., (2007), Effort during visual search and counting: Insights from pupillometry. *Q. J. Exp. Psychol.*, **60**(2), 211–229, DOI: 10.1080/17470210600673818.

Qin S., Hermans E. J., Van Marle H. J. F., and Fernández G., (2012), Understanding Low reliability of memories for neutral information encoded under stress: alterations in memory-

related activation in the hippocampus and midbrain. *J. Neurosci.*, **32**(12), 4032–4041, DOI: 10.1523/JNEUROSCI.3101-11.2012.

Raker J. R., Trate J. M., Holme T. A., and Murphy K., (2013), Adaptation of an instrument for measuring the cognitive complexity of organic chemistry exam items. *J. Chem. Educ.*, **90**(10), 1290–1295, DOI: 10.1021/ed400373c.

Reteig L. C., Van Den Brink R. L., Prinssen S., Cohen M. X., and Slagter H. A., (2019), Sustaining attention for a prolonged period of time increases temporal variability in cortical responses. *Cortex*, **117**, 16–32, DOI: 10.1016/j.cortex.2019.02.016.

Richardson M., Abraham C., and Bond R., (2012), Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychol. Bull.*, **138**(2), 353–387, DOI: 10.1037/a0026838.

Richland L. E., Stigler J. W., and Holyoak K. J., (2012), Teaching the conceptual structure of mathematics. *Educ. Psychol.*, **47**(3), 189–203, DOI: 10.1080/00461520.2012.667065.

Robert J. Hockey G., (1997), Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biol. Psychol.*, **45**(1–3), 73–93, DOI: 10.1016/S0301-0511(96)05223-4.

Rose H. and McKinley J. eds., (2020), *The Routledge handbook of research methods in applied linguistics*, Routledge.

Russell A. and Hannon D., (2012), A cognitive load approach to learner-centered design of digital instructional media and supporting accessibility tools. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, **56**(1), 556–560, DOI: 10.1177/1071181312561116.

Salame I. I., Casino P., and Hodges N., (2020), Examining challenges that students face in learning organic chemistry synthesis. *Int. J. Chem. Educ. Res.*, **3**(3), 1–9, DOI: 10.20885/ijcer.vol4.iss1.art1.

Sevian H. and Talanquer V., (2014), Rethinking chemistry: a learning progression on chemical thinking. *Chem. Educ. Res. Pract.*, **15**(1), 10–23, DOI: 10.1039/C3RP00111C.

Shelby A. and Ernst K., (2013), Story and science: how providers and parents can utilize storytelling to combat anti-vaccine misinformation. *Hum. Vaccines Immunother.*, **9**(8), 1795–1801, DOI: 10.4161/hv.24828.

Sim J. H. and Daniel E. G. S., (2014), Representational competence in chemistry: a comparison between students with different levels of understanding of basic chemical concepts and chemical representations. *Cogent Educ.*, **1**(1), 991180, DOI: 10.1080/2331186X.2014.991180.

Solhjoo S., Haigney M. C., McBee E., Van Merrienboer J. J. G., Schuwirth L., Artino A. R., et al., (2019), Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. *Sci. Rep.*, **9**(1), 14668, DOI: 10.1038/s41598-019-50280-3.

Stensaker B., Bilbow G. T., Breslow L., and Van Der Vaart R. eds., (2017), *Strengthening teaching and learning in research universities*, Springer International Publishing, DOI: 10.1007/978-3-319-56499-9.

Stieff M., Ryu M., and Yip J. C., (2013), Speaking across levels – generating and addressing levels confusion in discourse. *Chem. Educ. Res. Pract.*, **14**(4), 376–389, DOI: 10.1039/C3RP20158A.

Stoyanovich C., Gandhi A., and Flynn A. B., (2015), Acid–base learning outcomes for students in an introductory organic chemistry course. *J. Chem. Educ.*, **92**(2), 220–229, DOI: 10.1021/ed5003338.

Straumanis A. R. and Ruder S. M., (2009), New bouncing curved arrow technique for the depiction of organic mechanisms. *J. Chem. Educ.*, **86**(12), 1389, DOI: 10.1021/ed086p1389.

Subbiah A. S., Mathews N., Mhaisalkar S., and Sarkar S. K., (2018), Novel plasma-assisted low-temperature-processed sno_2 thin films for efficient flexible perovskite photovoltaics. *ACS Energy Lett.*, **3**(7), 1482–1491, DOI: 10.1021/acsenergylett.8b00692.

Sweller J., (1988), Cognitive load during problem solving: effects on learning. *Cogn. Sci.*, **12**(2), 257–285, DOI: 10.1207/s15516709cog1202_4.

Sweller J., (2010), Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educ. Psychol. Rev.*, **22**(2), 123–138, DOI: 10.1007/s10648-010-9128-5.

Sweller J., Ayres P., and Kalyuga S., (2011), Measuring cognitive load, in *Cognitive Load Theory*, Springer New York, pp. 71–85, DOI: 10.1007/978-1-4419-8126-4_6.

Sweller J., van Merrienboer J., and Paas F., (1998), Cognitive architecture and instructional design. *Educ. Psychol. Rev.*, **10**(3), 251–296.

Taber K. S., (2009), College students' conceptions of chemical stability: the widespread adoption of a heuristic rule out of context and beyond its range of application. *Int. J. Sci. Educ.*, **31**(10), 1333–1358, DOI: 10.1080/09500690801975594.

Taber K. S., (2013), Revisiting the chemistry triplet: drawing upon the nature of chemical knowledge and the psychology of learning to inform chemistry education. *Chem. Educ. Res. Pract.*, **14**(2), 156–168, DOI: 10.1039/C3RP00012E.

Talanquer V., (2014), Chemistry education: ten heuristics to tame. *J. Chem. Educ.*, **91**(8), 1091–1097, DOI: 10.1021/ed4008765.

Talanquer V., (2006), Commonsense chemistry: a model for understanding students' alternative conceptions. *J. Chem. Educ.*, **83**(5), 811–816.

