

Exploring students' interpretations of reactions and self-efficacy beliefs in organic chemistry in a
redesigned organic chemistry curriculum

Keith Lapierre

A thesis submitted in partial fulfillment of the requirements for the
Master's degree in Chemistry

Department of Chemistry and Biomolecular Science
Faculty of Science
University of Ottawa

© Keith Lapierre, Ottawa, Canada, 2019

Abstract

Organic Chemistry has been described as a challenging and confusing course for undergraduate students. Novices in the field have been struggling to understand fundamental concepts relating to organic mechanism and organize their knowledge around surface features such as functional groups rather than deep underlying features. At the University of Ottawa, a new “Mechanistic patterns and principles” curriculum was designed and implemented, organized by the underlying mechanistic patterns that govern reactions rather than the traditional surface features approach. The redesigned curriculum emphasizes principles of reactivity in organic chemistry and is organized in a gradient of difficulty. The three studies included within this work act as part of a larger evaluation of the redesigned curriculum, specifically investigating an instrument to assess the formation of expertise in organic chemistry and an instrument to capture self-efficacy beliefs in organic chemistry as students progress throughout the curriculum.

In the first two studies, an open and closed online categorization task was delivered to Organic Chemistry II students at both the beginning and end of their course. The open sort provided insights regarding how participants choose to sort, while the closed sort measured participants ability to categorize reactions according to their underlying mechanistic pattern. In the first study, we provide an in-depth analysis of the changes in expertise that occur with respect to the expertise of their choices and ability. Findings from this work demonstrated a positive shift from students attending to surface to process-oriented features in the open sort, as well as an increase in students’ ability in the closed sort. The following-up study investigates the relationship between the expertise demonstrated by participants in the open and closed sorts. Additionally, this work compares these measures of expertise against various other metrics, including a high-stakes categorization task, and academic performance to increase the validity, and probes at the reliability of findings. Findings from this work demonstrate a strong relationship between the expertise

demonstrate in the online task and academic performance, as well as describe an evolving relationship between the expertise demonstrated in students' choice and ability as they progress throughout the course. While previous work in the evaluation of the curriculum demonstrated that students possess greater ability, it is unknown whether this also translates to an increase in their beliefs about their abilities.

The last study included within this work moves beyond cognitive outcomes of the curriculum to investigating the role of self-efficacy beliefs in the curriculum. Self-efficacy beliefs are defined as an individual's belief in their capability to perform a specific task or objective successfully. This work intends to construct and validate a task-specific, multi-dimensional self-efficacy beliefs instrument for undergraduate students in the domain of organic chemistry. Pre-administration validity evidence, including test content and response process validity, was collected. Data for internal structure validity evidence was collected from a single administration with Organic Chemistry I students ($N=78$) to 7-factor structure within the final 39 item instrument. Due to the small sample size, these results are interpreted with extreme caution. Future work with this instrument aims to improve the validity evidence collected by expanding the sample size and evaluate the influence curriculum on self-efficacy beliefs, and who, based on demographic variables, may be benefiting the most from the transformed curriculum.

Keywords: Chemistry Education Research, Categorization Task, Expertise, Self-Efficacy, Curriculum and instruction.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor, Dr. Alison Flynn, for her continued support, encouragement, and patience throughout my studies. I have benefitted significantly from your knowledge, expertise, constructive feedback, and insightful questions. Your passion for learning is truly inspiring and I cannot thank you enough for acting as my mentor and friend.

I would like to express my deepest appreciation to my thesis committee, Dr. Jeffrey Keillor, and Dr. David Trumpower, for your passionate participation, and the invaluable feedback provided to shape this work.

I am thankful to have been given the opportunity to build on the previous work of Dr. Kelli Galloway and Michael Min Wah Leung. Their extensive work in the development and analysis of the physical categorization task laid the foundation for me to proceed with the online implementation. I would also like to thank all the experts across the fields of chemistry, and education who provided me with their time and unique insights during the construction of the self-efficacy beliefs instrument.

I am indebted to the amazing members of the Flynn Research Group I had the opportunity to work with, especially Amanda, Nik, Myriam, Jacky, Emily, Natalie, and Denzel. I am blessed to have worked with such a supportive group, who continuously dedicated their time to providing me with feedback and helping me work through problems, whether personal or academic.

To my friends and family, I would not have been able to complete this work without your continued encouragement and support. I am grateful to have such an amazing network of people around me, who pushed me to become a better person and researcher.

Lastly, this research would not have been possible without the students who graciously volunteered their time and effort in the participation of the studies.

Table of Contents

ABSTRACT	II
ACKNOWLEDGEMENTS	IV
LIST OF TABLES	VII
LIST OF FIGURES	VIII
LIST OF ABBREVIATIONS:	X
CHAPTER ONE: INTRODUCTION	1
CHAPTER TWO: AN ONLINE CATEGORIZATION TASK TO INVESTIGATE CHANGES IN STUDENTS' INTERPRETATIONS OF ORGANIC CHEMISTRY REACTIONS	4
COPYRIGHT STATEMENT	4
INTRODUCTION	5
TRANSFORMED CURRICULUM: MECHANISTIC PATTERNS AND PRINCIPLES	7
DIFFERENCES BETWEEN NOVICES AND EXPERTS	9
THEORETICAL FRAMEWORK: INFORMATION PROCESSING THEORY	14
RESEARCH QUESTIONS	17
METHODS	17
MODIFICATIONS TO THE CATEGORIZATION TASK AND MIGRATION TO AN ONLINE SETTING	17
INSTRUCTIONAL CONTEXT AND IMPLEMENTATION OF THE ONLINE CATEGORIZATION TASK	18
DATA ANALYSIS: VISUALIZING THE CATEGORIZATIONS WITH GEPHI	19
DATA ANALYSIS: LEVELS OF INTERPRETATION	20
DATA ANALYSIS: ABILITY TO IDENTIFY MECHANISTIC PATTERNS IN A REACTION	22
INTERVIEWS TO INVESTIGATE RESPONSE PROCESS VALIDITY	23
SUMMARY OF EVIDENCE TO SUPPORT THE REPRESENTATIVENESS OF THE POPULATION	24
RESULTS AND DISCUSSION	27
RQ 1: HOW DO INDIVIDUALS' CATEGORIZATIONS QUALITATIVELY COMPARE AT AN EARLY AND A LATE ADMINISTRATION?	27
RQ 2: HOW DO STUDENTS' CHOICES IN ORGANIZING REACTIONS CHANGE BETWEEN AN EARLY AND A LATE ADMINISTRATION IN AN ORGANIC CHEMISTRY II COURSE?	31
RQ 3: HOW DO STUDENTS' ABILITIES TO CATEGORIZE REACTIONS MECHANISTICALLY CHANGE BETWEEN AN EARLY AND A LATE ADMINISTRATION, WHEN CUED TO DO SO?	34
i) INDIVIDUAL PARTICIPANTS' CATEGORIZATION ABILITIES.	34
ii) CHANGES IN PARTICIPANTS' ABILITY TO IDENTIFY MECHANISTIC PATTERNS IN THE REACTIONS, ANALYZED FOR EACH MECHANISM TYPE.	35
iii) PARTICIPANTS' ABILITY TO IDENTIFY MECHANISTIC PATTERNS IN THE REACTIONS, ANALYZING EACH REACTION CARD.	38
LIMITATIONS	40
CONCLUSIONS	41
IMPLICATIONS FOR TEACHING AND LEARNING	43
IMPLICATIONS FOR RESEARCH	44
CHAPTER THREE: INVESTIGATING THE ROLE OF MULTIPLE CATEGORIZATION TASKS IN THE EVALUATION OF A REDESIGNED CURRICULUM	45
INTRODUCTION	45
ASSESSING THE VALIDITY AND RELIABILITY OF AN INSTRUMENT	46
RESEARCH QUESTIONS:	50
METHODS:	50
PARTICIPANTS AND INSTRUCTIONAL CONTEXT:	50
DATA COLLECTION	51

LEVEL OF INTERPRETATION: QUANTIFYING THE EXPERTISE DEMONSTRATED IN HOW A PARTICIPANT CHOOSES TO SORT	52
MATCH WITH EXPERT: QUANTIFYING A PARTICIPANT’S ABILITY TO SORT BY THE MECHANISTIC PATTERN GOVERNING A REACTION.	53
DATA ANALYSIS:	54
RESULTS AND DISCUSSION:	56
RQ1. WHAT RELATIONSHIP, IF ANY, CAN BE OBSERVED BETWEEN CATEGORIZATION TASKS IN VARIED CONTEXTS?	56
RQ2. HOW DO MEASURES OF EXPERTISE DETERMINED THROUGH A CATEGORIZATION TASK RELATE TO ACADEMIC PERFORMANCE?	61
RQ 3. TO WHAT EXTENT ARE FINDINGS FROM VARIOUS CATEGORIZATION TASKS STABLE AND REPRODUCIBLE?	64
LIMITATIONS:	67
CONCLUSIONS:	68
IMPLICATIONS FOR TEACHING AND LEARNING:	69
IMPLICATIONS FOR RESEARCH:	69
<u>CHAPTER FOUR: DEVELOPING AND VALIDATING A SELF-EFFICACY BELIEFS INSTRUMENT FOR UNDERGRADUATE STUDENT IN ORGANIC CHEMISTRY</u>	<u>71</u>
INTRODUCTION:	71
THEORETICAL FRAMEWORK	72
SOURCES OF SELF-EFFICACY BELIEFS	73
RELATED CONSTRUCTS	74
GUIDELINES FOR CONSTRUCTING A SELF-EFFICACY INSTRUMENT	77
MEASURING SELF-EFFICACY BELIEFS IN STEM EDUCATION	79
EXPLORING SELF-EFFICACY BELIEFS IN STEM	86
RESEARCH GOALS:	88
INSTRUMENT DEVELOPMENT:	89
PHASE ONE: INITIAL SCALE DEVELOPMENT AND PILOT STUDY	89
PHASE TWO: ASSESSING THE VALIDITY EVIDENCE RELATING TO THE INTERNAL STRUCTURE	95
PHASE TWO: EXPLORATORY FACTOR ANALYSIS RESULTS	103
LIMITATIONS	110
CONCLUSIONS AND FUTURE WORK	111
<u>CHAPTER FIVE: CONCLUSIONS</u>	<u>113</u>
<u>REFERENCES</u>	<u>117</u>
<u>APPENDICES</u>	<u>129</u>
APPENDIX A: EXPERT SORT OF THE REACTION CARDS BASED ON THE MECHANISTIC PATTERNS IN THE CURRICULUM	129
APPENDIX B: RESPONSE PROCESS VALIDITY INTERVIEW PROTOCOL	132
APPENDIX C: HIGH-STAKES CLOSED-SORT LIKE CATEGORIZATION TASK INCLUDING ON FINAL EXAMINATION (2017 AND 2018)	133
APPENDIX D: ALIGNMENT OF THE SEBOC ITEMS WITH THE 3D-LAP CRITERION	134
APPENDIX E: SELF-EFFICACY BELIEFS IN ORGANIC CHEMISTRY INSTRUMENT	136

List of Tables

TABLE 1.DESRIPTIVE STATISTICS OF CATEGORIZATION TASK COMPLETION TIMES.....	18
TABLE 2. DESCRIPTIVE STATISTICS RELATING TO ACADEMIC PERFORMANCE ON MAJOR ASSESSMENTS.	24
TABLE 3. SUMMARY OF RELEVANT SELF-EFFICACY BELIEFS INSTRUMENT, INCLUDING VALIDITY AND RELIABILITY EVIDENCE COLLETED. MARK (✓) DENOTES EVIDENCE WAS COLLECTED.....	81
TABLE 4.ORGANIC CHEMISTRY SELF-EFFICACY (OCSE) ITEMS, ADAPTED FROM VILLAFANE ET AL. (2016)	85
TABLE 5.PARALLEL ANALYSIS INDICATING TO RETAIN 7 FACTORS	105
TABLE 6. SEBOC FACTOR LOADINGS (FACTOR LOADINGS >0.3 OMITTED). SEBOC ITEMS COLOR CODING WITH HAVE BEEN COLOR CODED WITH INTENDED DIMENSIONS, COGNITIVE, EVERYDAY APPLICATIONS AND EXPERIMENTAL PRACTICES), FACTOR LOADINGS COLOR CODED WITH NEW DIMENSIONS).....	105
TABLE 7. SEBOC INSTRUMENT RELIABILITY EVIDENCE	107
TABLE 8. SEBOC FACTORS DESCRIPTIVE STATISTICS	109
TABLE 9.SEBOC FACTORS CORRELATION MATRIX, CORRELATION <0.1 OMITTED	110

List of Figures

FIGURE 1. KEY FEATURES OF AN EXPERT.....	7
FIGURE 2.OVERVIEW OF THE SECTIONS IN THE MECHANISTIC PATTERNS AND PRINCIPLES CURRICULUM.....	9
FIGURE 3.METHODS OF CATEGORIZING ORGANIC CHEMISTRY REACTIONS (GALLOWAY ET AL., 2019).....	16
FIGURE 4. STUDY’S TIMELINE. OCII = ORGANIC CHEMISTRY II.....	19
FIGURE 5. OVERVIEW OF DATA ANALYSIS.	21
FIGURE 6. DISTRIBUTION OF INTERPRETATION LEVELS OF LONGITUDINAL AND NON-LONGITUDINAL PARTICIPANTS, SHOWN AS PERCENTAGE OF CARDS SORTED PER LEVEL OF INTERPRETATION. LONGITUDINAL PARTICIPANTS N = 600 CARDS, NON-LONGITUDINAL EARLY ADMINISTRATION N= 1250 CARD, NON-LONGITUDINAL LATE ADMINISTRATION N =500	26
FIGURE 7.DISTRIBUTION OF PARTICIPANTS ABILITY TO CATEGORIZE REACTIONS BY MECHANISM TYPE. LONGITUDINAL PARTICIPANTS N = 24 PARTICIPANTS, NON-LONGITUDINAL EARLY ADMINISTRATION N = 50 PARTICIPANTS, NON-LONGITUDINAL LATE ADMINISTRATION N = 20 PARTICIPANTS.....	27
FIGURE 8. GEPHI VISUALIZATIONS OF PARTICIPANTS’ REACTION GROUPINGS AT EARLY AND LATE ADMINISTRATIONS. (N = 24).	28
FIGURE 9. COMMONLY OBSERVED GROUPS THAT LACKED MECHANISTIC SIMILARITIES.....	30
FIGURE 10. COMMONLY OBSERVED GROUPS REFLECTIVE OF THE EXPERT CATEGORIZATIONS.....	30
FIGURE 11. CHANGES IN INTERPRETATION LEVELS FROM EARLY TO LATE ADMINISTRATION, SHOWN AS PERCENTAGE OF CARDS SORTED PER LEVEL OF INTERPRETATION. N = 600 CARDS SORTED.	32
FIGURE 12. CHANGES IN PARTICIPANTS' OVERALL ABILITY TO CATEGORIZE REACTIONS BY MECHANISM TYPE. (N = 24.)	35
FIGURE 13. CHANGES IN PARTICIPANTS' ABILITY TO IDENTIFY MECHANISTIC PATTERNS IN THE REACTIONS, ANALYZED BY MECHANISTIC PATTERN. N = 24,	36
FIGURE 14. REPRESENTATIVE REACTIONS CARDS THAT HIGHLIGHT AREAS OF HIGH AND LOW SUCCESS, N = 24.	38
FIGURE 15. KEY MECHANISTIC STEPS OF A NUCLEOPHILE REACTING WITH A π ELECTROPHILE BEARING A LEAVING GROUP.	40
FIGURE 16. SOURCES OF VALIDITY EVIDENCE	47
FIGURE 17. STUDY TIMELINE.....	52
FIGURE 18. RELATIONSHIP BETWEEN HOW PARTICIPANT CHOSE TO SORT (LEVEL OF INTERPRETATION) AND THEIR ABILITY TO SORT ACCORDING TO THE DEEP FEATURES (MATCH WITH EXPERT) AT BOTH EARLY (N = 65) AND LATE (N = 43) ADMINISTRATION	58
FIGURE 19.RELATIONSHIP BETWEEN CUED ABILITY IN A LOW-STAKES ONLINE CATEGORIZATION TASK AND HIGH-STAKE FINAL EXAMINATION CATEGORIZATION TASK. (N=43), PEARSON’S R = -.29, P=.06.....	59
FIGURE 20. MAJOR TRENDS OBSERVED BETWEEN PARTICIPANT RESPONSES OF A HIGH AND LOW STAKES CATEGORIZATION TASK. STAR IN THE FIGURE DENOTES CATEGORIES WITH SIGNIFICANT DIFFERENCES, $p < .05$ BASED ON CHI-SQUARED. CATEGORIES POSSESSING <5% DISTRIBUTION, AND “OTHER” CATEGORY FROM THE LOW STAKES SORT WERE OMITTED FROM THE FIGURE FOR SIMPLICITY.	60
FIGURE 21. CONCURRENT RELATIONSHIP BETWEEN MAJOR ASSESSMENT SCORE AND A) LEVEL OF INTERPRETATION IN THE ONLINE OPEN SORT AND B) MATCH WITH EXPERT SORT IN THE ONLINE CLOSED SORT AT BOTH THE EARLY ADMINISTRATION (N = 65) AND LATE ADMINISTRATION (N = 43).	62
FIGURE 22. PREDICTIVE RELATIONSHIP BETWEEN MAJOR ASSESSMENT SCORE AND A) LEVEL OF INTERPRETATION IN THE ONLINE OPEN SORT AND B) MATCH WITH EXPERT SORT IN THE ONLINE CLOSED SORT AT BOTH THE EARLY ADMINISTRATION (N = 65).....	63
FIGURE 23. REPRODUCIBILITY OF DISTRIBUTION OF DATA FROM THE ONLINE CATEGORIZATION TASK, + INDICATING MEAN.....	65
FIGURE 24. TEST-RETEST RELIABILITY OF 2017 AND 2018 GROUPS, USING PAIRS MATCHED THROUGH FINAL EXAMINATION GRADE Z-SCORES, WITH SIGNIFICANT CORRELATIONS DENOTED	66
FIGURE 25. STABILITY OF TRENDS OBSERVED IN A FINAL EXAMINATION CATEGORIZATION TASK. A CHECKMARK IN THE FIGURE DENOTES CORRECT MECHANISTIC CATEGORY. NO SIGNIFICANT DIFFERENCES, $p < .05$ BASED ON CHI-SQUARED.	67
FIGURE 26. FOUR CONTRIBUTING SOURCES OF SELF-EFFICACY BELIEFS	73
FIGURE 27.RECIPROCAL CAUSATION MODEL WITH SNOWBALL EFFECT.	85
FIGURE 28. THREE PHASES OF DEVELOPMENT FOR SEBOC INSTRUMENT, OC = ORGANIC CHEMISTRY	89
FIGURE 30. EXPERT POPULATIONS SELF-IDENTIFIED AREA(S) OF EXPERTISE.....	93

FIGURE 31. ESSENTIAL STEPS DURING EXPLORATORY FACTOR ANALYSIS.....	98
FIGURE 32. SCREE PLOT INDICATING TO RETAIN 2 OR 7 FACTORS	104
FIGURE 33. PERCENTAGE OF SELF-EFFICACY BELIEFS IN ORGANIC CHEMISTRY INSTRUMENT ITEMS EXPLICITLY ASSOCIATED WITH EACH SCIENCE PRACTICE CRITERION AS DEFINED BY THE 3D-LAP	134
FIGURE 34. PERCENTAGE OF SELF-EFFICACY BELIEFS IN ORGANIC CHEMISTRY INSTRUMENT ITEMS EXPLICITLY ASSOCIATED WITH EACH CROSSCUTTING CONCEPT CRITERION AS DEFINED BY 3D-LAP	135

List of Abbreviations:

3D-LAP	Three-Dimensional Learning Assessment Protocol
95% CI	95% Confidence Interval
ANOVA	Analysis of Variance
Avg.	Average
C=C	A double bond between two carbon atoms
C=O	A double bond between a carbon atoms and oxygen atom
χ^2	Chi-squared
CAEQ	Chemistry attitudes and experiences questionnaire
CBI	Comparison-Based Index
CBSEI	College biology self-efficacy instrument
CCSS	College chemistry self-efficacy scale
CFA	Confirmatory Factor Analysis
<i>d</i>	Cohen's <i>d</i> effect size measure
DMF	Dimethylformamide
EFA	Exploratory Factor Analysis
Elec.	Electrophile
EtOH	Ethanol
EWG	Electron Withdrawing Group
HCA	Hierarchal Cluster Analysis
IR	Infrared
KMO	Kaiser-Meyer-Olkin measure of sampling adequacy
LG	Leaving group
M	Mean
<i>Mdn</i>	Median
Me	A methyl group (<i>i.e.</i> , CH ₃)
MeCN	Acetonitrile
Min.	Minutes
MSLQ	Motivated learning strategies questionnaire
NGSS	Next Generation Science Standards
Nuc.	Nucleophile
OC	Organic chemistry
OCSE	Organic chemistry self-efficacy
OR	Odds ratio
PAF	Principal Axis Factoring
PCA	Principal Component Analysis
Ph	Phenyl

<i>r</i>	Pearson's R value
R	A general alkyl group
RQ	Research question
SD	Standard deviation
SEB	Self-efficacy beliefs
SEBOC	Self-efficacy beliefs in organic chemistry
SPSS	Statistical Package for the Social Sciences
STEM	Science, technology, engineering, or mathematics
THF	Tetrahydrofuran
<i>W</i>	Test statistic for a Wilcoxon Signed Rank test

Chapter One: Introduction

Science education at the tertiary level plays an essential role in equipping students with knowledge and skills not only for their careers, but to become better problem solvers, communicators, and global citizens (National Research Council, 2012). Introductory Organic Chemistry courses have been referred to as a gatekeeper course for students continuing in science, technology, engineering, or mathematics (STEM) programs making it an area of particular interest (Fischer, Zhou, Rodriguez, Warschauer, & King, 2019). Traditionally, Organic Chemistry curriculums have been organized using a functional group approach, *i.e.*, by features of specific atoms or groups of atoms. With each functional group being about to proceed through multiple different reactions, this approach has learners in the domain of organic chemistry describe the course as challenging and perceive core concepts such as the symbolism of organic chemistry, organic mechanisms and synthesis difficult to learn (Anderson & Bodner, 2008; Bodé & Flynn, 2016; Galloway, Stoyanovich, & Flynn, 2017; O'Dwyer & Childs, 2017). At the University of Ottawa, a redesigned "Mechanistic patterns and principles" organic chemistry curriculum has been implemented organized around the underlying patterns in reactivity rather than similarities in surface features (Flynn & Ogilvie, 2015). This redesigned organic chemistry curriculum aims to address the challenges previously described by better equipping student with the tools to understand the principles of reactivity, solve for unknown reaction mechanism, transfer these skills to other disciplines such as biochemistry.

Discipline-Based Education Research, such as Chemistry Education Research, plays an essential role in providing evidence-based practices towards improving teaching and learning in the context of a specific field (National Research Council, 2012). Using three studies, this work explores the role of an design-based research approach (Cobb, Confrey, DiSessa, Lehrer, &

Schauble, 2003) as a means of evaluating the University of Ottawa's redesigned Organic Chemistry curriculum. This work aims to produce transferable instruments to enable practitioners and instructors better to evaluate their own organic chemistry curriculum. Although each study acts as a piece of a broader curriculum evaluation, they each possess unique goals requiring individual in-depth introductions and methodological approaches described in their respective chapters.

Chapter two and three of this work highlight studies investigating the utility of an online categorization tool as an approach for evaluating expertise in organic chemistry. Chapter two is focused on evaluating the influence of the redesigned Organic Chemistry curriculum on student's expertise through a longitudinal design. This study was recently published in the *Journal of Research for Science Teaching* but has been adapted to provide additional background information and context where necessary. This work used the online categorization task used to investigating how learners' interpretations of organic chemistry reactions in the redesigned curriculum change over time with respect to how they chose to organize their knowledge around organic chemistry reaction, as well as their ability to organize their knowledge around the underlying deep features of these reactions.

The study captured in Chapter three builds on the previous chapter expanding on the online categorization task, with a focus on evaluating the validity and reliability of the instrument. This work implements multiple categorization tasks with varied instructional contexts, and stakes to investigate the relationship between the expertise demonstrated in these tasks, as well as the establishing the validity by investigating the relationship between other metrics of expertise, specifically major assessment grades. The Reliability of the instrument is evaluated through a test-

retest model of similar samples (*i.e.*, Organic Chemistry II students), as well as an item-based approach (Arjoon, Xu, & Lewis, 2013; Brandriet & Bretz, 2014).

While the studies in Chapters two and three investigated a tool to evaluate domain-specific knowledge and skills as a result of the curriculum, the last study in Chapter four focuses on the development of an instrument to capture students' self-efficacy beliefs, *i.e.*, their beliefs in their ability to complete a given task in organic chemistry. (Bandura, 1997). This work establishes pre-administration validity evidence for the Self-Efficacy Beliefs in Organic Chemistry (SEBOC) instrument, as well as explores the internal structure of the instrument using factor analysis. Though chemistry self-efficacy instruments exist (Dalgety, Coll, & Jones, 2003; Uzuntiryaki & Aydin, 2009; Zusho, Pintrich, & Coppola, 2003), this work aims to construct a valid and reliable organic chemistry specific self-efficacy beliefs instrument with a strong emphasis on reactivity. Together, these instruments will better equip researchers to evaluate how the transformed curriculum influences desired content outcomes (*i.e.*, changes to expertise in organic chemistry), and non-content outcomes (*i.e.*, changes in students' self-efficacy beliefs about organic chemistry).

Chapter Two: An online categorization task to investigate changes in students' interpretations of organic chemistry reactions

The following chapter was adapted from the manuscript published in the *Journal of Research and Science Teaching*.

Copyright Statement

Republished (adapted) with permission from John Wiley and Sons Lapierre, An online categorization task to investigate changes in students' interpretations of organic chemistry reaction, *Journal of Research in Science Teaching*, Lapierre, K. R., & Flynn A. B *ASAP*. DOI:10.1002/tea.21586 - Copyright 2019; permission conveyed through Copyright Clearance Center, Inc.

Introduction

Science curricula that are structured by surface features do not provide opportunities to connect concepts and build expert-like understanding (Cooper & Stowe, 2018). Internationally, efforts are being made toward curricular reform, especially in high school and general chemistry (Cooper & Klymkowsky, 2013; Cooper, Stowe, Crandell, & Klymkowsky, 2019; National Research Council, 2012; Sevian & Talanquer, 2014; Talanquer & Pollard, 2010, 2017). In organic chemistry, curricula have historically been organized by surface features (*i.e.*, functional groups) and assessments have rarely asked students to make mechanistic connections across a variety of reaction types. At the University of Ottawa, a new “Mechanistic patterns and principles” curriculum was designed and implemented, organized by the underlying mechanistic patterns that govern reactions rather than surface features (Flynn & Ogilvie, 2015). The curriculum emphasizes principles of reactivity in organic chemistry and is organized in a gradient of difficulty. Previous work has revealed higher scores in the new context on familiar and unfamiliar questions related to reactivity (Webber & Flynn, 2018); however, that study did not reveal how students were organizing their organic chemistry knowledge. Subsequent qualitative studies revealed the variety of ways in which students and experts chose to organize their knowledge in organic chemistry at a single time point (Galloway, Leung, & Flynn, 2018, 2019). In the present study, we used an online categorization task to investigate what similarities students perceived between organic chemistry reactions in the context of a redesigned organic chemistry curriculum. We also investigated how those categorization choices and abilities changed throughout an organic chemistry course, reflective of changes in how students were organizing their knowledge. Generally, experts are characterized as having acquired more knowledge than novices (National Research Council, 2000). Chi has described several key differences in how experts both excel and fall short (Chi, 2006). Among these differences, expertise in a domain is often attributed to how

experts recognize meaningful patterns and organize their knowledge (Figure 1). When experts are faced with a problem, they often recognize meaningful patterns or features overlooked by novices. These features are typically representative of the underlying, deep feature in the problem (Chi, Feltovich, & Glaser, 1981b). Additionally, experts are more likely to organize their knowledge around big ideas and concepts, demonstrating a better understanding of the domain. This organization of knowledge or knowledge-structure becomes increasingly interconnected as expertise grows (Acton, Johnson, & Goldsmith, 1994). Additionally, research in expertise has demonstrated that experts in a domain possess conditionalized knowledge. Conditionalized knowledge is described as an expert's ability to identify the specific subset of knowledge required to answer a problem, of their extensive body of knowledge. Lastly, experts possess effortless retrieval of relevant information when compared to novices, and are fluently able to adapt their knowledge to new situations (Chi, Glaser, Farr, & Chi author, 1988; National Research Council, 2000). Foundational work by Chi and Glaser in physics education used a categorization task to provide insight into the differences in expertise, comparing eight PhD students in physics against eight undergraduate students in physics (Chi et al., 1981b). In that work, they illustrated a novice's tendency to focus surface features such as keywords that are presented when categorizing problems (e.g., problem is attributed to an inclined plane), whereas experts possessed a more interconnected knowledge structure organized by deeper conceptual features and recognized patterns attributed to said deep features such as the underlying principles presented in the problem (e.g., problem is attributed to conservation of momentum). We extend that work in an organic chemistry context in the present study.

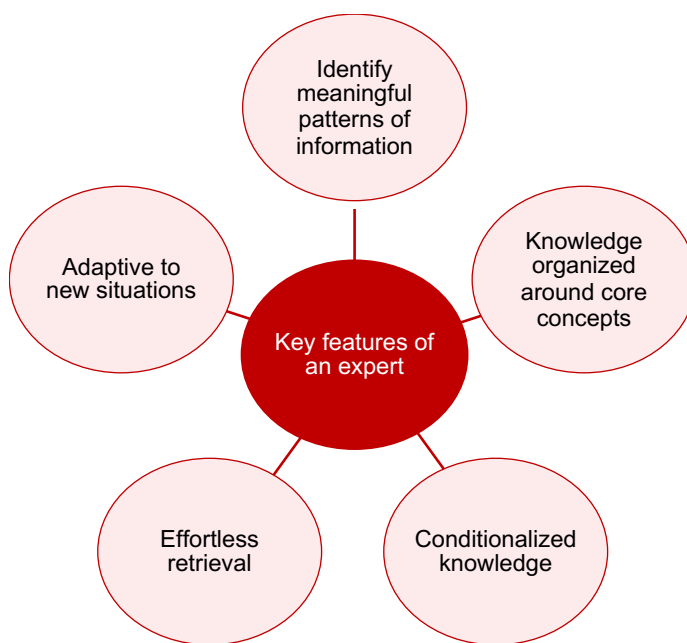


FIGURE 1. KEY FEATURES OF AN EXPERT

Transformed Curriculum: Mechanistic patterns and principles

Organic Chemistry serves as a prerequisite course for many undergraduate programs across STEM and is taken by students with wide-ranging career goals. The course has been described as a difficult and confusing course for undergraduate students (Anderson & Bodner, 2008; Bradley, Ulrich, Jones, & Jones, 2002) and often has high attrition rates (Grove, Hershberger, & Bretz, 2008; Rowe, 1983). Novices in the domain of organic chemistry may initially struggle with the fundamentals of chemistry and yet can become proficient, including in areas of interpreting functionality of chemical structures (Bodé, Caron, & Flynn, 2016; Flynn et al., 2014), placing meaning to the symbolic language of organic chemistry (*i.e.*, the electron-pushing formalism) (Flynn & Featherstone, 2017; Galloway et al., 2017), and organizing their knowledge around the deep underlying concepts of organic reactions (Galloway, Leung, & Flynn, 2018b; Galloway et al., 2019). Moreover, students' understanding can be based on surface level features rather than process or reactivity-oriented concepts (Cruz-Ramírez De Arellano, Towns, Cruz, & Towns, 2014). While many factors contribute to how learners develop expertise in a discipline, the

curriculum is one of the foundational components for achieving high-quality learning outcomes (Stabback, 2016).

Typically, organic chemistry curricula are arranged according to the features of specific atoms or groups of atoms, referred to as functional groups (Clayden, Greeves, Warren, & Wothers, 2012). This organization stems from historical reasons when compounds and reactions were investigated and identified by testing for the presence of these functional groups. Since instrumentation and experimental techniques have improved, chemists can now deeply investigate the reactions' underlying chemical mechanisms. This functional group organization does not reflect modern approaches to chemistry and can pose a learning barrier, since each functional group can undergo many different types of reactions, with different mechanisms. Therefore, students who learn in a functional group-based curriculum will understandably *choose* to see similarities based on surface features (Graulich & Bhattacharyya, 2017).

The University of Ottawa implemented a transformed curriculum using a mechanistic approach, diverging from the traditional approach (Flynn & Ogilvie, 2015). In this new curriculum, “Mechanistic patterns and principles”, reactions are organized around the underlying patterns of mechanisms governing the reactions rather than by functional group (Figure 2). The overarching aim of this organization is to focus on principles of reactivity, such as the structure, properties, and reactivity of nucleophiles and electrophiles (the electron donor and acceptor in bond formation, respectively). Reaction mechanisms are then taught with gradient difficulty over two semesters, building off simple mechanisms such as those in acid–base reactions before encountering more complex ones. Ideally, students become better able to explain the chemical basis for reaction mechanisms (*i.e.*, causal mechanistic reasoning), analyze competing reaction pathways, predict the reactivity of unknown reactions, and transfer these principles of reactivity to other areas and

disciplines (Flynn & Ogilvie, 2015). A curriculum evaluation is underway using a design-based research approach (Cobb et al., 2003), in which studies have been performed to assess students' knowledge and skills in the context of the transformed curriculum (Galloway et al., 2018, 2019, 2017; Webber & Flynn, 2018). The present study extends this evaluation, investigating how students chose and were able to organize their knowledge around organic chemistry reactions and how this organization changed over time as a reflection of aspects of their expertise in organic chemistry.

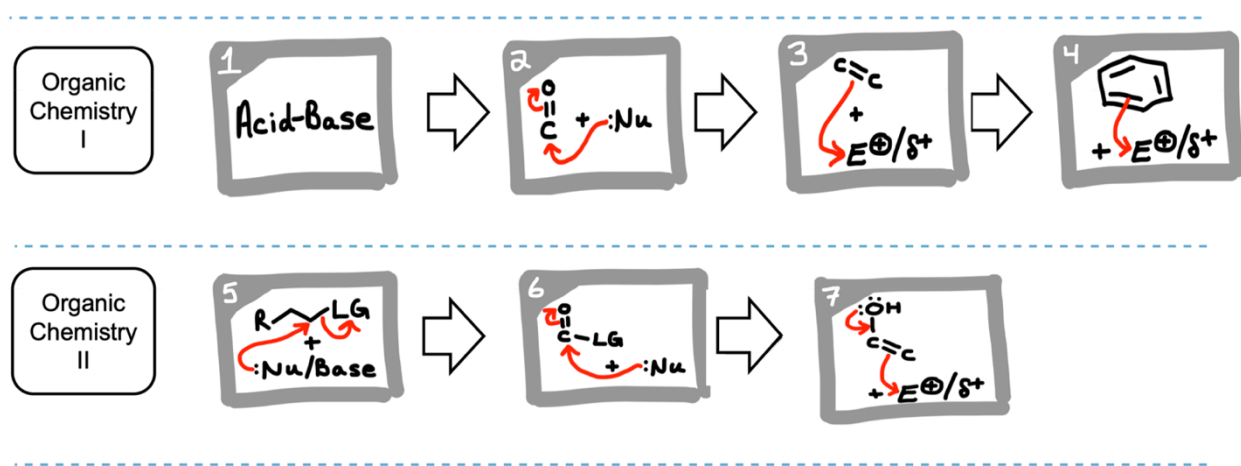


FIGURE 2. OVERVIEW OF THE SECTIONS IN THE MECHANISTIC PATTERNS AND PRINCIPLES CURRICULUM

Differences between Novices and Experts

Multiple different instruments and approaches have been developed to investigate the differences between novices and experts, including similarity ratings (Boster & Johnson, 1989), think-aloud interviews (Camacho & Good, 1989; Randles & Overton, 2015), and categorization (*i.e.*, card sorting) tasks.

Categorization tasks can take open or closed sort forms. An open (*i.e.*, unframed) sort is a categorization task used to investigate a participant's organization of knowledge and illustrates the connections a participant naturally chooses to see between items (*e.g.*, images on cards, reactions, questions). In a closed (*i.e.*, framed) sort, a participant is asked to sort items into a set of pre-determined categories. Studies using a closed sort task often use categories that aim to explicitly

cue participants to hypothesized deep features (Krieter, Julius, Tanner, Bush, & Scott, 2016; J. I. Smith et al., 2013). These studies conclusively demonstrated the utility of pre-determined categories for increasing a participant's probability of sorting based on these deep features in both novices and experts and gave insights into a participant's ability to sort in a given way (Krieter et al., 2016; J. I. Smith et al., 2013). Studies have also used an open sort in conjunction with a closed sort (or framed sort) to investigate a participant's choices and preferences and their ability to sort in a given way (Bissonnette et al., 2017; Krieter et al., 2016; J. I. Smith et al., 2013).

Categorization tasks have been used to probe the difference between experts and novices in many different disciplines of education research, including physics (Chi, Feltovich, & Glaser, 1981; Lin & Singh, 2010; Mason & Singh, 2011, 2016; Singh, 2009), biology (Bissonnette et al., 2017; Hoskinson, Maher, Bekkering, & Ebert-May, 2017; J. I. Smith et al., 2013; M. U. Smith, 1992), computer sciences (McCauley et al., 2005), and chemistry (Domin, Al-Masum, & Mensah, 2008; Graulich & Bhattacharyya, 2017; Irby et al., 2016; Kozma & Russell, 1997; Krieter et al., 2016; Stains & Talanquer, 2008). These studies may employ one-on-one interviews (Chi et al., 1981; Galloway et al., 2018; Snyder, 2000), but can also be administered in larger class settings (Krieter et al., 2016; Lin & Singh, 2010; Mason & Singh, 2011, 2016; J. I. Smith et al., 2013).