Talanquer V., (2017), Concept inventories: predicting the wrong answer may boost performance. *J. Chem. Educ.*, **94**(12), 1805–1810, DOI: 10.1021/acs.jchemed.7b00427.

Talanquer V., (2011), Macro, submicro, and symbolic: the many faces of the chemistry "triplet." *Int. J. Sci. Educ.*, **33**(2), 179–195, DOI: 10.1080/09500690903386435.

Tangen J. L. and Borders L. D., (2017), Applying information processing theory to supervision: an initial exploration. *Couns. Educ. Superv.*, **56**(2), 98–111, DOI: 10.1002/ceas.12065.

Taskin V. and Bernholt S., (2014), Students' understanding of chemical formulae: a review of empirical research. *Int. J. Sci. Educ.*, **36**(1), 157–185, DOI: 10.1080/09500693.2012.744492.

Thalmann M., Souza A. S., and Oberauer K., (2019), How does chunking help working memory? *J. Exp. Psychol. Learn. Mem. Cogn.*, **45**(1), 37–55, DOI: 10.1037/xlm0000578.

Urrestilla N. and St-Onge D., (2020), Measuring cognitive load: heart-rate variability and pupillometry assessment, in *Companion publication of the 2020 international conference on multimodal interaction*, ACM, pp. 405–410, DOI: 10.1145/3395035.3425203.

Van Der Heijden A. W. A. M., Bellière V., Alonso L. E., Daturi M., Manoilova O. V., and Weckhuysen B. M., (2005), Destructive adsorption of CCl_4 over lanthanum-based solids: linking activity to acid−base properties. *J. Phys. Chem. B*, **109**(50), 23993–24001, DOI: 10.1021/jp054689b.

Van Der Linden D., (2011), The urge to stop: the cognitive and biological nature of acute mental fatigue., in *Cognitive fatigue: multidisciplinary perspectives on current research and future applications.*, Ackerman P. L. (ed.), American Psychological Association, pp. 149–164, DOI: 10.1037/12343-007.

van der Heijden A. W. A. M., Garcia Ramos M., and Weckhuysen B. M., (2007), Intermediates in the destruction of chlorinated C1 hydrocarbons on La-based materials: mechanistic implications. *Chem. Eur. J.*, **13**(34), 9561–9571, DOI: 10.1002/chem.200700901.

Visser R. and Flynn A. B., (2018), What are students' learning and experiences in an online learning tool designed for cognitive and metacognitive skill development? *Collect. Essays Learn. Teach.*, **11**, DOI: 10.22329/celt.v11i0.5039.

Vollhardt K. P. C. and Schore N. E., (2018), *Organic chemistry: structure and function*, 8e ed. W.H. Freeman, Macmillan Learning.

Wang L., Zhang Z., McArdle J. J., and Salthouse T. A., (2008), Investigating ceiling effects in longitudinal data analysis. *Multivar. Behav. Res.*, **43**(3), 476–496, DOI: 10.1080/00273170802285941.

Webber D. M. and Flynn A. B., (2018), How are students solving familiar and unfamiliar organic chemistry mechanism questions in a new curriculum? *J. Chem. Educ.*, **95**(9), 1451–1467, DOI: 10.1021/acs.jchemed.8b00158.

Xie L., Vanlandeghem K., Isenberger K. M., and Bernier C., (2003), Kinetic enolate formation by lithium arylamide: effects of basicity on selectivity. *J. Org. Chem.*, **68**(2), 641–643, DOI: 10.1021/jo0263465.

Zagermann J., Pfeil U., and Reiterer H., (2016), Measuring cognitive load using eye tracking technology in visual computing, in *Proceedings of the sixth workshop on beyond time and errors on novel evaluation methods for visualization*, ACM, pp. 78–85, DOI: 10.1145/2993901.2993908.

Zhonggen Y., Ying Z., Zhichun Y., and Wentao C., (2019), Student satisfaction, learning outcomes, and cognitive loads with a mobile learning platform. *Comput. Assist. Lang. Learn.*, **32**(4), 323–341, DOI: 10.1080/09588221.2018.1517093.

Appendix A: Recruitment text and videos

Cc: Organic chemistry professor

Email subject: Study on EPF fluency & cognitive load/mechanistic reasoning

Dear [Professor X],

My name is Ahmed Youssef, and I am a graduate student working in Dr. Alison Flynn's chemistry education research group. My work focuses on EPF and its effect on students' cognitive load and mechanistic reasoning skills.

To help us understand how students in introductory organic chemistry courses reason through mechanisms and the role of EPF fluency, can you please post the following message on your Brightspace website by [date]? This study has received ethics approval at the University of Ottawa (File H01-22-7661).

[ENGLISH TEXT]

Hi uOttawa Orgo student!

My name is Ahmed Youssef, and I am a graduate student working in Dr. Alison Flynn's chemistry education research group. We are studying the role organic chemistry language (symbols, arrows, structures) proficiency plays in students' cognitive load, knowledge retention, and mechanistic reasoning.

Please consider participating in this study by clicking the following link: surveymonkey.ca/r/82MK6DP. We are seeking participants of all backgrounds and skill levels. We need people like you to help us improve science education at uOttawa. You will be rewarded with a $10–40 Amazon gift card, depending on how much of our study you complete.

More information about our group's research can be found at FlynnResearchGroup.com.

Thank you,

Ahmed Youssef

MSc Student | The University of Ottawa | Flynn Research Group

[FRENCH TEXT]

Bonjour, étudiant en chimie organique de l'Université d'Ottawa !

Je m'appelle Ahmed Youssef et je suis un étudiant diplômé travaillant dans le groupe de recherche sur l'enseignement de la chimie du Dre Alison Flynn. Nous étudions le rôle que joue la maîtrise du langage de la chimie organique (symboles, flèches, structures) dans la charge cognitive, la rétention des connaissances et la confiance des étudiants.