A recent study in the domain of physics education by Mason & Singh (2016) aimed to recreate the original Chi et al. (1981) study investigating the difference between novices and experts on a larger scale by increasing the number of participants from 8 novices and experts to 180 novices, 7 PhD students, and an additional 7 PhD faculty members. Similar to the original study, the participants were asked to perform an open sort of 25 problems based on the problems' similar features and provide a written sort explanation as to why they chose to sort them like this. The study differed from the original in the way the participants' answers were quantified. The study

separated categories based on "good" answers if the participant used underlying principles, "moderate" answers if the participant used underlying principles but not explicitly, and "poor" answers if the participant used surface features. Three physics professors determined what would constitute a good, moderate, or poor answer from a small sample of novices' answers. From this execution, the study concluded the same results hypothesized from the Chi study – still showcasing that novices rely heavily on surface features. The most notable finding from this study though was a large amount of overlap in good answers produced by the novices and the PhD students. The authors claim this overlap indicates that the novices were at a much higher level than what would be expected. Additionally, this work suggests that categorizations tasks may be a useful in-class activity to aid novices in learning the underlying principles. The authors recommend incorporating them into group activities to generate discussions amongst diverse levels of expertise in the classroom around why a specific categorization method might be more appropriate than others.

Similar to how the Mason & Singh study included participants with a range of education (from novice to PhD candidate to faculty PhD), work by Stains & Talanquer (2008) focused on the connections that students with varying levels of chemistry education (from first-year general chemistry to Chemistry PhD students) see between two different chemical representations. During a one-hour semi-structured interview process, students were presented with an open sort of nine symbolically represented chemical reactions and were asked to categorize them, followed by six microscopically represented reactions that they were then asked to place into their created categories. The results of this study showed that students in a general chemistry course would be more likely to generate categories with greater variety. The study also found that novices were more likely to categorize problems based on concrete/explicit features of the problems, whereas experts sorted based on implicit/deep features. An unintended consequence of the implementation

design led to an unanticipated sorting strategy by novice students who attempted to match microscopic representation with the symbolic rather than categorize. Overall, this work by Stains et al. (2008) effectively demonstrated how various individuals with levels of expertise organize their knowledge around different representations and highlighted limitations around the implementation strategy.

Krieter et al. (2016) used a quantitative approach in their effort to differentiate between the novices and experts' categorizations. The categorization task used intended to measure the level of conceptual expertise and difference within a population of 318 students with varying levels of chemistry education. Of this population, 162 students were non-chemistry students who were considered novices, and 31 full-time chemistry faculty members were considered experts. The authors noted that the select group of experts presents a limitation, as the faculty who participate in the study might not be representative of the level of expertise found at other institutions. A set of 16 cards were developed, each varying orthogonally in 2 dimensions: one dimension examines one of four hypothesized surface features, the other dimension examines one of four content domains. In context, if a participant generates a grouping four cards sharing a common the hypothesized deep feature, such as equilibrium, each card within this grouping will possess a unique hypothesized surface feature. The task first asked individuals to perform an open sort, and then a closed sort where students placed the same set of cards into one of four hypothesized deep feature categories.

The results of this study explored measures that required hypothesized sorts, as well as explored quantitative methods that did not require a hypothesized pairing. When exploring against a surface and deep hypothesized sort, the authors used measures of edit distance, and percent pairings. Edit distance captures the minimum number of card moves required before producing a

hypothesized sort, where the fewer moves required indicates the participant sorts are more like the hypothesized sort. Percent pairing describes the number of card pairings aligned with a hypothesized sort. In this approach, the higher the percent pairing, the more aligned with the hypothesized sort. From the results of the edit distance and percent pairings, the participants sorts demonstrated the difference of expertise throughout various stages of education quantitatively. When exploring different quantitative methods that did not require a hypothesized sort, the authors proposed the novel comparison-based index (CBI) and used a more traditional hierarchical cluster analysis (HCA) to develop a visual dendrogram. The CBI was meant to act as an empirical measure, that would compare a single sort against the distribution of both experts and novices. The CBI value is developed by subtracting the pairing frequency matrix of the novice sorts from the expert sorts. A CBI value close to zero indicated that the novices and experts' sorts were alike, a positive value indicated a card pairing was made more frequently by experts, and a negative indicated a pairing was more likely to occur by novices. The analysis of the CBI showed a gradual increase in the level of expertise between stages. The comparison between the results of the open and closed sort using the CBI values showed that when prompted, individuals can identify the deep features in the problem more effectively, showing more expert-like thinking. Through the hierarchical cluster analysis of the open sort, the authors observed that the categorizations created by experts would almost perfectly reflect the hypothesized deep categories even though a hypothesized deep sort does not inform the analysis. Contrastingly, the hierarchical cluster analysis of the novices did not reflect the hypothesized surface categories. The application of the above scoring methods provided some key limitations with respect to the edit distance and percent pairing: 1) these methods do not capture the qualitative differences in categorization methods used by students in the open sorts, and 2) these methods do not identify which concepts or features were

misaligned in the sort, 3) the closed sort method used in this work lacked an "unknown" category, which may force students who do not have a basis of which the correct grouping is to sort without a real rational or understanding, essentially guessing.

Related work in the domain of chemistry by Graulich and Bhattacharya (2017) used a series of categorization tasks revolving around differently substituted alkene molecules (*i.e.*, a double bond between two carbon atoms) to investigate student perceptions. The initial categorization task prompted students to sort the reaction cards however they deemed fit. A later categorization task asked students to complete the same task but now with a new set of cards containing variations on the starting materials. They found that students were likely to focus on the perceptually shared attributes, with a strong focus on the functionality displayed on the reaction cards. The findings of this work are consistent with those of Chi et al. (1981), which found that novices tended to attend to surface features of a problem rather than the conceptual or deep features.

Theoretical framework: Information processing theory

Modern information processing theory describes cognitive processes involved in learning and knowledge organization (Schunk, 2016). Information received from an external stimulus (whether audio, visual, etc.) is transferred into working memory. Working memory contains a very limited capacity for information and processes the information for encoding into the long-term memory. While information is in working memory, relevant prior knowledge is retrieved from the long-term memory for use in the working memory. If a meaningful connection is generated between the prior knowledge and the new information, it can be encoded into long-term memory. Once this information is encoded into long-term memory, it is available for retrieval to develop new connections (Johnstone, 2006, 2010; Mayer, 2012; Schunk, 2016). If no meaningful connections are made but the learner believes the information to be important, the information can

be entered into the long-term memory with rote learning and is referred to as unconnected. This type of information will be difficult to recall and develop new connections from (Johnstone, 2006). This theory is adapted from the classic version of information process theory that viewed learning as information acquisition, rather than knowledge construction (Mayer, 2012). A computer metaphor inspired the classic view with learners considered as an empty vessels to receive knowledge with no regard for the mechanism in which learners incorporate new information with prior knowledge (Mayer, 2012). For this study, both the reaction cards and pre-determined categories act as retrieval cues aiming to elicit participants' knowledge organization. How a participant generates connections and sees similarities acts as a reflection of the participant's knowledge structure at the time of the sort, capturing the organization of their conceptual knowledge and expertise in long-term memory.

The present study builds upon previous research on the differences in expertise and categorization methods within organic chemistry (Galloway et al., 2018b, 2019). In that earlier work, we developed a categorization task to investigate how students were organizing their knowledge regarding organic chemistry reactions taught within the “Mechanistic patterns and principles” curriculum. Experts in organic chemistry validated the categorization task to ensure all reactions included were feasible and reflected the content of the curriculum. Additionally, early qualitative interviews established response process validity, ensuring the task elicited the intended cognitive processes from the target population before being implemented, as well as addressing any ambiguity, or areas of concern about the task design (Arjoon et al., 2013; Galloway et al., 2019). That study highlighted the diversity of students' responses when interpreting organic chemistry reactions and identified four levels of interpretation that participants used when sorting organic reaction cards (Figure 3) (Galloway et al., 2019). A follow-up study used the

categorization task to compare how novices in Organic Chemistry II, graduate students, and experts (professors) organized the reactions (Galloway et al., 2018b). The findings from this work illustrated the differences between novices and experts, in which novices used diverse levels of interpretation including static features (*i.e.*, *Identical structural features*, and *Similar properties of structure*) and process-oriented features (*i.e.*, *Similar reaction type* and *Similar mechanism*), while experts primarily used the *Similar mechanism* level (Galloway et al., 2018b). Between these two endpoints, the graduate students' categorizations demonstrated a progression in the levels used, where graduate students in their master's beginning to incorporate more higher-level static features (*i.e.*, *Similar properties of structure*) and doctoral students relied solely on process-oriented features, but focused on *Similar reaction type* features more than experts. These findings are consistent with the previous literature across disciplines, demonstrating that experts approached problems differently than novices, including that experts tended to attend to the conceptual or deep features within a problem (Chi et al., 1981; Hoskinson et al., 2017; Krieter et al., 2016; Randles & Overton, 2015). The present study explored how the previous findings extend to a larger context, analyzing students' categorization choices, abilities, and changes over time.

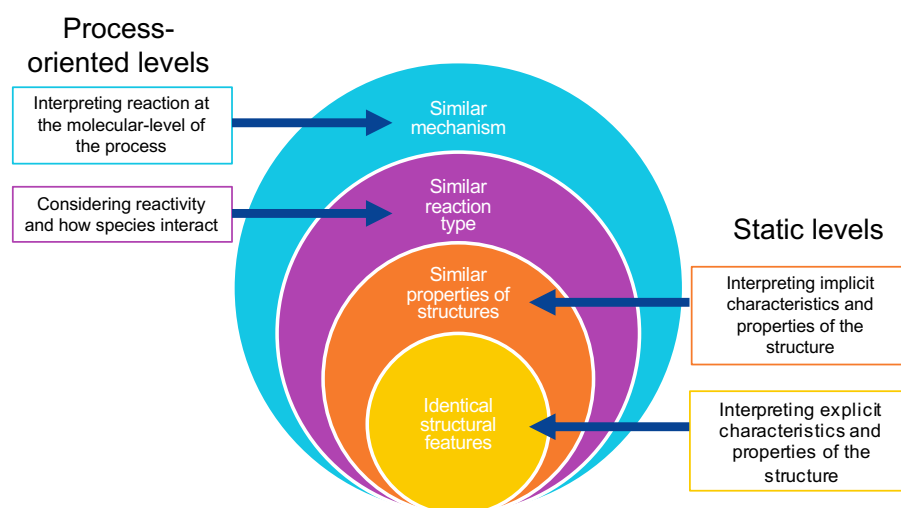


FIGURE 3. METHODS OF CATEGORIZING ORGANIC CHEMISTRY REACTIONS (GALLOWAY ET AL., 2019)

Research questions

This study exploring students' knowledge organization of organic chemistry reactions, was guided by three main research questions: 1) How do individuals' categorizations qualitatively change from an early and a late administration? 2) How do students' choices in organizing reactions change between an early and a late administration? 3) How do students' abilities to categorize reactions mechanistically change between an early and a late administration, when cued to do so?

Methods

Modifications to the categorization task and migration to an online setting

The previously-used categorization research instrument (Galloway et al., 2018b, 2019) was migrated to an online platform (Optimal Workshop, 2015) and made into two consecutive tasks: an open and a closed categorization task. During the first open categorization task, participants were prompted to place the reactions cards into categories of their choosing with no cues or restrictions. Participants were asked to name each category and explain their categorization choices. The open categorization task intended to investigate how participants *chose* to organize their knowledge around organic chemistry reactions and what similarities within reactions they naturally saw (Harper et al., 2003; Spencer, 2009). In the closed categorization task, participants were presented with an identical set of reaction cards and provided with eight categories that aligned with the organization of the curriculum (*i.e.*, the mechanistic patterns), presented in a condensed form (Figure 2). Additionally, an “Other” category option was given that participants could use if they felt that a card did not belong to any of the categories provided. While completing both categorization tasks, participants were instructed not to consult outside resources or peers. By working individually, we intended to gather responses that reflected students’ own choices and abilities. Although the online nature of the task prevented us from monitoring the extent of individual work, the relatively short median completion time of the tasks (Table 1) supports that

the participants' categorizations were a natural reflection of their choices and abilities and that they spent little time consulting outside resources.

TABLE 1. DESCRIPTIVE STATISTICS OF CATEGORIZATION TASK COMPLETION TIMES

Categorization task		<i>Mdn</i> (Min.)	1st Quart. (Min.)	2nd Quart. (Min.)
Early administration	Open categorization	13	8	19
	Closed categorization	7	5	13
Late administration	Open categorization	11	7	25
	Closed categorization	8	5	13

Mdn = Median, Quart. = Quartile, Min. = Minutes

There were a few reasons to migrate the previously used qualitative categorization instrument to an online setting. The manual categorization task was labour-intensive for both the researcher and participants and thereby restricted the sample size that could practically be accommodated (Harper et al., 2003). The online format of the categorization task allowed participants to complete the categorization task at their leisure. The online administration method also intended to alleviate some of the risks of post-administration data treatment errors that can occur with a large sample size (Harper et al., 2003). We did not change the total number of cards (25) or the reactions depicted on the cards.

Instructional context and implementation of the online categorization task

Participants of the study were students at a large Canadian university and enrolled in Organic Chemistry II structured in a flipped course format (Flynn, 2015, 2017). The organic chemistry sequence consists of four one-semester organic chemistry courses, one semester in each year. Completing the tasks were an optional part of the course, worth 1% towards the course's total participation mark (5%). Institutional research board approval was received for secondary use of data. Participants were asked to complete the open and closed categorization task hosted through the OptimalSort software (Optimal Workshop, 2015) at both an early and a late time point, one week from the start and end of the course, respectively (Figure 4). For this study, we only

included participants who completed both open and closed sorts at both time points, with the categorizations done in consecutive order (two participants completed the closed task before the open task). Of the 178 students enrolled in the course, 74 participants completed the early administration, and 44 participants completed the later administration; 24 participants met the criteria as mentioned earlier, and their data were included in the present study.

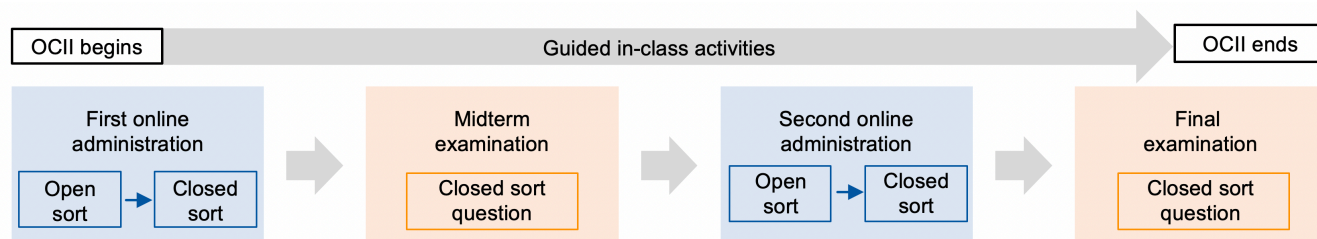


FIGURE 4. STUDY'S TIMELINE. OCII = ORGANIC CHEMISTRY II.

In addition to the online categorization tasks, the patterns of mechanisms described in the closed sort categories were incorporated explicitly throughout the course. While not part of this study, students used the patterns during guided in-class activities and encountered mechanistic pattern-based questions on the midterm and final exams. Guided in-class activities were used periodically throughout the term as well. On both a midterm and the final examination, closed categorization problems asked students to identify a reaction's mechanistic pattern. These activities and assessment items were designed to help students build connections between reactions and the underlying deep features of reactivity as they progressed through the course.

Data analysis: Visualizing the categorizations with Gephi

Gephi software was used to explore and visualize the relationships between the groups of reaction cards that participants created in both the open and closed tasks. Gephi is an open-access network visualization tool used to generate graphics that demonstrate the intricate relationships between items, often used in website design and user experience optimization (Bastian, Heymann,

& Jacomy, 2009; Neiles, 2014). We have used Gephi to explore connections between reaction cards (Galloway et al., 2018b, 2019).

In Gephi, reaction cards are represented as nodes (*i.e.*, circles). These nodes are coloured to represent the underlying mechanism that governs the reaction from Figure 2. An edge (*i.e.*, line) connecting nodes represents instances when two reaction cards were placed together within a category. The more frequently that two specific cards were placed in the same category, the thicker and darker the edge appears in the network (Gephi, 2010). The visualizations are constructed using a force atlas algorithm, which places nodes with a higher frequency of pairings toward the center of the graphic and places less commonly paired nodes apart, to the periphery (Gephi, 2010, 2011). The method has been used in other contexts to demonstrate conceptual networks (Markham & Lindgren, 2014) and in our previous work (Galloway et al., 2018b, 2019). Groupings that were classified in an *unknown* or *other* category were excluded during analysis to avoid the assumption that participants viewed connections between reactions in those groups.

Data analysis: Levels of interpretation

Data analysis began by identifying the levels of interpretation (Figure 3) that participants used to categorize reactions card in the open sort (Figure 5). The participants' card grouping names and explanations were assigned a level of interpretation by applying the previously established coding scheme to the participants' explanations. When assigning the level of interpretation present in a grouping, only the approach was considered, not the correctness of the group names and explanations. This method intended to place weight on the way students organize their knowledge around reactions and not their ability with respect to specific content knowledge. For example, a participant could incorrectly identify a reaction, but still, demonstrate a process-oriented level of interpretation. Categories were coded as "Unknown" when the participant explicitly stated that the group contained cards with unknown connections to other cards. Categories were coded as "No

data" when the participant created a sort but gave no explicit chemical rationale (*e.g.*, Group names such as Organic Chemistry One Content or "Orgo 1"). This "No data" category was not previously observed in the qualitative interviews, illustrating a potential limitation of the online tool with respect to the inability to probe participants knowledge further with follow up questions. The first author initially coded the participants' categorizations, then discussed with the principal investigator, reaching consensus on any changes to coding.

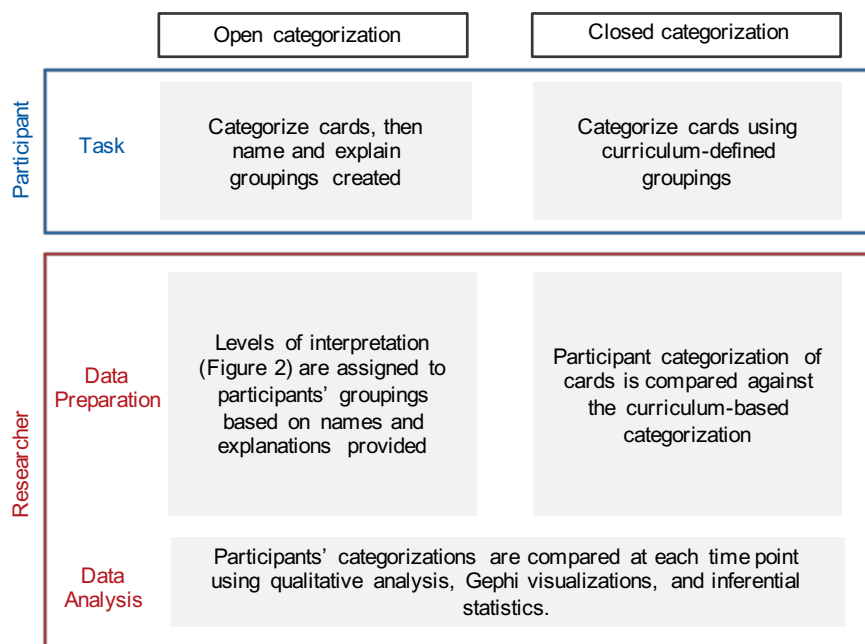


FIGURE 5. OVERVIEW OF DATA ANALYSIS.