Veuillez envisager de participer à cette étude en cliquant sur le lien suivant : surveymonkey.ca/r/82MK6DP. Nous recherchons des participants de tous les horizons et de tous les niveaux de compétence. Nous avons besoin de personnes comme vous pour nous aider à améliorer l'enseignement des sciences à l'Université d'Ottawa. Vous serez récompensé par une carte cadeau Amazon de 10 à 40 $, le montant dépendant de la partie de notre étude à laquelle vous aurez participé. Veuillez noter que l'étude sera menée uniquement en anglais.

Vous trouverez de plus amples renseignements sur les recherches de notre groupe à FlynnResearchGroup.com.

Merci,

Ahmed Youssef

Étudiant en MSc | Université d'Ottawa | Groupe de recherche Flynn

Video Links

https://youtu.be/8XAenz2VYh4 (EN)

https://youtu.be/DYbgBgaz00Q (FR)

Video Transcripts

[ENGLISH]

Hi everyone! My name is Ahmed and I'm a graduate student doing my Masters in the Flynn Research Group. My research focuses on investigating the importance of organic chemistry language, that is organic chemistry symbols representations that you're taught during your course, and how it affects student success in the subject. Specifically, we're looking at how organic chemistry language proficiency affects three specific facets, the first being cognitive load, that is how much information is processed. The second is knowledge retention, how well that information is retained over time and three, the ability of students to mechanistically reason or explain why a reaction proceeds the way it does. In the past, we've used information collected from our research to aid and improve student learning. Examples of this include creating new learning tools and also modifying the curriculum to maximize student success.

I am looking for students in Ottawa currently enrolled in Organic Chemistry I to participate in my study. Participation is totally optional and will not affect your academic standing in the course. You have my word that your participation along with any information you share will remain strictly confidential. We will NOT share the names of any participant with anyone else at any time. To thank you for your participation, you will be compensated with an Amazon gift card of value up to $20. If that doesn't convince you, I don't know what will. If you're looking to contribute to the betterment of organic chemistry education, please scan the following QR code or click the link provided in your course's webpage for more details. Feel free to also email me at the email shown on the screen if you have any questions/concerns. Thank you for taking the time to watch this video and wishing you the absolute best!

[FRENCH]

Bonjour à tous, je m'appelle Allison et je travaille dans le groupe de recherche Flynn. Voici Ahmed, il est étudiant en maîtrise au sein du groupe. Ses recherches portent sur l'importance du langage de la chimie organique, c'est-à-dire les représentations des symboles de la chimie organique que l'on vous enseigne pendant votre cours, et sur la façon dont cela affecte la réussite des étudiants dans cette matière. Plus précisément, il étudie comment la maîtrise du langage de la chimie organique affecte trois facettes spécifiques, la première étant la charge cognitive, c'est-à-dire la quantité d'informations traitées. La deuxième est la rétention des connaissances, c'est-à-dire la manière dont ces informations sont retenues au fil du temps, et la troisième est la capacité des étudiants à raisonner de manière mécaniste ou à expliquer pourquoi une réaction se déroule comme elle le fait. Dans le passé, nous avons utilisé les informations recueillies dans le cadre de nos recherches pour faciliter et améliorer l'apprentissage des élèves. Par exemple, nous avons créé de nouveaux outils d'apprentissage et modifié le programme d'études pour maximiser la réussite des étudiants. Nous recherchons des étudiant.es de l'Université d'Ottawa actuellement inscrits au cours de Chimie organique I pour participer à cette étude, qui se déroulera en anglais. La participation est totalement facultative et n'affectera pas votre niveau académique dans le cours. Vous avez mes assurrances que votre participation ainsi que toute information que vous partagerez resteront strictement confidentielles. Nous ne communiquerons à aucun moment le nom des participants à qui que ce soit. Pour vous remercier de votre participation, vous recevrez une carte cadeau Amazon d'une valeur maximale de 20 dollars. Si cela ne vous convainc pas, je ne sais pas ce qui le fera. Si vous souhaitez contribuer à l'amélioration de l'enseignement de la chimie organique, veuillez scanner le code QR suivant ou cliquer sur le lien fourni dans la page Web de votre cours pour plus de détails. N'hésitez pas non plus à envoyer un courriel à l'adresse indiquée à l'écran si vous avez des questions ou des inquiétudes. Merci d'avoir pris le temps de regarder cette vidéo et je vous souhaite le meilleur !

Appendix B: Study introduction information

The following information was shared/discussed with each participant before they started the pre-test.

Eye tracker

- We will use an eye tracker to track pupil dilations (a sign of cognitive processing)
- The eye tracker will have a camera attached to record your field of view should we need to reference any data
- Once the eye tracker is on and calibrated, try not to move the instrument
- You won't need to come in to do the delayed post-test

Test (referred to as "exercise" for students to minimize the stress associated with the word "test")

- The exercise is eight questions broken up into three types of questions:
    - Draw the arrows
    - Draw the products
    - Predict how the reaction will proceed and explain your reasoning
        - You will be prompted to use chemical factors or experimental values for reasoning questions. Chemical factors are those principles you learned in class, like electronegativity, atom size, inductive effect, etc. Experimental values are just pK_a values (you will be provided with a pK_a table if you request one)
        - You must be as detailed as possible in your "explain your reasoning answers." Without a detailed explanation, I can't properly analyze your data.
        - For these questions, you can click the keyboard button to type out your explanation or use the drawing tablet to write out your responses.
- You will be using a drawing tablet
    - Four buttons were programmed to make the drawing process easier (you can undo/redo and scroll without using your mouse)

Module

- While working through the module, I will share your screen on *Zoom* throughout the process in case I need to reference something from the module.
- After a certain time, I will get you to move on to the last part of the module. Let me know when you finish LO2.
- Your camera/audio won't be on, and I won't be recording through Zoom.

Delayed post-test and demographic questionnaire

- One week later, you will receive an email from me with the exercise. I will just ask you to complete the exercise once more. You can do this on your own time, and you won't have to come in
- Reminder that you must complete these to obtain the full $40.