Once coded, the frequency of each level of interpretation was compared for all participants between each time point. Changes in the frequency at each individual level between time points (*i.e.*, early and late administration) were investigated using the nonparametric Wilcoxon Signed Rank test due to the paired nature and lack of normality in the data. When quantifying the frequency, the level of interpretation for each individual card was considered instead of the number of groups generated at each level. This approach accounted for variations in participants' group sizes, as well as diversity amongst participants' propensity to use specific levels of interpretation. For example, a participant might have generated two groups, one large group with twenty cards

assigned at a lower *Identical structural features* level of interpretation and one small group at a higher *Similar reaction types* level of interpretation with five cards. For that sort, we recorded frequencies of 20 and 5 at each respective level, capturing the propensity of that participant to sort by *Identical structural features* over *Similar reaction types* level. If we had quantified frequencies by groups only (the participants' two groups of unequal size), we would have observed equal weight for levels used (*i.e.*, one lower level and one higher level group) and lost the information about the number of cards per group (*i.e.*, the greater propensity to sort by *Identical structural features*).

Data analysis: ability to identify mechanistic patterns in a reaction

To quantify categorizations and changes observed in the closed sort, we used an empirical approach based on work in physics education that gives each sort a score (Mason & Singh, 2011). We classified the reaction cards based on the most mechanistically specific pattern presented within the card, in line with the mechanistic patterns used in the curriculum (Flynn & Ogilvie, 2015). Experts in the field then validated the mechanistic categorizations. This categorization constituted the expert sort based on the curriculum's mechanistic patterns (expert sort is available in Appendix A). We assigned a score to each participant's closed sort based on the degree of match to the expert sort (%). In doing so, we could readily compare the participants' changes in ability to identify mechanistic patterns in the reactions in three ways: i) For each individual participant ii) For each mechanism type (Figure 2), and iii) For each reaction. This analysis was performed for each administration time and across both administrations, similar to a pre-test/post-test design. We then compared the participants' scores at each administration for reaction cards using a paired *t*-test. Additionally, the participants' ability to correctly categorize the cards according to the expert sort was compared at both the level of specific mechanistic patterns and the specific reaction cards using the McNemar test. The McNemar test is analogous to a repeated measure chi-square, used

to determine the changes in dichotomous variables, in this case, correct versus incorrect (Adedokun & Burgess, 2012). Effect sizes for the McNemar test were calculated using the odds ratio (OR). The odds ratio represents the odds that one outcome will be more likely with a particular exposure, that is, exposure to the “mechanistic patterns” curriculum. Rosenthal describes OR effects as small = 1.5, medium = 2.5, large = 4, very large = 10 (Rosenthal, 1996).

Interviews to investigate response process validity

To investigate response process validity, interviews ($N = 2$) were conducted following the original online administration, using a semi-structured interview format. Although the physical manipulation of the reaction cards had been validated through qualitative interviews, there were some key differences between the previous method and the online instrument that required further investigation, including the impact of delivering all 25 cards at once, how Organic Chemistry II students used the online categorization instrument, and how these students interpreted the instructional prompts and cues. The qualitative study had delivered the 25 reaction cards in a two-task design, 15 and 10 cards sequentially, as participants were initially found to be overwhelmed with all 25 at once (Galloway et al., 2019). During the response process validity interviews with the online instrument, students did not exhibit fatigue, nor expressed that they felt overwhelmed, supporting the transition to the online platform. Overall, the semi-structured think-aloud interviews demonstrated that students interpreted the goals of both the open and closed categorization task as intended using the instructional cues provided to trigger the intended cognitive processes; the interviews also illustrated potential limitations within the online design that will be discussed in depth in the limitations section. The detailed procedure used during semi-structured interviews is available in Appendix B.

Summary of evidence to support the representativeness of the population

The 24 students who participated at both time points were determined to be a robust representation of the Organic Chemistry II student population. This conclusion regarding the representativeness of the sample was made through investigating the students who participated at both time points ($N = 24$) using i) academic performance, ii) distribution in open sort, and iii) match with expert in the closed sort.

i) *Sample included demonstrated no significant difference with respect to academic performance*

In order to determine if the students who participated at both time points were representative of those who participated at one time point ($N = 60$) or did not participate at all but completed the course ($N = 90$) a repeated measure factorial ANOVA was performed to investigate the differences with respect to academic performance at both their first midterm examination and final examination. Descriptive statistics of each group with respect to mean grade, and standard deviation are summarized in Table 2. This investigation revealed no main effect with respect to the participation in a categorization [$F(2,171) = 1.54, p = .217$]. This lack of difference demonstrates that the students who participated at both time points possessed no statistical difference with respect to academic performance on major assessments compared to the other populations of students enrolled in the Organic Chemistry II course. This finding aids in establishing the participants included in the study as a robust reflection of the students enrolled in the course.

TABLE 2. DESCRIPTIVE STATISTICS RELATING TO ACADEMIC PERFORMANCE ON MAJOR ASSESSMENTS.

Group	Midterm Examination Score (%)	Final Examination Score (%)
Participated longitudinally (N=24)	$M = 65\%, SD = 22\%$	$M = 64.5\%, SD = 22\%$
Participated in one categorization task (N=60)	$M = 60\%, SD = 16\%$	$M = 64.5\%, SD = 22\%$
Did not participate (N=90)	$M = 64.5\%, SD = 22\%$	$M = 64.5\%, SD = 22\%$

ii) *Both groups demonstrated the same shift to process-oriented levels of interpretation*

The students who participated at both time points demonstrated a shift from relying on static features (*i.e.*, *structural features* and *similar properties*) when generating sorts in the early administration to relying on process-oriented (*i.e.*, *reaction type* and *mechanism*) features in the late administration (Figure 6). This trend is found to be consistent when comparing the two independent groups at early ($N = 50$) and late ($N = 20$) administration. While participants in the non-longitudinal group were observed to be more likely to rely on static features at both time points, both the students included in the study, and the independent sample of students who compared one administration demonstrated a similar increase in their propensity to use process-oriented levels of interpretation, of 24% and 28% respectively. As observed with the participants included in the study, a small number of sorts, 4% of reaction cards, were observed at the highest level of interpretation (*i.e.*, *similar mechanism*) at the late administration in the independent groups. While participants in the independent groups were more likely to use lower levels of interpretation at the later administration compared to the sample included in the study, the decrease of 15% in the "No data" sorts towards features rooted in a chemical rationale, whether static or process-oriented, acts as an additional indicator of expertise formation. By comparing the sample of participants included in the study to those who did not meet the criteria, we can emphasize the trend observed of students increasing their likelihood to rely on process-oriented features. This reliance on process-oriented features further supports the finding students progressing through the curriculum were developing a more coherent knowledge structure with respect to the domain of organic chemistry.

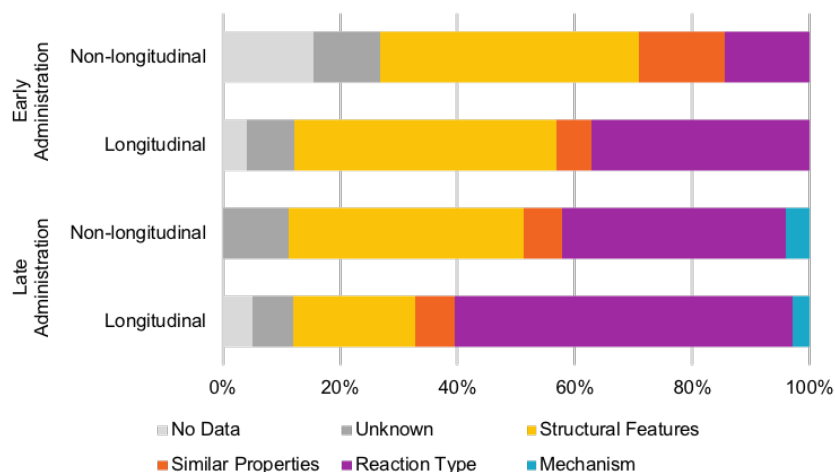


FIGURE 6. DISTRIBUTION OF INTERPRETATION LEVELS OF LONGITUDINAL AND NON-LONGITUDINAL PARTICIPANTS, SHOWN AS PERCENTAGE OF CARDS SORTED PER LEVEL OF INTERPRETATION. LONGITUDINAL PARTICIPANTS $N = 600$ CARDS, NON-LONGITUDINAL EARLY ADMINISTRATION $N = 1250$ CARD, NON-LONGITUDINAL LATE ADMINISTRATION $N = 500$

iii) No significant difference with respect to match with expert sort at either timepoint

Two independent-samples t-tests were conducted to compare the students who participated at both time points to those who only completed the closed categorization at either the early and late time point (Figure 7). At the early administration, there was no significant difference observed in the match with expert sort (%) between the those who participated in both time point ($N = 24$, $M = 25\%$, $SD = 13\%$) and a single timepoint ($N = 50$, $M = 22\%$, $SD = 10\%$); $t(72) = -1.18$, $p = .24$. Additionally, at the late administration there was no significant difference observed in the match with expert sort (%) between those who participated in both time point ($N = 24$, $M = 45\%$, $SD = 23\%$) and a single timepoint ($N = 20$, $M = 39\%$, $SD = 20\%$); $t(42) = -1.03$, $p = .31$. These results demonstrate no significant difference with respect to student's ability between those who participated at one or both of the timepoints.

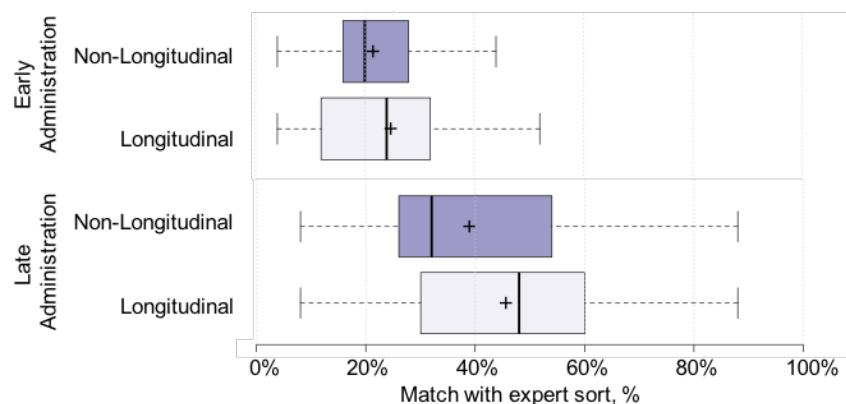


FIGURE 7. DISTRIBUTION OF PARTICIPANTS ABILITY TO CATEGORIZE REACTIONS BY MECHANISM TYPE. LONGITUDINAL PARTICIPANTS $N = 24$ PARTICIPANTS, NON-LONGITUDINAL EARLY ADMINISTRATION $N = 50$ PARTICIPANTS, NON-LONGITUDINAL LATE ADMINISTRATION $N = 20$ PARTICIPANTS.

In summary, the discussed results aid in establishing the robust nature of the sample included when compared to students who did not participate, or only participated in at a single time point. The analysis illustrates that students enrolled in the course widely possessed no statistical difference with respect to academic performance, while those who only participated both administrations of the categorization task showed strong similarities in the levels of interpretation used in the open sort, and no difference with respect to ability in the closed sort compared to those who only participated at a single time point.

Results and discussion

RQ 1: How do individuals' categorizations qualitatively compare at an early and a late administration?

Through Gephi visualizations, we visually inspected the trends in participants' groupings (Figure 8). In the early administration open and closed sorts, we observed very little clustering, indicating that we could not isolate any prominent participant-generated categorizations of reaction cards. In the open categorization task, this intricate network reflected the fact that participants used multiple levels of interpretation and multiple different types of rationales possible within a single level of interpretation. For example, students could sort by different types of surface level features (e.g., solvent or functional group), which each lead to different groupings. Similar to the open sort,

the Gephi visualization of the early administration's closed sort showed very few similarities between participants' categorization choices, indicating that participants generated diverse connections around the pre-determined mechanistic categories with little similarity to the expert sort.

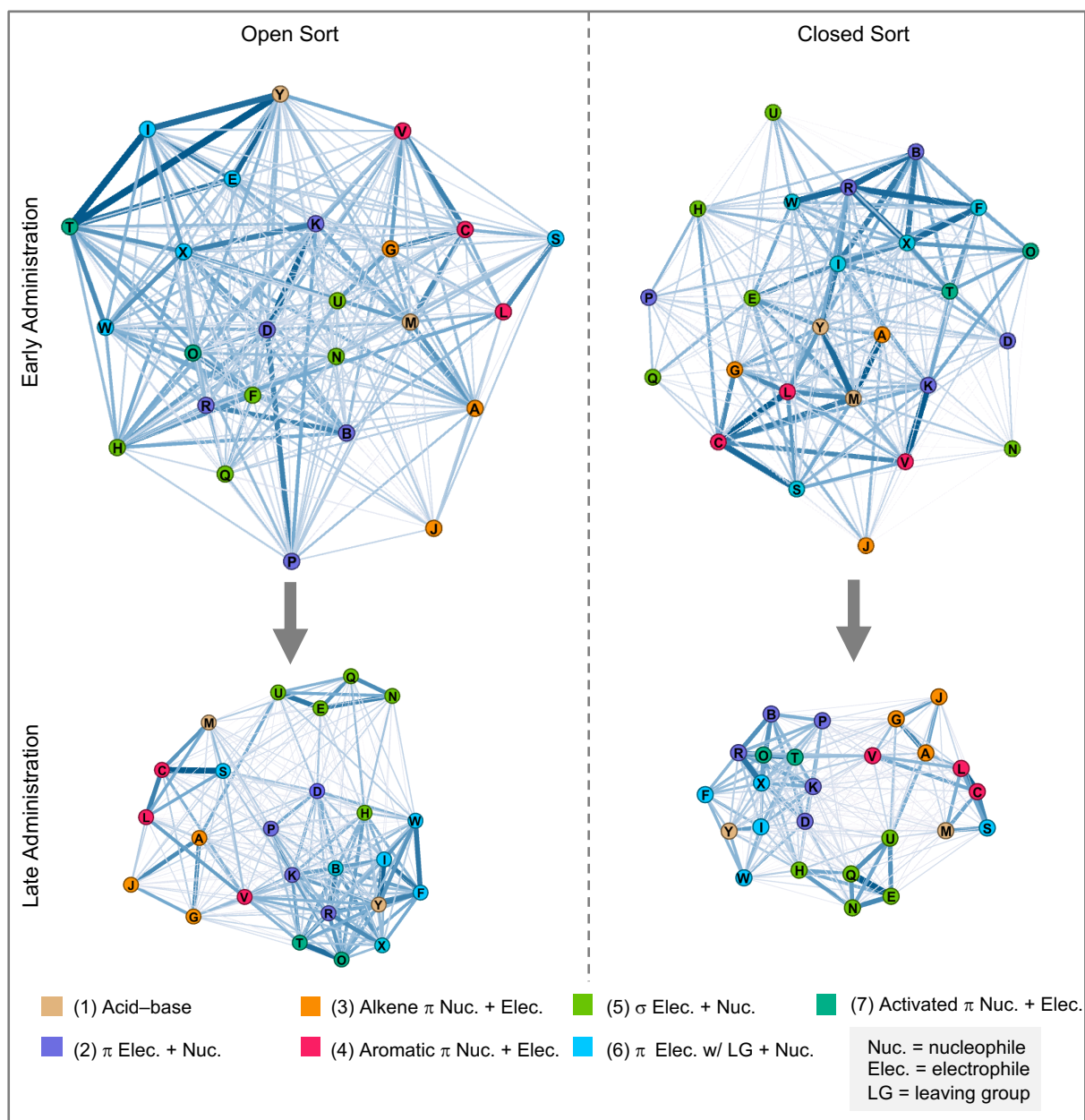


FIGURE 8. GEPHI VISUALIZATIONS OF PARTICIPANTS' REACTION GROUPINGS AT EARLY AND LATE ADMINISTRATIONS. (N = 24).

For the closed categorization task, the large variations in the sorting decisions of the early administration reflect the students' lack of ability in matching the reactions to the mechanistic patterns in the curriculum. We had originally hypothesized that the students would categorize their reactions using the more familiar patterns of reactions taught in the Organic Chemistry I course. However, the sorting of these reactions qualitatively appeared to be sorted as diversely as the unfamiliar reactions from Organic Chemistry II.

Students identified some connections between cards that were different from those predicted or identified by experts. Cards EITY formed one of the most prominent groupings in the first administration, open sort (Figure 9). The cards most commonly sorted together (in approximately 70% of participant sorts) were I with T and T with Y, as identified by the thick edges connecting the nodes representing those cards. These cards, EITY, contain four separate and distinct mechanistic patterns but were often sorted together by participants, who gave reasons of similarities in surface and static features, such as the presence of a hydroxyl (OH) group (whether in the reagent or solvent) or base, rather than using process-oriented features. In the late administration, more distinctive groupings appeared in both the open and the closed categorization tasks, meaning we were able to identify distinct clusters of nodes that are frequently sorted together. The more defined groups suggest that students were collectively starting to see similarities between reactions sharing the same deep features, although the groupings that were observed were not identical to the experts' categorizations. In both the open and the closed sorts of the late administration, the groupings AGJ and ENQU were prominent; these groupings reflect the mechanistic patterns of Category 2 π nucleophiles with electrophiles (AGJ) and Category 5 σ electrophiles with nucleophiles (ENQU), respectively (Figure 10). A large portion of the remaining cards fell into one large grouping of reaction cards containing a carbonyl group.

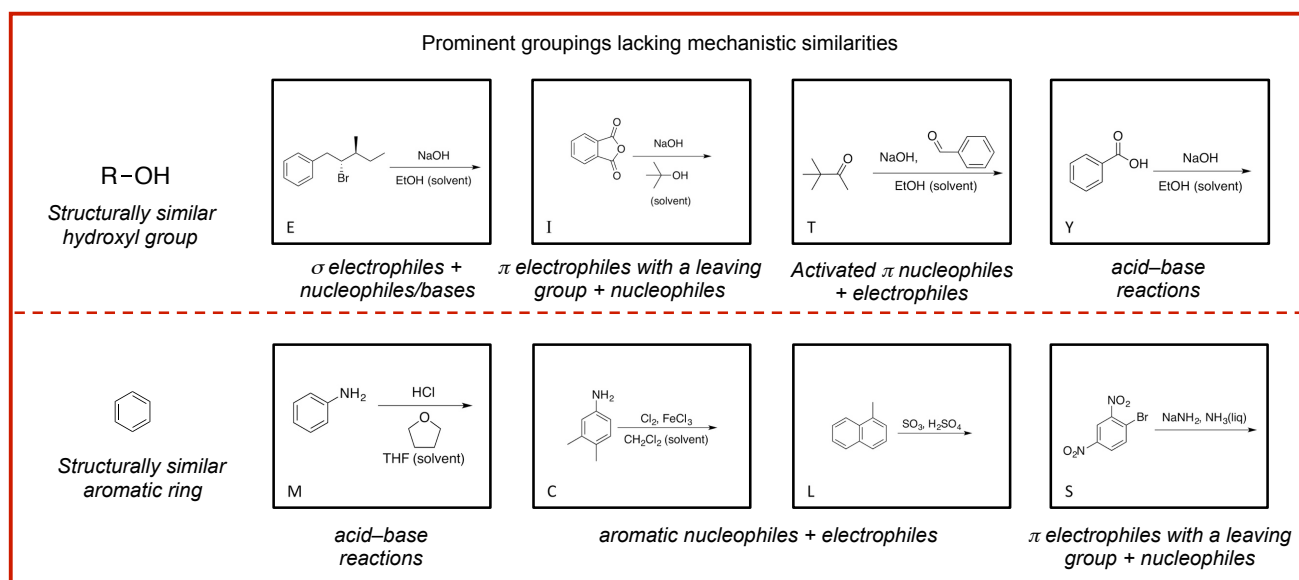


FIGURE 9. COMMONLY OBSERVED GROUPS THAT LACKED MECHANISTIC SIMILARITIES.

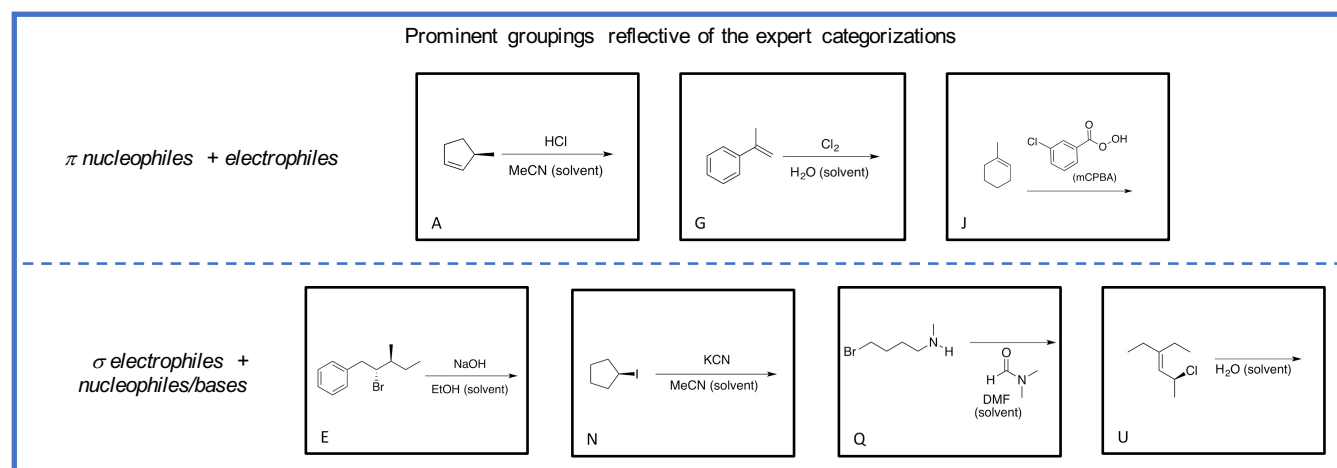


FIGURE 10. COMMONLY OBSERVED GROUPS REFLECTIVE OF THE EXPERT CATEGORIZATIONS.

While the observed groupings were more similar to those of an expert's mechanistic categorization, we also found some reaction groupings that might be of concern. For example, cards M and S were commonly sorted with reaction cards containing aromatic nucleophiles and electrophiles (C, L) (Figure 10). Although both M and S contain an aromatic ring, this structural feature plays no role in the reaction mechanism. In the late administration open sort, 70% of participants grouped C and S together, making it the most commonly occurring pairing although

the reactions proceed by different mechanisms. This trend suggests that while students may have been starting to see more expert like patterns, they may also have preferred to sort by surface features (*e.g.*, the aromatic ring).

RQ 2: How do students' choices in organizing reactions change between an early and a late administration in an Organic Chemistry II course?

The Gephi visualizations allowed us to identify trends in the participants' categorizations but cannot provide any rationales or explanations for the categorization decisions. To explore the participants' rationales, we next investigated the reasons that participants gave for their groupings and changes in those reasons over time in the open categorization task. Based on the participants' groupings and explanations, their categorizations were placed within one of the four levels of interpretation (Figure 3). As a result of a limitation of the online administration, the participants commonly provided brief and sometimes generalized explanations for their sorts, with an average explanation length of 39 words and a standard deviation of 31 words. Although these generalized explanations may not capture the full body of knowledge used by participants, they provided support to the levels of interpretation used and additional insights into why the categories were generated. For example, Sam (a pseudonym) generated categories using three levels of interpretation: structural features (*e.g.*, possessing a carbonyl), structural properties (*e.g.*, possessing a leaving group), and reactions type (*e.g.*, electrophilic aromatic substitution, epoxides/Baeyer-Villiger). Sam provided the following explanation:

"To organize the groups, I looked at the reactants used to decide which pathway they were going to take, so first I organized them by reaction but I could not figure out all the reactions, so then I started to organize them by their structure (whether they were a carbonyl or whether they had a leaving group)."

While Sam did not provide a complete explanation of their rationale for each of their sorts, we do gain a deeper understanding of how they derived each grouping and why they decided to use multiple levels of interpretation.

We compared the frequencies of each level of interpretation present for each sort using a Sankey diagram (Figure 11). The Sankey diagram shows the observed changes to students' categorizations between the early administration and the late administration but does not capture any additional changes in students' expertise that may have occurred at other time points in the semester.

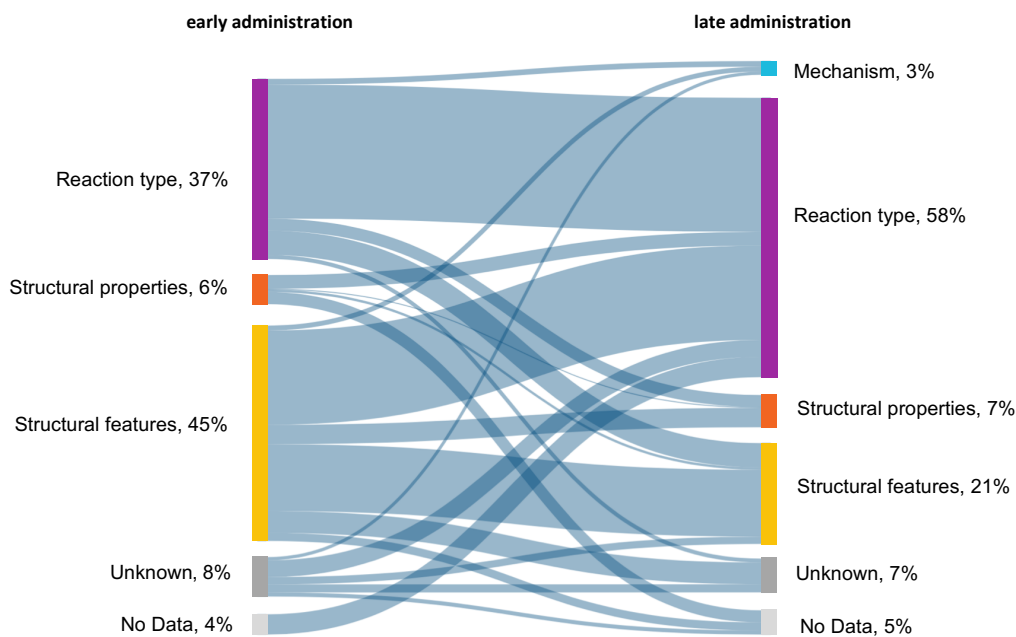


FIGURE 11. CHANGES IN INTERPRETATION LEVELS FROM EARLY TO LATE ADMINISTRATION, SHOWN AS PERCENTAGE OF CARDS SORTED PER LEVEL OF INTERPRETATION. $N = 600$ CARDS SORTED.

In the early administration, the majority of categorizations were based on the explicit static features visible on the reaction cards. Common explanations provided in these lower level categorizations included sorting by structural features, such as reagents, with 14% of all categorizations based on a single feature that was present (*i.e.*, solvent). None of the students initially described sorts at the highest level, *i.e.*, at the molecular level using language of how

electrons moved, language that is required for the sort to be coded as “*Similar Mechanism*”. The prevalence of the lower, surface-level categorizations may indicate that the curricular organization by mechanistic patterns was initially too implicit for students to notice or incorporate into their thinking. In other words, the mechanistic patterns may not have been explicit enough at that point in the students’ studies, including a lack of opportunities to practice identifying more sophisticated patterns in a variety of contexts (*e.g.*, reactions).

In the late administration, there was a statistically significant decrease in the frequency of sorting by “*Identical Structural Features*” from 45% to 21%, with the median number of reaction cards sorted at this level per participant decreasing from 11 to 5 cards out of a possible 25 ($W = 20, p = .007, r = .39$). There was a concomitant increase in cards sorted by “*Similar Reaction Type*” from 37% to 58%, with the median number of reaction cards sorted at this level per participant increasing from 9 to 14 cards ($W = 38, p = .022, r = .33$). Relating these findings to the Gephi visualizations, we confirmed that the diversity in the sorting patterns was a result of multiple different interpretations, at multiple levels and between individual levels. Moreover, at the second time point, students began sorting by increasing levels of sophistication.

While only three participants provided justifications for their categorizations based on similarities of mechanisms, participants in the late administration began to incorporate features of mechanistic thinking into their categorizations even at the reaction level. For example, one participant described reactions in which the alkene (described as C=C) acted as a nucleophile. In this example, the participant identified similar reactions based on a key mechanistic feature but made no explicit mention of the flow of electrons or molecular level required to code this explanation as being based on similarities of mechanisms. Additionally, a promising indicator for the development of more coherent knowledge organization was the fact that students were

choosing more often to attend to process-oriented features over static features. This change toward focusing on process-oriented features more often also supports the utility of the explicit instruction in the patterns of mechanisms within the curriculum, which had been done during the semester through explanations, in-class activities, and assessments.

RQ 3: How do students' abilities to categorize reactions mechanistically change between an early and a late administration, when cued to do so?

To better understand the changes in categorizations that participants made in the closed categorization tasks and consequently the students' demonstrated abilities, we compared the students' degree of match to the expert sort (%) at both the early and late administration in three ways:

i) Individual participants' categorization abilities. We observed a very large, significant improvement in individual participants' ability to correctly categorize reaction cards based on the mechanistic patterns from the early ($M = 23\%$, $SD = 13\%$) to the late administration ($M = 46\%$, $SD = 23\%$), $t(23) = -7.27$, $p < .001$, $d = 2.1$ (Figure 12). This shift provides evidence that as students were progressing through the "Mechanistic patterns and principles" curriculum, they became better equipped to identify and draw similarities between deep features in reactivity. These characteristics are likely attributable to developing a more interconnected knowledge structure like that of an expert, possessing greater conceptual expertise in organic chemistry.

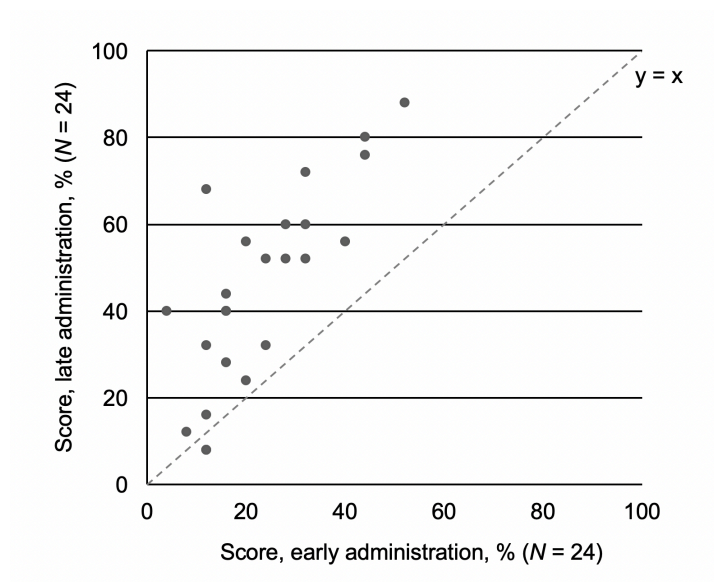


FIGURE 12. CHANGES IN PARTICIPANTS' OVERALL ABILITY TO CATEGORIZE REACTIONS BY MECHANISM TYPE. ($N = 24$.)

ii) Changes in participants' ability to identify mechanistic patterns in the reactions, analyzed for each mechanism type. We next analyzed the scores between the early and late administrations based on the mechanistic categories in the curriculum. In doing so, we identified the mechanistic patterns that participants were exceling in or struggling to identify. While the early administration contained no clear groupings in the qualitative visualization, the quantitative investigation revealed that students at the early administration were more adept at categorizing reactions from Organic Chemistry I ($M = 30\%$, $SD = 20\%$) compared to Organic Chemistry II ($M = 15\%$, $SD = 10\%$) $t(23) = 3.9$, $p < .001$. (Figure 13). For categorizing reactions of Category 4: aromatic nucleophiles with electrophiles (*i.e.*, Organic I reactions), students excelled in both the early and late administrations, demonstrating no significant change between time points, $\chi^2 (1, N = 72 \text{ cards}) = 2.89$, $p = .089$, OR [95% CI] = 2.09 [0.87, 5.02] (OR = odds ratio, CI = confidence interval). In comparison, students demonstrated a significant improvement in their ability to categorize the following mechanistic patterns:

- Organic Chemistry I, Category 2: π electrophiles + nucleophiles, $\chi^2 (1, N = 120 \text{ cards}) = 29.5$, $p < .001$, OR [95% CI] = 7.87 [3.68, 16.79]

- Organic Chemistry I, Category 3: π nucleophiles + electrophiles, $\chi^2 (1, N = 72 \text{ cards}) = 14.7$, $p < .001$, OR [95% CI] = 6.2 [2.37, 16.23]
- Organic Chemistry II, Category 5: σ electrophiles + nucleophiles, $\chi^2 (1, N = 120 \text{ cards}) = 41.40$, $p < .001$, OR [95% CI] = 13.76 [6.1, 31.04]

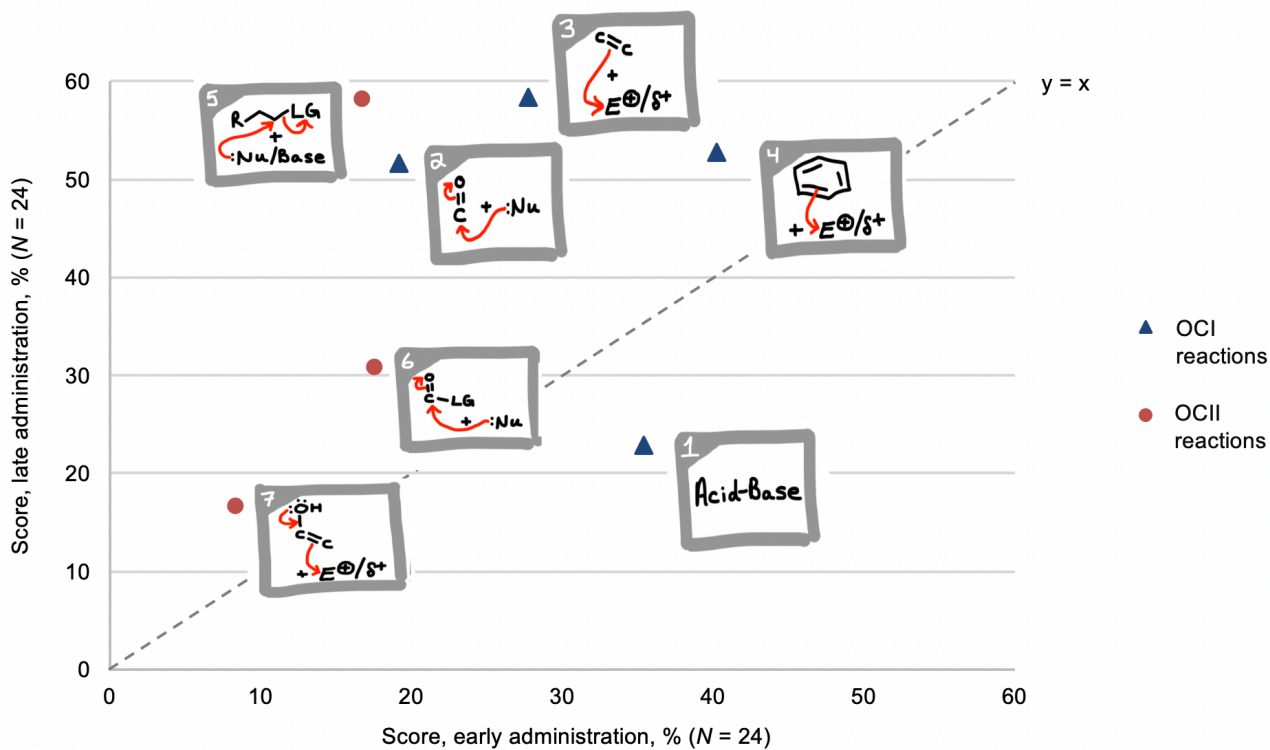


FIGURE 13. CHANGES IN PARTICIPANTS' ABILITY TO IDENTIFY MECHANISTIC PATTERNS IN THE REACTIONS, ANALYZED BY MECHANISTIC PATTERN. $N = 24$,

Contrasting this success, students had low scores and demonstrated no significant changes in their ability with respect to the following mechanistic patterns, which were taught toward the end of the Organic Chemistry II course:

- Organic Chemistry II, Category 6: π electrophiles with leaving groups + nucleophiles, $\chi^2 (1, N = 96 \text{ cards}) = 1.90$, $p = .16$, OR [95% CI] = 1.67 [0.79, 3.52]
- Organic Chemistry II, Category 7: Activated π nucleophiles with electrophiles, $\chi^2 (1, N = 48 \text{ cards}) = 0.75$, $p = .39$, OR [95% CI] = 1.57 [0.54, 4.6].

Students had a low ability to recognize the mechanistic pattern of activated π nucleophiles with electrophiles (Organic Chemistry II, Category 7, reaction cards O and T). The reaction cards O and T were sorted together by ~60% of participants in both the open and the closed categorization tasks (Figure 8) but were placed in categories other than Category 7 (activated π nucleophiles with electrophiles). This finding indicates that participants saw similarities between the cards, but these similarities could have been surface level instead of the patterns in the underlying reaction mechanism. Moreover, no changes in students' abilities were detected between the beginning and end of the semester. This lack of change throughout the semester may be related to the reactions within this category being the last mechanistic patterns taught in the curriculum. While the reactions would have been the most recent students had seen, they may not have been provided with enough opportunities within the curriculum to develop the skills required to identify the mechanistic pattern and see beyond these surface features.