Appendix C: *Pupil Core* validity test

To validate the Pupil Core's ability to measure pupil diameter regarding cognitive load change, we performed a vigilance task adapted from Klinger (2010). In this test, a participant was presented with a sequence of numbers ranging from 1 to 20. They were informed that the sequence could either progress normally (i.e., count from 1 to 20) or contain errors at specific numbers (6, 12, and/or 18). Each trial could have 0, 1, 2, or 3 errors, and the participant knew when the targets might appear. The participant was asked to quickly press a button when they noticed an error in the sequence. The principle behind this test is that pupil dilations are supposed to be highest when there may be potential errors in the sequence to reflect increased cognitive processing. Rather than looking at the pupil diameter change like Klinger (2010) did, we elected to look at raw pupil diameter. Results for the reference and replicate experiments are shown in **Figure C-1** and **Figure C-2**.

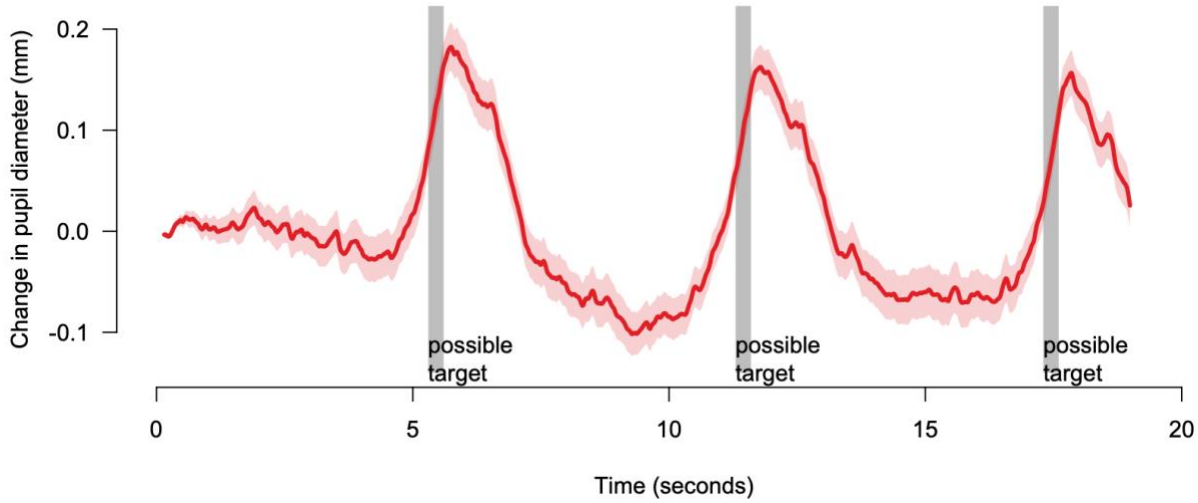


***Figure C-1**. Reference pupil diameter results from Klinger (2010). Regions highlighted in grey represent time points at which an incorrect number in the sequence may be present (6, 12, and/or 18 seconds). Sharp spikes in pupil diameter were observed near regions where errors in the sequence may occur.*

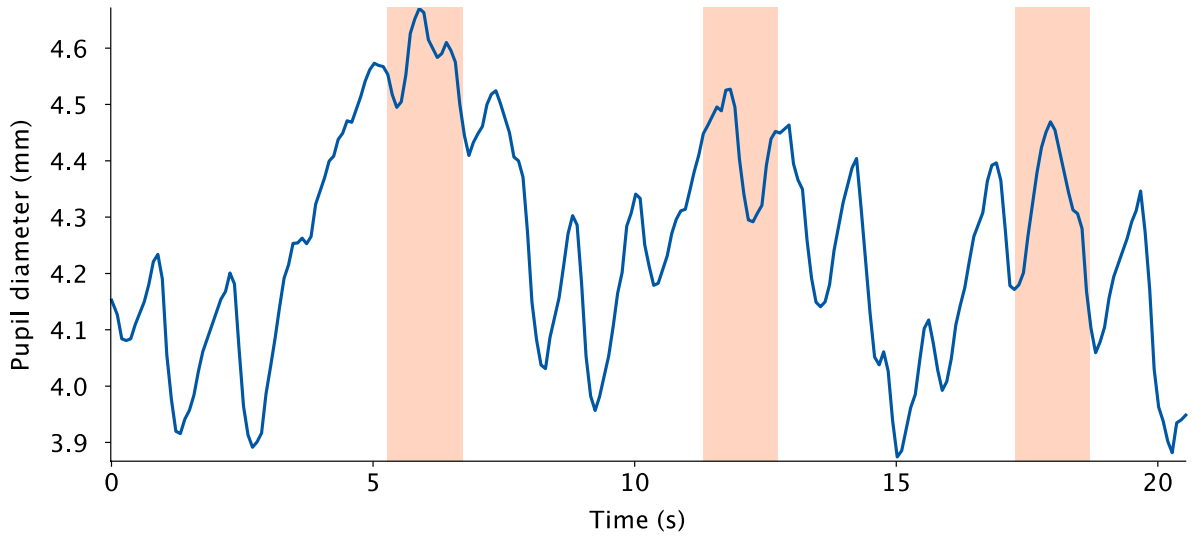


***Figure C-2**. Pupil diameter results obtained using the Pupil Core. Data were obtained from an undergraduate student who did not participate in the study.*

For both figures, we see a similar result, where the peaks of pupil diameters correspond to regions where errors may occur. There seem to be many more turbulent fluctuations than the reference test, likely due to the difference in y-axis units. Overall, we see similar trends: an increase in diameter until the error target is reached, a decrease in diameter, then an increase until the next target is reached.

Appendix D: *Bach Filter* code used to clean and analyze pupil data

The following appendix provides an overview of how this study's pupil data was processed. The *Python* code below will outline steps that involve removing low-confidence data, combining data from the left and right eye, plotting the data and generating measures of interest (e.g., average pupil diameter change, maximum pupil diameter change).