Finally, there was a significant decrease in the students' ability to identify acid–base reactions, the simplest reactions they learn in the course, which were frequently placed in more complex reaction categories, χ^2 (1, N = 48 cards) = 4.08, p = .043, OR [95% CI] = 2.99 [0.98, 9.09]. These reactions were most frequently put in Category 2 (π electrophiles with nucleophiles) for reaction card Y or Category 4 (aromatic nucleophiles with electrophiles) for reaction card M. This decrease could be caused by students trying to find more advanced reactivity (*i.e.*, looking beyond what is actually present), suggesting errors in how students are identifying key features and interpreting reacting partners.

To summarize these findings, students were more capable of recognizing simpler mechanistic patterns that could be identified through surface features (*e.g.*, carbon–carbon π bond or a carbonyl functional group) at the late administration. Although students used more process-

oriented levels of explanations in the late administration when describing their categorizations, they were often unable to identify reactions that relied on interpreting the more implicit mechanistic features of the reactions (*e.g.*, identifying hidden leaving groups in reactions of acetals or recognizing the formation of activated π nucleophiles). These areas of difficulty are further elaborated in the next section. The lowest scores were attributed to reaction types that are learned latest in the course, and ergo these lower scores could be attributable to the lesser familiarity with (less time to have learned) these reactions and/or to the greater complexity of these reaction types.

iii) Participants' ability to identify mechanistic patterns in the reactions, analyzing each reaction card. In this section, we highlight three reactions to support and elaborate on students' rationale in the areas of difficult or success found in the mechanistic patterns (Figure 14). Key issues in the closed sort for the three reactions below can be summarized in the two following ways: a) The participants relied on surface features to categorize the reaction cards, and b) The participants struggled to identify more implicit aspects of the cards, such as relevant elements of reactivity.

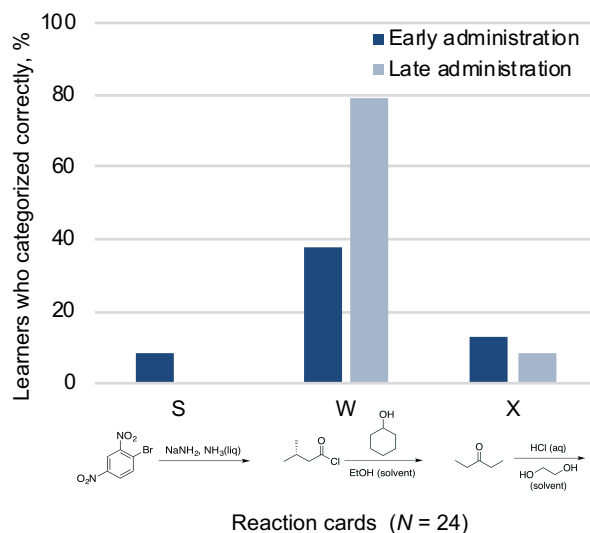
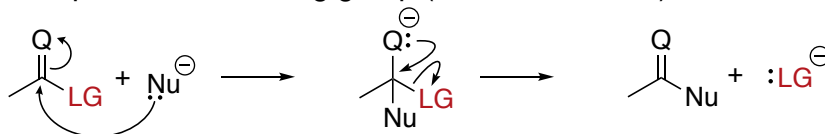


FIGURE 14. REPRESENTATIVE REACTIONS CARDS THAT HIGHLIGHT AREAS OF HIGH AND LOW SUCCESS, $N = 24$.

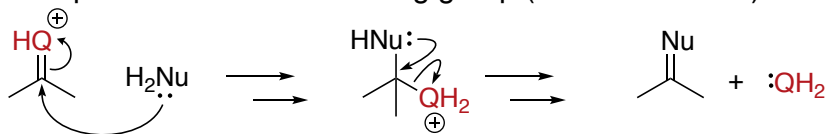
Of the reactions selected, students consistently demonstrated difficulty categorizing all but one: reaction card W. This card depicts a common reaction in Category 6 (π electrophiles with a leaving group) – this particular card has explicitly visible reactive features (*i.e.*, a reaction of a nucleophile with an electrophile that contains a familiar and readily identifiable acyl chloride group). Once students possessed the knowledge of the reaction (*i.e.*, by the late administration), there was a significant improvement from 42% to 80% of students categorizing the reaction correctly. Contrasting this improvement, reaction cards S and X were not well categorized at either time point. Although reactions S and X have the same key mechanistic pattern as W, the relevant reactive features are implicit and therefore require a deeper understanding of the reaction mechanism. Specifically, reaction card X bears a hidden leaving group present in the acetal reaction (represented as QH in Figure 15b). While, reaction S (a nucleophilic aromatic substitution) requires making a connection between two reactions that look different on the surface but proceed by the same fundamental mechanism (nucleophile reacting with a π electrophile bearing a leaving group) (Figure 15). Students were unable to draw the connections between these three reactions, with X often being sorted in Category 2 – simple π electrophile reacting with a nucleophile and S in Category 3 – aromatic nucleophile (despite having a nucleophilic reacting partner) reacting with an electrophile. These sorts emphasized a reliance on surface features when students were faced with reactions in which the key mechanistic step was located beyond the first. Although the participants were relying on surface features for these reactions in the closed sort, this did not necessarily mean that they would use static features in the open sort. Returning to the reaction card X, of the participants who categorized the reaction as π electrophiles and nucleophiles in the closed sort, six of those students also recognized this as an acetal reaction. In other words, students could be sorting by more advanced process-level methods but seeing different process-level similarities

than experts. Students may also be using various levels of mechanistic reasoning; however, we would need different or additional tasks to elicit the presence and nature of any mechanistic reasoning.

a) π electrophile with a leaving group (Reaction card W)



b) π electrophile with a hidden leaving group (Reaction card X)



c) aromatic π electrophile with a leaving group (Reaction card S)

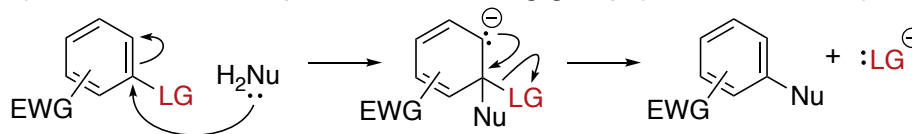


FIGURE 15. KEY MECHANISTIC STEPS OF A NUCLEOPHILE REACTING WITH A π ELECTROPHILE BEARING A LEAVING GROUP.

These findings can be related to information processing theory, to which students attend to, think about, and make connections with their long-term memory that are relevant to them (whether surface features or more process-oriented aspects). By directing attention to sorting by mechanistic (or other relevant) patterns and providing explicit opportunities to practice (*i.e.*, bringing concepts purposefully into working memory), students' skills can improve, and they can connect their knowledge in different and more advanced ways.

Limitations

The goal of this study was to understand how participants interpret reactions and mechanistic patterns but did not investigate how students can apply these mechanistic patterns and use them as a tool to solve problems in organic chemistry. These interpretations provided insights

into how students naturally draw connections but not how students attribute meaning to these different levels of interpretation; different study designs are needed for the latter purposes.

Findings from the response process validity interviews illustrated that participants might have had difficulties interpreting the symbolism used in the closed sort mechanistic categories in the online medium. For instance, participants struggled to recognize the abbreviation "LG" for leaving group, a challenge they may similarly face in classrooms in interpreting the many different representations used (Flynn & Featherstone, 2017; Johnstone, 1991). With the specific online program used, students explained their groupings after making and naming their groups (*i.e.*, on the following screen). Although the instructional prompts indicated this separation to students and asked them to record their group names before continuing, no participant followed this direction in the response process validity interviews. Students, therefore, had to recall their groupings, which increased their cognitive load and may have led to explanations being briefer and more generalized than they might otherwise have been. In the response process validity interviews, participants' explanations were reflective of the level of interpretation used in their sorts; however, this alignment may not have existed for all participants. The brief online explanations provide limited insight into the participants' reasoning and do not provide as detailed of a representation of the knowledge that students were leveraging while they were generating their categorizations when compared to qualitative interviews (Galloway et al., 2018b, 2019).

Conclusions

In this study, we used an online categorization task to investigate the changes in students' organization of knowledge around reactions in a redesigned organic chemistry curriculum. Early in the course, students completed (i) an open categorization task sorting a series of cards bearing chemical reactions to probe what similarities between the reactions they chose to see and (ii) a closed categorization task to investigate their ability to categorize reactions according to different

mechanism types. Late in the course, students repeated the categorization tasks to investigate changes in their choices and abilities over time in relation to their knowledge organization in organic chemistry. Such categorization tasks could be used online and in-class in a variety of different contexts, both in organic chemistry and other STEM fields.

Both open and closed categorization tasks were used to elicit students' knowledge in different ways. An open categorization (sorting) task allowed the participants to identify categories that may not have been anticipated by the researchers. The open task also captured the participants' choices at a given time, which can vary based on factors such as context, purpose, and ability. The closed categorization task identified the participants' abilities to categorize items in a specific way and could have given entirely different results from an open categorization task. In our work, we presented the open task to participants before the closed task to avoid issues related to cueing specific types of thinking.

Three main findings emerged from this study: (1) As the course progressed, students' categorizations became more organized and similar to those of experts, based on changes found in Gephi visualizations of both the open and closed categorization tasks (Figure 8).

(2) Based on the findings from the open categorization tasks at the end of the semester, participants chose to categorize reaction cards at more advanced levels more frequently than early in the semester, *i.e.*, by process-level reaction types and mechanisms rather than static and surface level features. The majority of the process-oriented features that were used were centred on the third level of interpretation (*Similar reaction types*), with few sorts generated at the highest level of interpretation (*Similar mechanism*). Overall, the diversity of levels used by Organic Chemistry II students aligns with previous findings from the qualitative study of Organic Chemistry II students at the end of the transformed Organic Chemistry curriculum (Galloway et al., 2018b).

These changes serve as an indicator for the formation of a more coherent and expert-like organization of knowledge around the presented reactions.

(3) Following explicit instruction of the mechanistic patterns, students demonstrated large significant increases in their abilities to identify reactions at the highest interpretation level (mechanistic patterns) in the closed categorization task (Figure 12). From the closed categorization task, specific challenges students may face were identified including: a) lowest ability connected to reactions learned latest in the course, perhaps because of having the least time to practice and learn these and/or the higher level of complexity of these reactions, b) difficulty identifying relevant elements of reactivity and c) a tendency to rely on surface features to categorize the reaction cards when the key mechanistic step is implicit (*e.g.*, occurs after the first reaction step). These challenges illustrate the need for students to have opportunities to explore possible and plausible reaction mechanisms based on fundamentals of reactivity and identify and connect knowledge using different types of features (implicit/explicit, static/process, etc.).

Implications for Teaching and Learning

Previous work has demonstrated that novices tend to rely on surface level features when faced with a problem (Chi et al., 1981; Krieter et al., 2016; Mason & Singh, 2011). This present study found predominant sorting by surface features initially but an increase in sorting by deeper, more process-oriented methods later in the curriculum, following explicit instruction and practice as part of the curriculum. We also found evidence that additional practice and development is needed with some cards still being sorted by surface features (Figure 14). Selected quotes also emphasize the impact of explicit instruction and the need for additional development, for example, in the explanation provided by Alex. In the first administration, Alex wrote: *“I looked to see if there was a solvent in the reactants or if there was not a solvent in the reactants and categorized accordingly”*. Alex demonstrated a shift in thinking in the late administration: *“I used the*

mechanistic patterns that [the professor] always uses even though I still don't really understand what they mean and wished they were more detailed." The consistent and explicit use throughout the redesigned curriculum of these mechanistic patterns helped promote an understanding of the importance of mechanistic thinking, but we can see with Alex that students may continue to struggle to recognize these patterns as a tool in explaining reactivity as an expert does. As we identify ways to improve the curriculum, students need to be able to interpret structure-property-reactivity features, including a clear understanding of the possibilities for and roles of nucleophilic and electrophilic species in reactivity. In organic chemistry and other STEM fields, categorization tasks (online and in-person) could be used in a variety of ways to help students develop connections and see patterns within and across topics.

Implications for Research

The current study captured the varying levels of expertise used to interpret reactions within a singular subset of students and demonstrated the use of an online card sorting task in tracking the formation of a well interconnected knowledge structure in a "Mechanistic patterns and principles" curriculum. To fully understand the impact of this redesigned curriculum, further work could compare the present curriculum with different curriculum designs. Before moving forward, further work is required to completely understand where, how, and why students are placing meaning to features of the reactions presented within the categorization task – how do specific distractors or positioning or reagents influence their ability to interpret the features of reactivity. Additionally, findings from this work support a need for investigations into how students understand these various methods of categorization, including patterns of reactivity.

Chapter Three: Investigating the role of multiple categorization tasks in the evaluation of a redesigned curriculum

Introduction

As continued efforts towards curricular reforms arise (Cooper & Klymkowsky, 2013; Cooper et al., 2019; Grove et al., 2008; Talanquer & Pollard, 2017), so does the increased need for the development of effective instruments and assessments for instructors to evaluate the efficacy of these changes. The University of Ottawa's recently implemented redesigned "Mechanistic patterns and principles" organic chemistry curriculum aims to emphasize the underlying mechanistic patterns that govern reaction to promote expertise development in organic chemistry (Flynn & Ogilvie, 2015). The role of pattern recognition as a fundamental skill to the nature of science has been highlighted in the Next Generation Science Standards (NGSS) framework, a curricular reform for K–12 science education in the USA (Council, 2012). This framework places the importance not merely on identifying patterns but using these patterns as a tool to seek the underlying cause of a phenomenon and formulate questions about this phenomenon. The ability to create meaningful patterns in a domain has long been attributed as a key feature of expert in a domain, whereas an expert will often notice deep features that novices will overlook (National Research Council, 2000). The previous work in Chapter two demonstrated the efficacy of the online categorization task in evaluating changes in expertise with respect to how a participant chooses to organize their knowledge and cued ability to see the underlying mechanistic patterns. Additionally, that work illustrated a potential lack of alignment between the expertise demonstrated in how a participant chooses to sort and their ability when cued, posing the question of what relationship, if any, exists between various sorting contexts. This study aims to extend on the work described in Chapter two, investigating the connections of multiple different categorization tasks including open and closed sorts, as well as varied stakes in a "Mechanistic

patterns and principles" curriculum, as well as the relationship to academic performance. In doing so, this work aims to increase the validity and reproducibility of the online categorization task as a tool for instructors to measure the formation of expertise as students' progress through the Organic Chemistry curriculum, and potential to identify traditionally "at-risk" students.

When evaluating an educational intervention, such as a curriculum, it is important to consider the *quality* of the assessment or instruments used when measuring the achievement of the learning outcomes. Frequently, low stakes assessments, such as the online categorization task, are used to capture this invaluable information for both practitioners, and the university about the academic performance of learners. Cole & Osterlind (2008) define these low stakes assessments as an assessment having no meaningful consequences to the students, whereas a high stakes assessment possesses some meaningful consequences. When using low stakes assessments, there is a concern that a lack of consequences leads to students not performing to their best effort, and may not be an accurate reflection of what a student knows (Wise & DeMars, 2005). Extending these ideas of low-stakes and high stakes assessment allow us to reflect on the online categorization task. The context that students completed the open and closed sort possessed no extrinsic motivation, and low effort could be a limiting factor of the identified demonstrated expertise. To make evidence-based decisions regarding the curriculum, it is important to ensure that the findings from the online categorization task are valid and reliable.

Assessing the validity and reliability of an instrument

Validity is analogous to the accuracy of the instrument, where various forms of validity evidence are used to demonstrate that the instrument is measuring the intended construct (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Testing, & Joint Committee on Standards for Educational and Psychological, 2014; Barbera & VandenPlas, 2011; Creswell, 2012). The validity of the

instrument refers to the validity of the data collected by the instrument within a specific population, rather than the instrument itself. Meaning that while the instrument may have proved valid in our context, the same instrument can produce different results in a new context, if implemented with a new group. In previous work, the reaction cards used within the online categorization task had been evaluated for evidence-based on test-content by a panel of experts in Organic Chemistry (Galloway et al., 2019), as well the physical and online categorization tasks were evaluated for responses process validity through qualitative interviews. Other methods to establish validity include evidence based on the internal structure and evidence-based on relations to other variables (Figure 16). When assessing validity based on relations to other scores, the intent is to relate the score of the instrument against similar (convergent validity) or dissimilar (divergent validity) tests to see if the relationship is positive or negative. Arjoon et al. (2013) note while this analysis is often performed and investigated, it is seldom described as an aspect of validity.

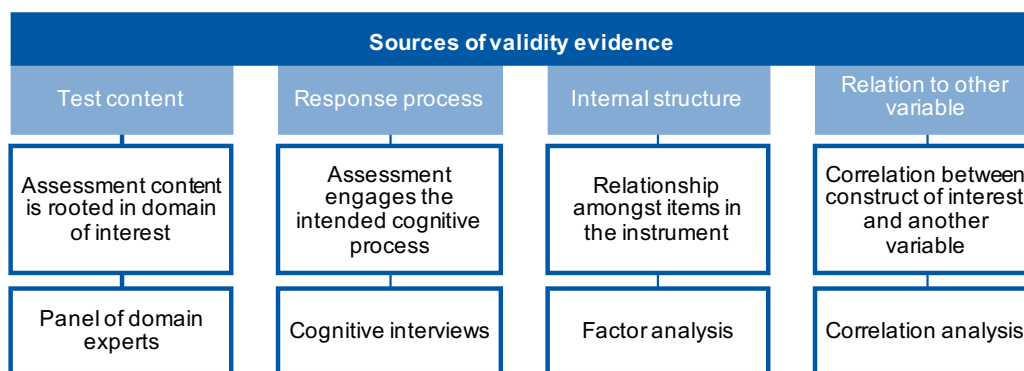


FIGURE 16. SOURCES OF VALIDITY EVIDENCE

As described in Chapter two, changes in expertise can be investigated through multiple different methods. In this work, we implemented categorization tasks to elicit students' organization of knowledge around organic chemistry reaction. To our knowledge, no work has attempted to assess the validity of the expertise demonstrated in a categorization task against other variables. The following work will look specifically at assessing the validity of expertise demonstrated in various

categorization tasks in relation to other variables intended to measure the expertise of students, specifically academic performance on examinations.

Previously work has implemented metrics from network analysis to investigate students' knowledge organization and its relation to academic performance (Acton, Johnson, & Goldsmith, 1994). While the work in chapter two implemented network analysis qualitatively (*i.e.*, Gephi visualization) to view the connections and groupings no quantitative information was extracted. Quantitative measure to accompany network analysis can include coherence, path length correlation and neighbourhood similarity. Neighbourhood similarities is a measure of the degree a concept has a similar neighbourhood between two networks (Neiles, 2014). In work by Acton et al. (1994), structural knowledge networks of individual instructors, individual non-instructor experts, average experts, and average "good" students were investigated in comparison to novice networks as predictors academic performance. This work demonstrated the utility of comparing expert networks to novices as a metric for predicting achievement on examinations. This work supports the potential relationship between the online categorization task and academic achievement.

In addition to categorization tasks, concept mapping is a commonly implemented tool for investigating and measuring a student's knowledge structure. Concept maps are graphical representations typically constructed by the participant, containing nodes and lines. Previously, in the Gephi visualizations, nodes represented specific cards in the sort and lines represented the frequency of pairings. In a concept map, these nodes represent a concept or important term and lines represent a relationship between concepts. Similar to their knowledge structure, a more expert-like concept map is thought to be more interconnected using deep features than that of a novice. Similarly to work described in Chapter two with respect to analyzing the online

categorization task, one of the key features of a concept map is the ability for it to be scored accurately and consistently (Ruiz-Primo, Shavelson, & Shultz, 1997). Concept maps can be used to monitor and track changes in knowledge structure throughout the semester as the online categorization task was (Cook, 2017), but unlike the online categorization tasks require more training to prepare students to use the task.

Work by Szu et al. (2011) investigate the relationship between how knowledge structure established through concept maps relate to academic performance determined by final grade in organic chemistry. Concept maps were generated by students in first semester organic for five of the common core areas. The links were scored ordinally, from 0, indicating incorrect or scientifically irrelevant links, to 3 indicating scientifically correct and scientifically stated links. These links were then summed to provide a score on the overall concept map. This work showed that conceptual understanding as determined by scored concept maps, specifically the two relating to 1) introduction to reactions and 2) substitution and elimination reactions were strong predictors of the final grade. While this work suggested the role of conceptual knowledge as a predictor of academic performance, the study suffered from a small sample size ($N = 20$) leaving little room for generalizability.

Reliability is analogous to the precision of the instrument and is attributed to the consistency of a measure. Reliability evidence aims to demonstrate the scores of an instrument are not due to chance and are reproducible (Arjoon et al., 2013; Barbera & VandenPlas, 2011). Reliability of an instrument is often demonstrated through evidence-based on replicate administrations, or through internal consistency. When gathering evidence-based on replicate administration, individuals from the same group will be administered the assessment twice. Following the repeated administration, the correlation coefficient between the two scores provides

evidence of test-retest reliability(American Educational Research Association et al., 2014). While the correlation coefficient helps identify if students are achieving the same scores on tests, it provides no context at the item level. Work by Brandriet and Bretz (2014) demonstrated the utility of a chi-squared test to address this limitation. This method investigated the consistency of student responses at an item level during test-retest reliability. Transitioning these concepts to our context, we aim to further improve on the validity of the online categorization task as an instrument for measuring expertise in organic chemistry while additionally investigating the reproducibility of the findings from the instrument in a near-identical context.

Research Questions:

The present study aims to investigate the relationship between various categorization tasks, probing the relationship between different categorization tasks in different contexts as well as the relationship to other indicators of expertise. Additionally, this study aims to demonstrate the stability of these findings. The following research questions guided this goal:

1. What relationship, if any, can be observed between categorization tasks in varied contexts?
2. How do measures of expertise determined through a categorization task relate to academic performance?
3. To what extent are findings from various categorization tasks reproducible?

Methods:

Participants and instructional context:

This study was performed at a large Canadian research-oriented university. All students enrolled in Organic Chemistry II were invited to participate the low stakes' online categorization tasks. Students who chose to participate in a minimum of one of the two low stakes online categorization task were incentivized with a bonus of 1% towards a participation mark accounting for a total of 5% of their total grade, through this approach a convenience sampling was used (L.

Cohen, 2010b). Students entering Organic Chemistry II had been previously taught the symbolism of organic chemistry (*i.e.*, the electron pushing formalism) and had introduced to reactions governed by the following mechanistic patterns as described in Figure 2: acid–base reaction, π -electrophiles + nucleophiles, alkene π nucleophiles + electrophiles, and aromatic nucleophiles + electrophiles. The Organic Chemistry II course was delivered to students in a flipped format (Flynn, 2015, 2017) in the falls semester of the 2017 and 2018 academic year.

Data Collection

Throughout the Organic Chemistry II course, categorization tasks were incorporated using three components: low stakes online instruments, guided in-class activities, and high stakes major assessments (Figure 17). During the course, students are first recruited to complete a low stakes online categorization task hosted through the Optimal Sort software (Optimal Workshop, 2015). The low stakes online categorization task followed an identical procedure to that described in the Chapter two method section. The two additional components, used during major assessments and guided in-class activities, asked all enrolled students to complete a closed-sort like categorization task. These tasks asked students to sort up to eight reaction cards into the pre-determined categories (Figure 2) included in the online instrument. These reactions aimed to be of similar difficulty and to reactions asked within the low stakes task. While only the closed-sort tasks provided on the final examination are included in this study, the guided in-class activities and midterm closed sort like problem were used to emphasize the importance of the underlying mechanistic patterns discussed throughout the curriculum. All closed-sort like tasks required students to be knowledgeable on mechanistic patterns previously seen in Organic Chemistry I, as well those learnt in Organic Chemistry II. For the complete list of questions implemented in the final examination, please see Appendix C. In addition to the previously described sampling criteria requiring students to complete both the open and closed administration in subsequent order, participants in this study

were required to have completed both major assessments. This additional sampling criteria lead to the exclusion of nine participants in the early administration, and one in the late. The midterm assessment and final examination were delivered to the learners approximately two weeks after completing the low stakes card sorts; these assessments contributed 10–20% and 40–60 %, respectively, towards their final course grade. All assessments included the courses were designed in collaboration with an expert in chemistry education to reflect the content of the “Mechanistic patterns and principles” curriculum, with a strong emphasis on student’s understanding and ability to reason through principles of reactivity.

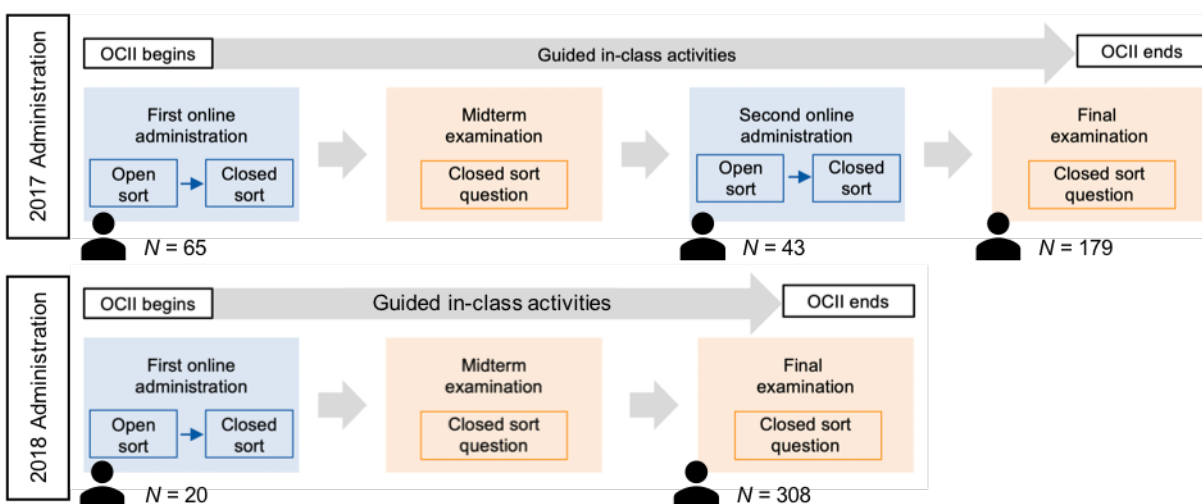


FIGURE 17. STUDY TIMELINE

Level of interpretation: Quantifying the expertise demonstrated in how a participant naturally chooses to sort

In the open sort, the group names as well as explanations of the categories were used to assign the level of interpretation, as describe previously described in Chapter two. Once all participant generated grouped were assigned a level of interpretation, the sorts were quantified for statistical analysis. To effectively capture the distribution of levels used by each participant, levels of interpretation were quantified per card. Each card was allocated a numeric value increasing from 1–4 based on the level of interpretations in the features attended to, where a value of 1 was

assigned if a card was at the most novice level of “Similar Structural Features” increasing to a value of 4 if a card was at the highest level of “Similar Mechanism”. This scheme was applied to all reaction cards, and the level of each individual card was summed to give a total possible demonstrated level of interpretation out of 100. Participant groupings that contained no chemical rational, *i.e.*, “no data” group or were assigned to an “Unknown” category were given no weight towards their overall level, *i.e.*, assigned a value of 0. For example, if a participant generated a sort with all cards placed at the “Similar Structure Features” level, all cards would receive a scoring of 1 for the level of interpretation used, leading to a summed score of 25/100. This method aimed to mimic the scoring procedures used in concept mapping, but rather than capture accuracy and correctness of the individual links made, the scores were an indication of the expertise capture through the connections and explanations observed at each level of interpretation used (Ruiz-Primo et al., 1997). Again, this work placed no weight to the correctness of the groupings provided as the interest of this work was on how they organized their knowledge around the cards.

Match with Expert: Quantifying a participant’s ability to sort by the mechanistic pattern governing a reaction.

In the online categorization tasks closed sorts, participants’ cued ability was captured through comparison to an expert sort to generate a “Match with expert” score (%) for each individual administration, as previously described in Chapter two. Additionally, a second expert sort was developed for the high stakes’ final examination categorization task. This approach for quantifying the high stakes task allowed for a similar comparison of a participant's ability to categorize reaction cards like an expert would, giving participants a “Match with Expert” Score out of 6. A “Match with Expert” score of six indicates that participants correctly categorized all reaction cards included in the problem, including the reactions in part A of the task that are assigned by participants (four reactions) and those in part B that required participants to correctly

identify the two reaction cards belonging to a category 7 – activated π nucleophiles + electrophiles. These participant-specific scores allowed for us to readily make comparisons against the level of interpretation score determined from the open sort, and other metrics such as academic performance.

Data Analysis:

Characterizing the relationships between multiple categorization tasks (RQ1)

To investigate the relationship between the expertise demonstrated in the online categorization tasks (*i.e.*, level of interpretation and match with expert), as well as demonstrated in closed-sort like task (match with expert sort) in a higher stake setting, a correlation analysis was performed using SPSS (“IBM SPSS Statistics,” 2019). Mirroring procedures used within concept mapping for dealing with summed score of qualitative characteristics, after investigating normality, Pearson R was implemented in all correlation analysis, the strength of the correlation (*r*-value) are qualitatively described as small = .1, medium = .3, and large = .5 (J. Cohen, 1988). In addition to the correlation analysis, trends in how participants sorted similar reactions cards were investigated between low stakes, online closed sort, and higher stakes major assessment closed sort-like problems. Participants included in this section of this analysis were those who had completed the late administration of the online categorization task ($N = 43$), and the final examination closed sort ($N = 179$). Each card in Question 1 of the high-stake categorization task possessed a *matched* card in the low stakes online task. Reaction cards between the two categorization tasks were matched based on similarities in the underlying mechanistic patterns. Trends were first visually inspected, followed by the quantitative investigation using a Chi-Squared test of independence to identify any differences in frequency of use between the pre-determined mechanistic categories (Franke, Ho, & Christie, 2012). By evaluating the relationship

between the low-stakes online categorization task, and high stakes major assessment tasks, we aimed to increase the validity of scores collected through the low-stakes task and support its ability to assess a participant's expertise without fear of losing efficacy due to lack of motivation as common in low stakes assessments.

Evaluating the relationship to academic performance (RQ2) and reproducibility (RQ3)

All measures of expertise were investigated in relation to the participants' academic performance on major assessments throughout the course. In the early and late administration, the relationship between participants' level of interpretation score (/100) and match with expert (%) were correlated against major assessment scores, including the first midterm and final examination score, respectively using Pearson R. This approach was used to investigate the link between the expertise observed in each categorization task with respect to the external variable of the students' overall expertise in the course in a concurrent fashion (*i.e.*, at approximately the same time) or a predictive manner (*i.e.*, at a later time). Throughout this investigation, we acknowledge that students' performance on a major assessment is not an omnibus assessment of expertise. Additionally, the relationship between the final examinations closed sort-like problem's "Match with expert" score (/6), and the relationship to the major assessment score (%) on the final examination was investigated.

To ensure the reliability and reproducibility of the previously established findings, both the early administration of online categorization task, as well as the final examination question were implemented in Organic Chemistry II of the following year. Students in this reproduction group were in a near-identical context to those of the main group *i.e.*, used the same curriculum, flipped course organization, and course instructor. Similar to the original group, students in the reproduction group were given explicit instruction of the mechanistic patterns and a final

examination that possessed a similar structure, and many identical problems. Due to the similarity of the final examination, participants in the reproduction group who completed the online categorization task ($N=20$) were subsequently matched with a participant from the main study according to standardized z-scores from the final examination, with an average difference in z-scores of 0.026 units. While this method of matching does not allow for us to capture the changes that participants might have incurred in expertise from the early to late time points in the course, investigating the similarities on the examination act at the most objective measure of expertise available. The online categorization tasks and final examination categorization task were quantified using the aforementioned procedures. Using the two population samples, we are able to determine the reproducibility of the findings with respect to the expertise demonstrated in choice and ability. Additionally, the stability of results relating to the final examination categorization task of the main group ($N=179$) and reproduction group ($N=308$) were compared using a chi-squared analysis to compare the use frequency of mechanistic categories used. Lastly, the matched sample was used to illustrate the test-retest reliability of the categorization task for probing expertise related to how participants choose to sort and their ability when cued to do so.

Results and Discussion:

RQ1. What relationship, if any, can be observed between categorization tasks in varied contexts?

The following aims to investigate the relationships between categorization tasks in varied context, including the relationship between choice and ability, as well as ability in various stakes, in two ways: i) broadly with a correlation analysis exploring the relationships and using ii) an item-based approach to inspecting trends between different categorization contexts.

1) Investigating the relationship between multiple categorization tasks

Although we hypothesized that participants who chose to sort by features requiring higher expertise would also be able to see the underlying mechanistic patterns, this was not found to be a consistent trend across the various administration. We observed a small but significant correlation ($N = 65$, $r = .28$, $p = .026$) between how learners *chose* to organize their knowledge (*i.e.*, level of interpretation, /100) and their *cued ability* (*i.e.*, match with expert, %) at the beginning of the Organic Chemistry II course. This small correlation provides weak evidence to the relationship between the two categorization contexts (*i.e.*, open and closed sort) at the early administration. Moving forward to the late administration, we observed a large and significant correlation between a participant's demonstrated choice and cued ability ($N = 43$, $r = .53$, $p = .000$) (Figure 18). Together, these findings suggest an evolving relationship between the multiple facets of a learner's expertise. While in the early stage of the course, several participants ($N = 46$) demonstrated a well-interconnected knowledge structure rooted in process-oriented features, attributed to participants who generated sorts around similar reaction types. The way they organized their knowledge in the open sort, did not immediately imply an ability to recognize the underlying mechanistic pattern of the reaction in the closed sort. While the content in Organic Chemistry I was organized by the governing mechanism, the course did not previously promote the patterns explicitly. This lack of explicit instruction may be a contributing factor to why the students were unable to identify the most mechanistically correct category when cued during the closed sort task. As students progressed through the curriculum there was a greater alignment with the two modes of investigating expertise, with students who demonstrated more expertise with respect to how they chose to sort being more likely to recognize the underlying mechanism of the reaction. To summarize, while initially, learners may possess this conceptual expertise as captured in the open

sort, we can see how the explicit instruction of the mechanistic patterns may improve the ability to retrieve information easily and apply this knowledge in novel situations such as the context of the closed sort.

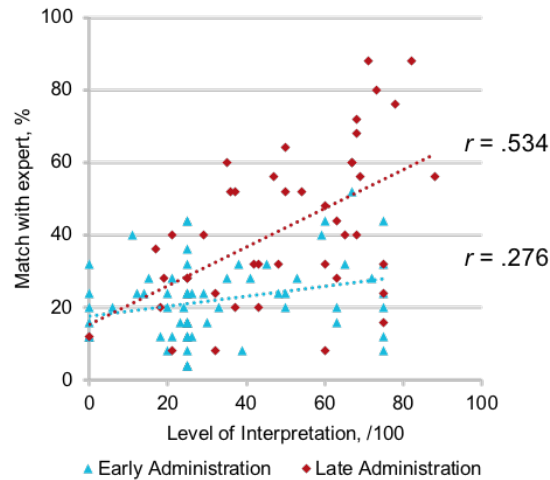


FIGURE 18. RELATIONSHIP BETWEEN HOW PARTICIPANT CHOSE TO SORT (LEVEL OF INTERPRETATION) AND THEIR ABILITY TO SORT ACCORDING TO THE DEEP FEATURES (MATCH WITH EXPERT) AT BOTH EARLY ($N = 65$) AND LATE ($N = 43$) ADMINISTRATION

The investigation into the relationship between participants' cued ability in various stakes, including the consequence-free, low stake online categorization tasks at the late administration and the high stakes major assessment categorization task at the late administration, yielded a small, non-significant correlation ($N = 43$, $r = -.29$, $p = .061$) (Figure 19). The lack of relationship between cued ability in varied stakes may be a result of changes in a participant's ability from between the two-time points. Alternatively, this can serve as an indicator that the smaller eight card instrument is not able to capture the high levels of diversity in participants ability compared to the large 25 card instrument. While card sorting tasks are unavoidably variant from administration to administration (Harper et al., 2003), this lack of continuity between the varied stakes requires further investigation to determine what factors influenced the change in participants' expertise and how can we ensure that participants who could categorize by the deep

underlying features in the low stakes setting were learning the content meaningfully. To further investigate the relationship between the varied context, an item-based approach was used to examine trends between specific reaction cards.