1. The first step is to import necessary libraries for the coding process to occur, such as *pandas* for data manipulation and analysis, matplotlib.pyplot for creating visualizations, numpy for numerical computations, and csv for reading and writing CSV files.

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import csv
```

2. The next step is to read the CSV file containing the blink data generated from *Pupil Player*. My blink files followed the naming convention "*PX_pre/post_blinks.csv*" (e.g., P2_post_blinks.csv for P2's post-test blink data). After that, we assign the contents of the CSV file to a variable named blink_data.

```
blink_data = pd.read_csv("./PX_pre_blinks.csv")
```

3. The following code then reads the '*start_frame_index*' and '*end_frame_index*' columns from the blink_data DataFrame and creates a list called blinks_array containing pairs of start and end frame numbers. It then adds all the frame numbers between each start and end pair to the blinks list.

```
blinks_array = [(start, end) for start, end in
zip(blink_data['start_frame_index'], blink_data['end_frame_index'])]
blinks = []
for start, end in blinks_array:
    for frame_num in range(start, end+1):
        blinks.append(frame_num)
```

4. We then have to read the CSV file containing the actual pupil data. The following code allows us to read the CSV file and assign its contents to a variable named pupil_pd_frame.

```
pupil_pd_frame = pd.read_csv("./P36_post_pupil_positions.csv")
```

5. Before analyzing the data, we must remove any low confidence values or blink data. The *Pupil Capture* program exports confidence interval values with the pupil data, so we can easily remove data that may not be accurate. Another column called “model confidence” also represents the confidence level for the 3D data. In the next set of codes, we assign a confidence interval of 0.99, meaning any data below 99% confidence will be removed. This cut-off may seem like a lot, but the eye tracker generates 250 readings per second, and most of the generated data is high confidence, so in my opinion, it is better to be more strict than lenient. After setting a confidence interval, the data in pupil_pd_frame get filtered by three conditions: confidence > 0.99, model confidence > 0.1 and whether or not it lies in any of the blink intervals from the blinks_array.

```
CONFIDENCE_INTERVAL = 0.99
pupil_pd_frame = pupil_pd_frame[(pupil_pd_frame.confidence >
CONFIDENCE_INTERVAL) & (pupil_pd_frame.model_confidence >= 0.1) &
(~pupil_pd_frame['world_index'].isin(blinks))]
```

6. Because the first few seconds of each recording are often extraneous data, we usually cut this segment when exporting. This action generates timestamps that do not start at zero. To fix this issue, we normalize all data timestamps by subtracting the first data point’s timestamp from all other values.

```
pupil_pd_frame.pupil_timestamp = pupil_pd_frame.pupil_timestamp -
pupil_pd_frame.pupil_timestamp.iloc[0]
```

7. Since *Pupil Capture* produces separate data for the left and right eye, we must average the data from the two eyes at any given time. This process will take a few steps. First, we create two separate DataFrames (one for each eye).

```
eye_groups = pupil_pd_frame.groupby('eye_id')
eye0 = eye_groups.get_group(0)
eye1 = eye_groups.get_group(1)
```

8. We then create an empty list to store each pupil's x- and y-axis values. The x-axis values are for the timestamps and the y-axis values are for the pupil diameter.

```
x_axis = []
y_axis = []
```

9. Because there are extraneous data between questions (e.g., participants scrolling between questions), we cannot just take the start and end points and calculate the average. Instead, we must map the specific sections of data by cross-referencing the data with the world camera recording to determine when each interval starts and ends (e.g., Q1 starts at frame 671 and ends at frame 4880). The following section accomplishes this by assigning each question to a specific HEX code colour and label. HEX codes were used to give us more freedom of colour selection. Using *Python*'s built-in colour labels instead of the HEX codes (e.g., red, blue, green) would also be sufficient.

```
q1start = 671
q1end = 4880
q2start = 2111
q2end = 3295
q3start = 3508
q3end = 4880
lookup_dict = {(q1start, q1end):["#0058A4", "question 1"]}, (q1end+1,
q2start-1):["#AAA9A9", "extraneous"], (q2start,q2end):["#C63D91",
"question 2"], (q2end+1, q3start-1):["#AAA9A9", "extraneous"], (q3start,
q3end):["#F16621", "question 3"]}
```

10. Because the readings alternate between the left and right eye, averaging two data points at any given moment is difficult. To effectively average the data points from each eye, we created equally spaced data points and generated a figure and axis object for plotting. After experimentation, we decided that 3,000 data points divided by the number of generated intervals from the previous step (five in our case) was the most accurate representation. Using fewer data points led to a loss in data, and using more data points led to less smooth plots.

```
NUM_OF_POINTS = 3000 // len(lookup_dict)
total_rows = len(pupil_pd_frame)
step_size = int(total_rows / NUM_OF_POINTS)
fig, ax = plt.subplots()
```

11. After making the new data points, we create empty lists to store the x- and y-axis data for plotting.

```
x_axis, y_axis = [], []
```

12. Now that the new data points have been generated and the x- and y-axes have been created, the process of interpolation (creating new data points based on the magnitude of obtained data points) can begin. The first step is to filter the eye DataFrames by the current interval in lookup_dict to get the x- and y-coordinates for plotting.

```
for lrange, (color, label) in lookup_dict.items():
    temp0 = eye0[(eye0.world_index >= lrange[0]) & (eye0.world_index <=
    lrange[1])]
    temp1 = eye1[(eye1.world_index >= lrange[0]) & (eye1.world_index <=
    lrange[1])]
```

13. We then obtain each eye's earliest and latest timestamp values in the current interval.

```
    a1_min, a1_max = temp0['pupil_timestamp'].min(),
    temp0['pupil_timestamp'].max()
    a2_min, a2_max = temp1['pupil_timestamp'].min(),
    temp1['pupil_timestamp'].max()
```

14. Next, we create a new set of timestamps and diameters that are evenly spaced.

```
    new_a1_x = np.linspace(a1_min, a1_max, NUM_OF_POINTS)
    new_a2_x = np.linspace(a2_min, a2_max, NUM_OF_POINTS)
    new_a1_y = np.interp(new_a1_x, temp0['pupil_timestamp'],
    temp0['diameter_3d'])
    new_a2_y = np.interp(new_a2_x, temp1['pupil_timestamp'],
    temp1['diameter_3d'])
```

15. We then produce the average of both eyes at the given timestamps and subtract the calculated baseline pupil diameter (3.33 mm in the example below).

```
    BASELINE = 3.33
    midx = [np.mean([new_a1_x[i], new_a2_x[i]]) / 60 for i in
    range(NUM_OF_POINTS)] # divide by 60 to get minutes in x-axis
    midy = [np.mean([new_a1_y[i], new_a2_y[i]]) for i in
    range(NUM_OF_POINTS)]
    midy = [x - BASELINE for x in midy]
```

16. We officially created our new data points! But now we need to plot the data, which can be done by assigning data points to the x- and y-axis.

```
    x_axis += midx
    y_axis += midy
```

17. Finally, we plot the summarized pupil diameter measurements.

```
    ax.plot(midx, midy, color=color, label=label)
```

18. We then can calculate some summary statistics from the produced figures, including the maximum diameter, maximum diameter time and mean diameter.

```
max_diameter_time, max_diameter = max(zip(x_axis, y_axis), key=lambda
point: point[1])
mean_diameter = sum(y_axis) / len(y_axis)
print(f"Max Diameter: {max_diameter:.2f} mm, Max Diameter Time:
{max_diameter_time:.2f} s, Mean Diameter: {mean_diameter:.2f} mm")
```

19. If we want to export a CSV with the new data points, we can also use the following code:

```
data = [("Time (s)", "Pupil Diameter (mm)")] + list(zip(x_axis, y_axis))
with open('output.csv', 'w', newline='') as f:
    csv.writer(f).writerows(data)
```

20. I made minor adjustments to the graph's appearance, including changing the legend's location, reorganizing x-axis values, removing the top and right spines, and adding axes labels.

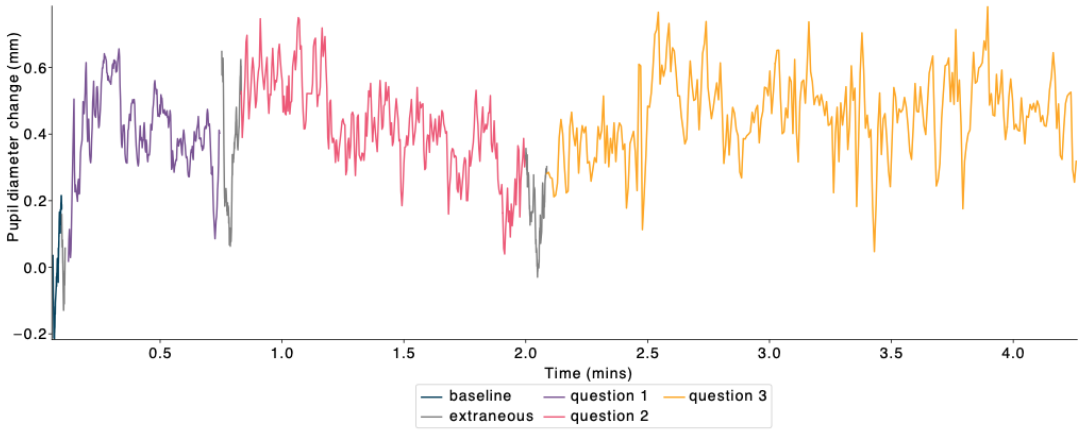
```
plt.legend(ncol=3, loc='lower center', bbox_to_anchor=(0.5,-0.28))
plt.margins(x=0.001)
ax = plt.gca()
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)
ax.set_xlabel('Time (mins)')
ax.set_ylabel('Pupil Diameter Change (mm)')
```

21. Lastly, we save a high-quality image and preview the graph!

```
# save high-quality image
fig.set_size_inches([15, 6])
fig.tight_layout(pad=0.5, h_pad=0, w_pad=0)
plt.savefig('PX_pre.pdf', bbox_inches='tight')
plt.show()
```

Three items will appear after you run the code. These items are described in **Figure D-1**.

1) The graph with the interpolated pupil data (the x-axis is in minutes because the data points were divided by 60 in step 15)



2) The output with the summary statistics

```
Python 3.10.6 (v3.10.6:9c7b4bd164, Aug  1 2022, 17:13:48) [Clang 13.0.0 (clang-1
300.0.29.30)] on darwin
Type "help", "copyright", "credits" or "license()" for more information.

= RESTART: /Users/ahmedyoussef/Desktop/pupil_data/bach_filter remaster copy.py =

= RESTART: /Users/ahmedyoussef/Desktop/pupil_data/bach_filter remaster copy.py =
Max Diameter: 4.67 mm, Max Diameter Time: 0.22 s, Mean Diameter: 4.19 mm
```

3) A CSV file containing the interpolated data and corresponding timestamps

	A	B
1	Time (s)	Diameter (mm)
2	0.00	0.00
...	...	...
3001	688	0.58

***Figure D-1**. The three outputs from the Bach Filter. (1) A plot containing the interpolated data. (2) The summary statistics (maximum diameter, maximum diameter time, and mean diameter). (3) A CSV file containing the interpolated data point values.*

Appendix E: Modes of reasoning coding instructions

The purpose of this appendix is to provide an in-depth explanation of how each case comparison question was assigned a mode of reasoning. At this stage, we will not consider whether a statement is scientifically/chemically accurate or correct. For each response, it was helpful to highlight text and draw a flowchart similar to what was done in the Deng and Flynn (2021) paper. I used the following colours for each component:

- Claim: Purple
- Evidence: Light Blue
- Causal Factor: Red
- Link: Yellow

In all cases, the option circled (A or B) was automatically the claim. Sometimes participants reiterated their claim in the textbox; sometimes, they did not. In some cases, participants provided evidence that almost seemed like a claim (e.g., option A is the better leaving group for Q7 (tetrahedral), or option A is the better nucleophile for Q8 (amine)). Regardless of how similar to a claim the sentence appeared to be, anything unrelated to the correct choice was always considered evidence.

The way the questions were set up, a link between the claim and evidence was not always apparent. A rule of thumb I followed was that if a chemical concept was used to support another present concept, it was considered at least a linear causal mode of reasoning. If there was no chemical concept that supported another one, the answer was relational only if the claim was reiterated in the text and linked to the evidence. **Figure E-1** provides example arguments for each mode of reasoning.

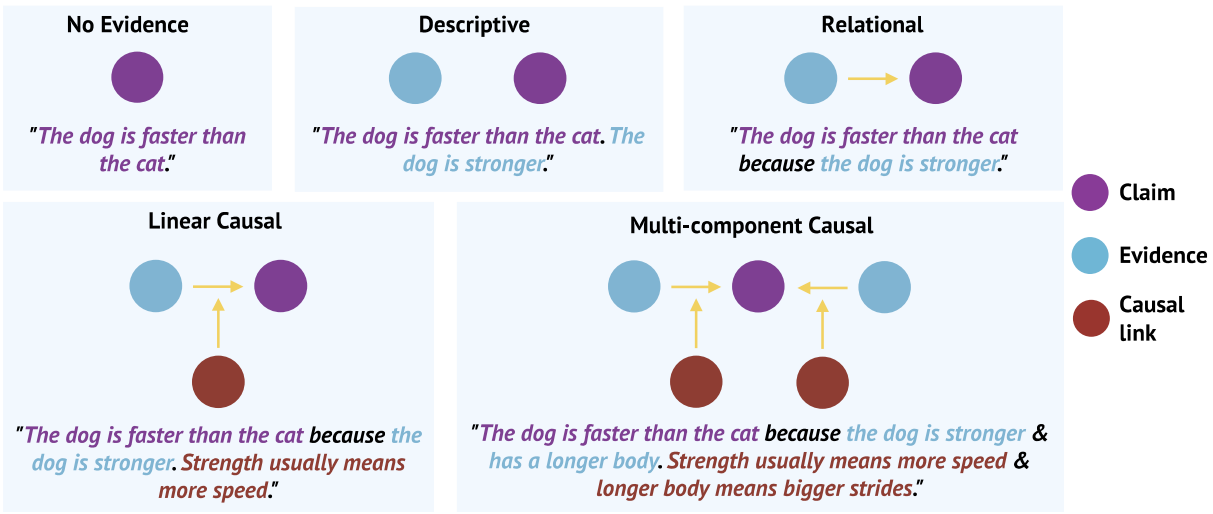


Figure E-1*. The different modes of reasoning. Claims, evidence, and causal links are colour-coded according to the legend on the right. An example is provided for each mode of reasoning.*

Whether an argument is descriptive or relational isn't significantly important to the study, so distinguishing between the two is unimportant.

The four modes of reasoning were written in short-form notation in each textbox:

- **D**: Descriptive
- **R**: Relational
- **LC**: Linear Causal
- **MC**: Multi-component Causal
- **NE**: No evidence (e.g., I picked option A, but I don't know why)

Figure E-2 provides an example of a coded response, considering everything that was mentioned.

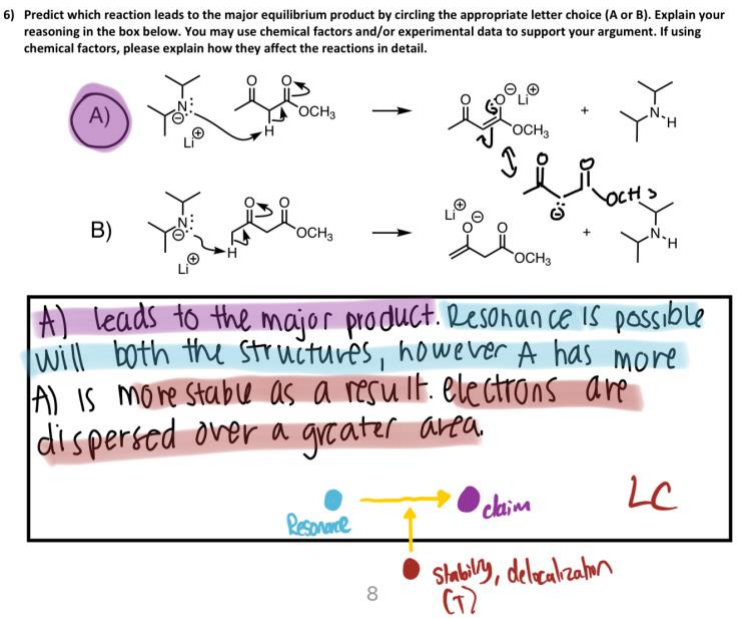


Figure E-2*. Example of coded response from P28 (OrgMech101). The claim is highlighted in purple, the evidence in blue and the causal link in red. LC indicates that the response was coded as "linear causal."*

Updates: from Inter-rater reliability analysis discussions:

- **The order doesn't matter**: participants sometimes listed a causal link before their evidence. Doing this is okay as long as the causal link and evidence can be distinguished.
- **An explanation of a concept is not necessarily a causal link**: If someone says, "A is more electronegative because it has a greater capacity to hold onto its electrons," their follow-up sentence is just a definition of electronegativity. In this case, only code for electronegativity.
- **No in-depth explanations of causal links are needed**: Usually, just mentioning "stability" warrants including it. If more detail/concepts are used to explain a piece of evidence, you can add them all under one causal link bubble separated by commas. We want to capture the participants' thoughts, so detail isn't essential.
- **If two independent arguments are used and have different modes of reasoning, code for the higher mode of reasoning**: For example, a participant may provide an argument using electronegativity and then make mention of more electronegative atoms being able to stabilize a charge better. They then might provide a separate argument about atom size. In this situation, both linear causal and descriptive arguments are used in the same response. In this case, we code for the higher mode of reasoning (linear causal).

Appendix F: Correctness of response coding guidelines

One of three categories of code that can get assigned to a response:

- **Correct (C): argument(s) used in response are scientifically accurate/correct.**
    - *e.g., "The F atoms pull electron density from the carbon, leading to a partially positive change that would destabilize reaction A b/c two positive charges would be near each other."*
- **Partially correct (PC): some (or not all) arguments used in response are scientifically accurate/correct.**
    - *e.g., "F's inductive effect allowing it to stabilize the H. It also is better because the positive charge is further away from the electronegative atom."*
- **Incorrect (I): wrong claim or argument(s) are scientifically incorrect.**
    - *e.g., "I think A is correct b/c CF_3 is very electronegative and can better stabilize the positive charge."*

Identifying correct and incorrect responses should be simple, but partially correct is the most challenging. If an argument is not entirely accurate, it gets coded as PC. Examples of responses that were coded as PC can be found in **Figure F-1** and **Figure F-2**.

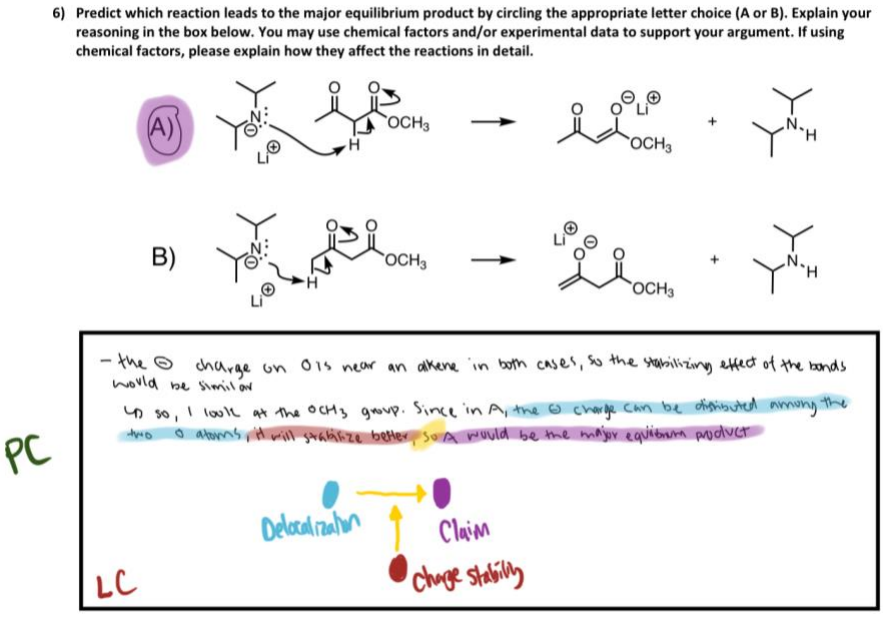


Figure F-1*. Example of coded response from P7 (Acid–Base). PC indicates that the response was coded as partially correct.*

In the former, the response was coded as PC because the negative charge can be distributed among the two O atoms, but the stabilizing effects of the double bonds would not be similar. The latter example was coded as PC because the product in A is more resonance-stabilized, but not easier to form.

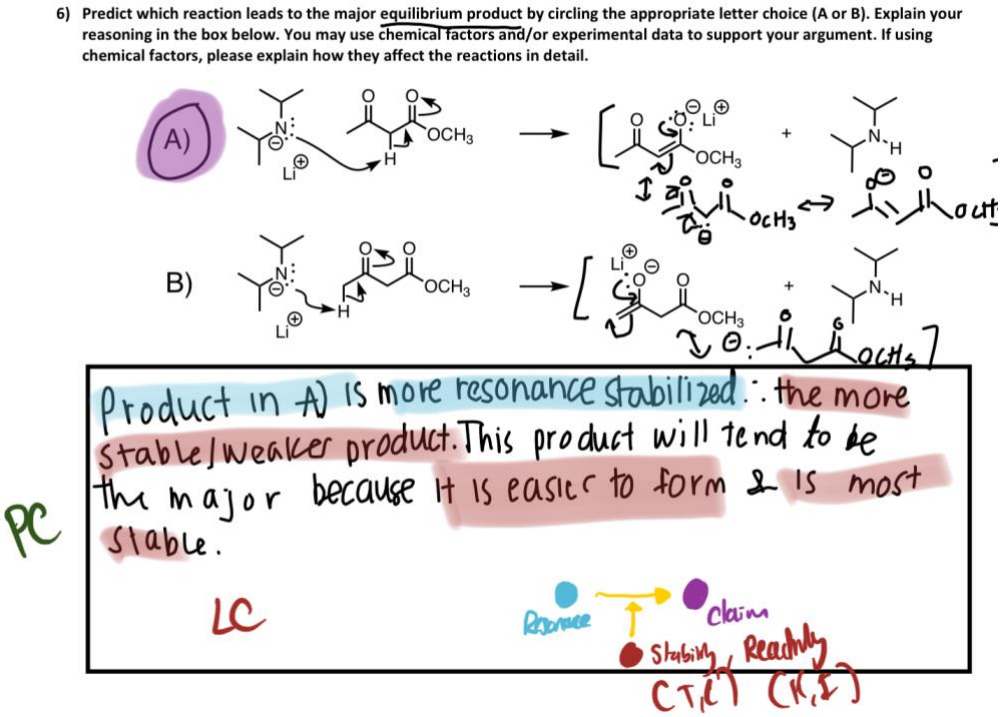


Figure F-2*. Example of coded response from P28 (OrgMech101). PC indicates that the response was coded as partially correct.*