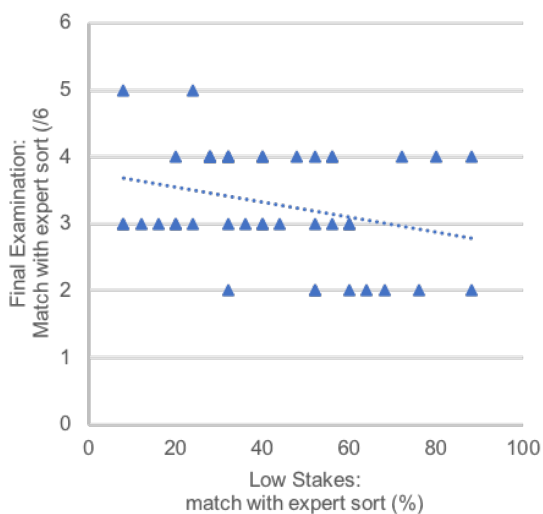


FIGURE 19. RELATIONSHIP BETWEEN CUED ABILITY IN A LOW-STAKES ONLINE CATEGORIZATION TASK AND HIGH-STAKE FINAL EXAMINATION CATEGORIZATION TASK. ($N=43$), PEARSON'S $R = -.29$, $P = .06$

ii) Item-based approach

When comparing the trends between the four reaction cards in question one of the final major assessment and their associated cards in the online categorization task, several key similarities can be concluded (Figure 20). Reaction pairs of A/F and D/X both follow the mechanistic pattern of a π electrophile with a leaving group and a nucleophile. We observed similarities in how participants were sorting the paired reaction card A/F and D/X, demonstrating no significant differences between the categories used (Figure 20). Alternatively, the reaction pairs of B/A and E/C demonstrated a significant difference in the distribution of the primary category used between the two different categorization tasks. Although, there was a significant difference in both categorization tasks of $\sim 20\%$, the shift demonstrated an increased likelihood in the high stakes to sort correct. From these two sets (B/A and E/C), we additionally observed similarities in the common errors that were made during categorization. This similarity is best illustrated in the

reaction pair of B/A, where ~20% of participants in each categorization task identified the reaction as an aromatic nucleophile reacting with an electrophile. Referring back to the reaction cards neither contains an aromatic component, yet we observed consistency the same error. In general, these similarities between the two sorts aid in demonstrating the reproducibility of the results and the generalizability of the low stakes assessment. Providing greater evidence that the trends observed may not be limited by the lack of consequences, with areas of concern such as the potential lack of clarity on the fundamentals of aromaticity in A/B, or inability to recognize a hidden leaving group as observed in pair X/D observed in both stakes.

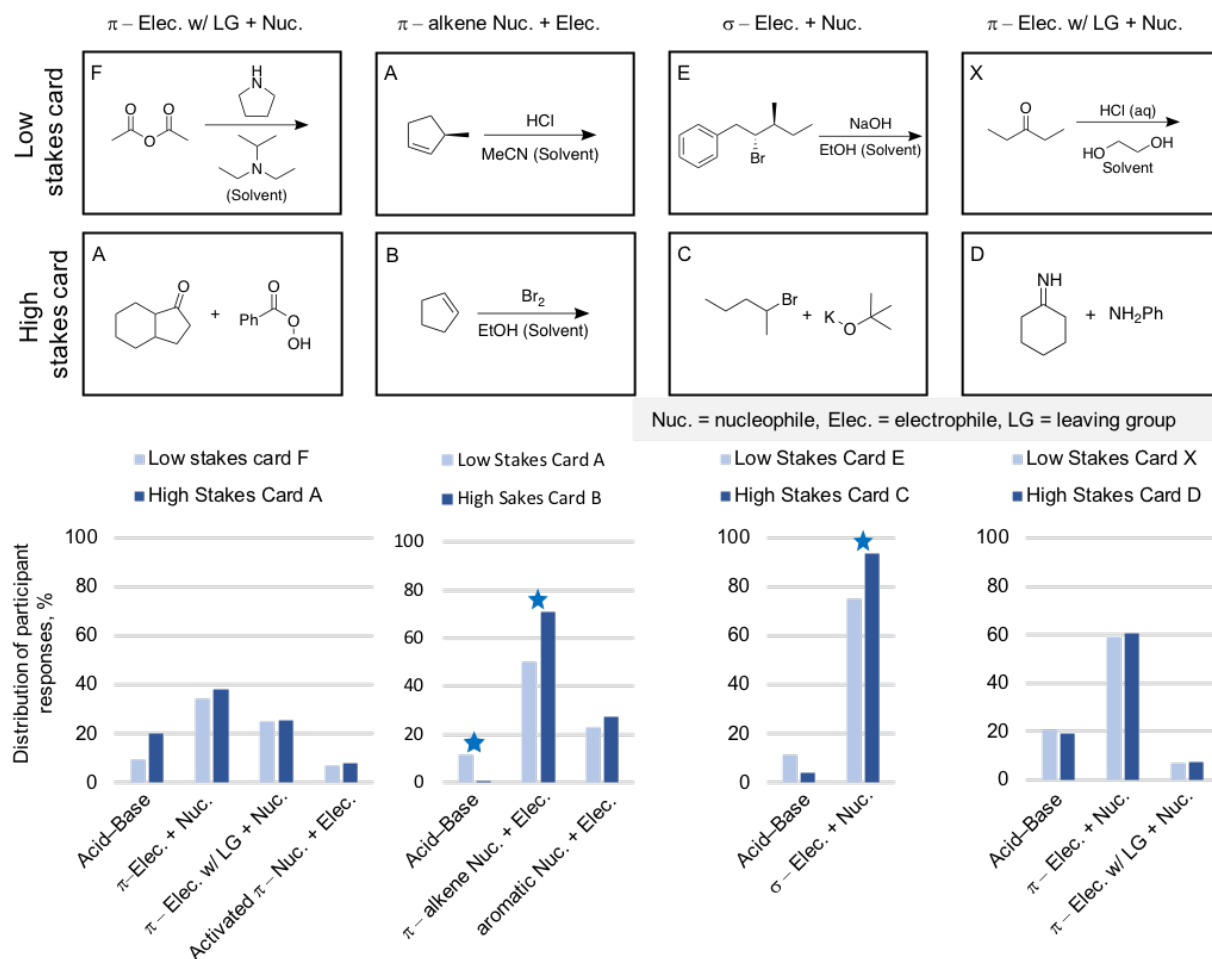


FIGURE 20. MAJOR TRENDS OBSERVED BETWEEN PARTICIPANT RESPONSES OF A HIGH AND LOW STAKES CATEGORIZATION TASK. STAR IN THE FIGURE DENOTES CATEGORIES WITH SIGNIFICANT DIFFERENCES, $P < .05$ BASED

ON CHI-SQUARED. CATEGORIES POSSESSING <5% DISTRIBUTION, AND "OTHER" CATEGORY FROM THE LOW STAKES SORT WERE OMITTED FROM THE FIGURE FOR SIMPLICITY.

RQ2. How do measures of expertise determined through a categorization task relate to academic performance?

Through the comparison of the various measures of expertise determined through the categorization task, we aimed to evaluate the criterion validity, in both a concurrent and predictive fashion, of these measures against a more tangible measure of expertise for students, major assessment grades. These measures of expertise were first compared to assessments given at approximately the same time point (*i.e.*, within two weeks). This approach aimed to capture the concurrent relationship between the measures of expertise capture on the low-stakes categorization tasks and an external but related measure. With respect to the measure of how a participant naturally chooses to sort (*i.e.*, level of interpretation), we observed a small correlation at the early administration ($N = 65, r = .28, p = .023$) and a medium correlation at the late administration ($N = 43, r = .38, p = .012$). With respect to the measure of a participant's ability, we observed a large correlation at the early administration ($N = 65, r = .53, p = .000$) and an even larger correlation at the final administration ($N = 43, r = .70, p = .000$) (Figure 21).

Understanding how a participant's choice and ability to categorize reactions are related to academic performance on the major assessment, allows us to understand which tools provide a stronger insight the expertise students possess. The open sort offers a unique insight into the connections participants naturally view as meaningfully and allow for instructors to observe any potential errors or weaknesses in the connections students are naturally making, such as students tendency to sort by solvent as seen in Chapter two. The expertise demonstrated in the open sort lacks a strong relation with academic performance, this may be due to the high variability in how a participant may choose to sort acting as a limiting factor to the utility of the online categorization task. The closed sort tasks provide a consistent and strongly correlated measure with student's academic performance and may act as a more accessible instrument for instructors to evaluate the

expertise in the course. In comparison to the open sort, the closed sort is a more accessible instrument for instructors to evaluate, as it does not require interpretation of qualitative data. While previous research had demonstrated the closed sort ability to promote both novices to sort by deep features, this finding establishes the ability to use these tools to measure academic performance in a concurrent fashion, with relevancy for instructors who wish to identify academically at-risk students ahead of the major assessment.

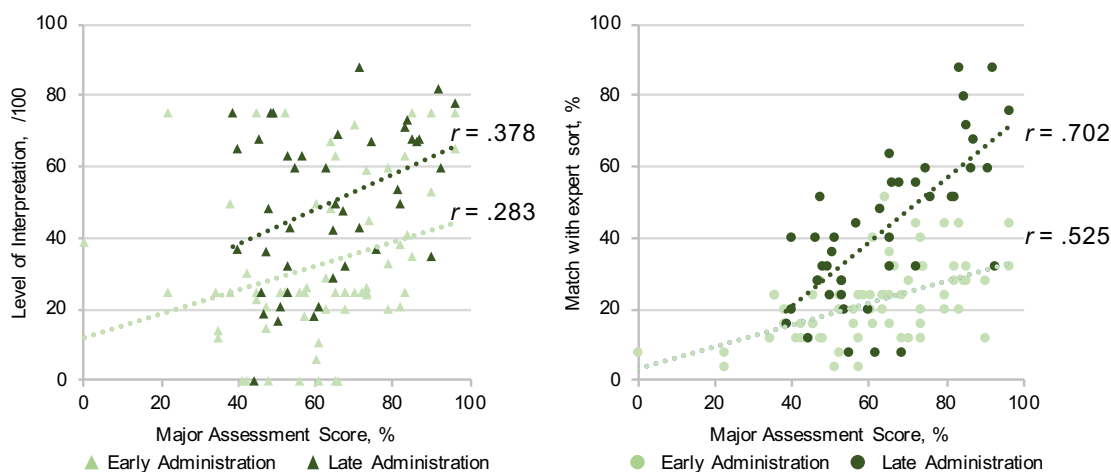


FIGURE 21. CONCURRENT RELATIONSHIP BETWEEN MAJOR ASSESSMENT SCORE AND A) LEVEL OF INTERPRETATION IN THE ONLINE OPEN SORT AND B) MATCH WITH EXPERT SORT IN THE ONLINE CLOSED SORT AT BOTH THE EARLY ADMINISTRATION ($N = 65$) AND LATE ADMINISTRATION ($N = 43$).

In addition to assessing the concurrent alignment of the open and closed online categorization task with students' academic performance, the viability of the online categorization task to predict future academic performance was investigated. Students who participated in the early administration demonstrated a moderate correlation between how they choose to sort (*i.e.*, level of interpretation) and final examination grade ($N = 65$, $r = .36$, $p = .003$) and a larger correlation between their demonstrated ability (*i.e.*, match with expert) and final examination grade ($N = 65$, $r = .48$, $p = .000$) (Figure 22). These findings demonstrate the predictive capacity of both the open and closed categorization tasks for detecting academic performance early in the course. The early administration of the open sort possessed a stronger correlation to the final major

assessment as opposed to the early major assessment. This finding may be attributed to the variable nature of the expertise an individual may choose to use in the open sort. Alternatively, this can be an indicator of the varied expertise that was required and measured by the major assessment; meaning the final examination may have had a stronger alignment with the expertise captured in the open sort's levels of interpretation than the midterm examination. Regardless, this finding aids in establishing validity evidence for the assessment for the overall expertise of an individual, and further promotes the utility of the closed sort for predicting academic performance and potentially at-risk students.

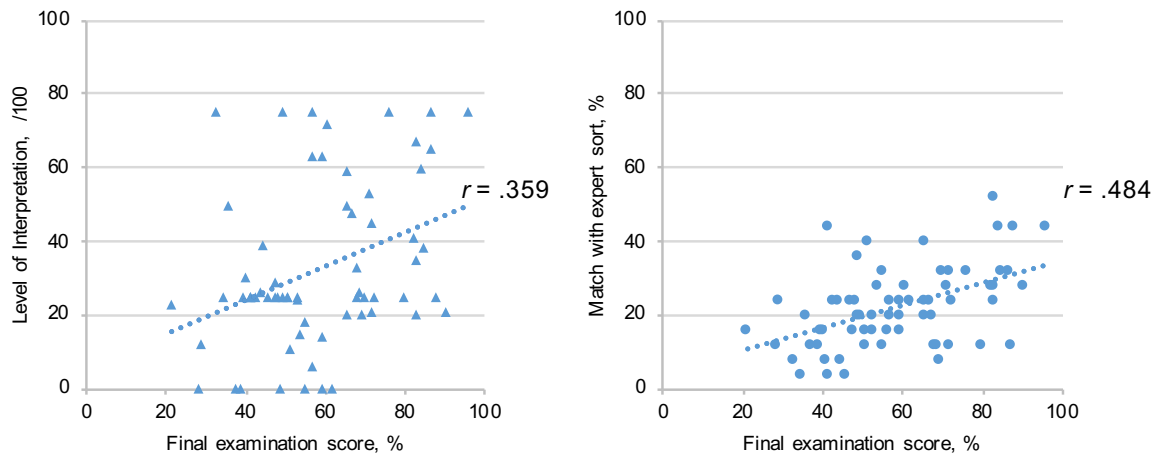


FIGURE 22. PREDICTIVE RELATIONSHIP BETWEEN MAJOR ASSESSMENT SCORE AND A) LEVEL OF INTERPRETATION IN THE ONLINE OPEN SORT AND B) MATCH WITH EXPERT SORT IN THE ONLINE CLOSED SORT AT BOTH THE EARLY ADMINISTRATION ($N = 65$)

Lastly, we observed no significant correlation between the participants' ability on the high-stakes categorization task of final examination and the overall final examination grade ($N = 179$, $r = .06$, $p = .41$). This finding can indicate multiple possibilities, such as an issue with the examination categorization task, where the problems may have been beyond the scope of knowledge expected of Organic Chemistry II students or supporting the previously discussed conclusion that the smaller eight item instrument was not able to capture the diversity in participant expertise. Alternatively, this can be an indicator that changes in expertise may have occurred

between the final examination and the online categorization task. The inconsistency of the relationship of the closed sort categorization tasks with the indicators of expertise such as academic performance is troubling for the transferability of the knowledge that we believe we are assessing with the of the online categorization instrument.

RQ 3. To what extent are findings from various categorization tasks stable and reproducible?

To ensure the reliability and reproducibility of the previous findings, the online categorization task was administered at the beginning of the subsequent year's Organic Chemistry II course. Additionally, the final examination included the identical closed sort problem. We investigated the reproducibility of the expertise demonstrated from year to year (*i.e.*, is the reliance on surface features consistent at the beginning of the course) and the stability of any trends determined within the final examination problem.

To draw comparisons between the main group and the reproduction group, we first looked broadly at the overall measures of expertise used at the beginning of the year, both the level of interpretation and match with expert (Figure 23). In the open categorization tasks, there was a slight increase in the mean level used, but this difference was not found to be significant as determined by independent t-test between students entering into Organic Chemistry II in 2017 ($N=65$, $M_{2017}= 33$ $SD_{2017}=22$) and 2018 ($M_{2018}= 39$ $SD_{2018}= 25$) ($t(83) = -1.1$, $p = .290$ 95% CI: -16.9, 13). Meaning, as students entered into OCII, they continued to rely on lower, static levels of interpretation. In the closed categorization, we again observed a no statistically significant difference from year to year between students entering into OCII in 2017 ($N=65$, $M_{2017}= 22\%$ $SD_{2017}=11\%$) and 2018 ($M_{2018}= 23\%$ $SD_{2018}= 11\%$) ($t(83) = -0.210$, $p = .834$, 95% CI: -6.1, 4.9). Through the independent t-test, we are able to determine that students entering the OCII show

consistency in the expertise demonstrated in the online categorization task. This result speaks to the generalizability of the previous findings in Chapter two.

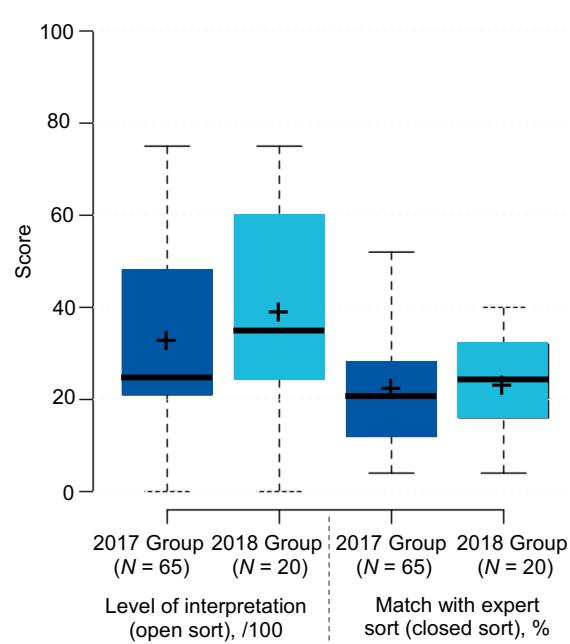


FIGURE 23. REPRODUCIBILITY OF DISTRIBUTION OF DATA FROM THE ONLINE CATEGORIZATION TASK, + INDICATING MEAN.

To further probe the stability of the categorization task, the test-retest reliability coefficient of the z-score matched participants between participants from the main group and reproduction group was investigated. With respect to the open sort and how a participant chooses to see connections, no significant relationship was observed between expertise demonstrated ($N=20$, $r = -.20$, $p = .413$) between the paired participants. (Figure 24) This lack of continuity, when paired on academic performance, serves as a support to the high level of variability found at the early administration in the term in how a participant *chooses* to sort. Contrasting this, the relationship between the ability of participants demonstrated between the two groups was highly correlated ($N=20$, $r = .67$, $p = .001$). These findings emphasize the stability of the categorization task for measuring a student's cued ability to identify underlying features of reactivity, as well as emphasis again the relationship between the student's measured ability and academic performance. This

finding suggests students who performed similarly on a final examination (based on z-scores) were likely to demonstrate similar levels of expertise in the closed sort at the beginning of the course.

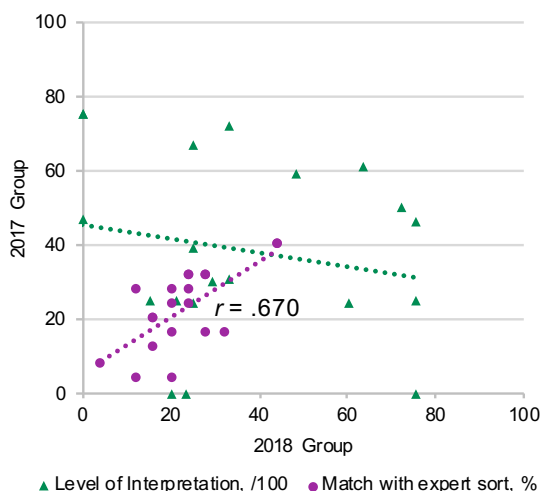


FIGURE 24. TEST-RETEST RELIABILITY OF 2017 AND 2018 GROUPS, USING PAIRS MATCHED THROUGH FINAL EXAMINATION GRADE Z-SCORES, WITH SIGNIFICANT CORRELATIONS DENOTED

Lastly, we can investigate the stability of any trends determined in the final examination, using an item-based analysis. Similar to the method previously described, but with the same item compared from year to year. As demonstrated in Figure 25, each item elicits the same trend from one administration to another in the distribution of categories used, demonstrating the reproducibility of the examination categorization task, with no significant difference ($p > .05$) observed between the main group (2017) and the reproduction group (2018). These findings also demonstrate the stability of trends previously discussed, *i.e.*, students incorrectly categorizing reaction card B as an aromatic nucleophile, students struggling to see beyond the surface features to the deep underlying mechanism as seen in reaction cards A and D.

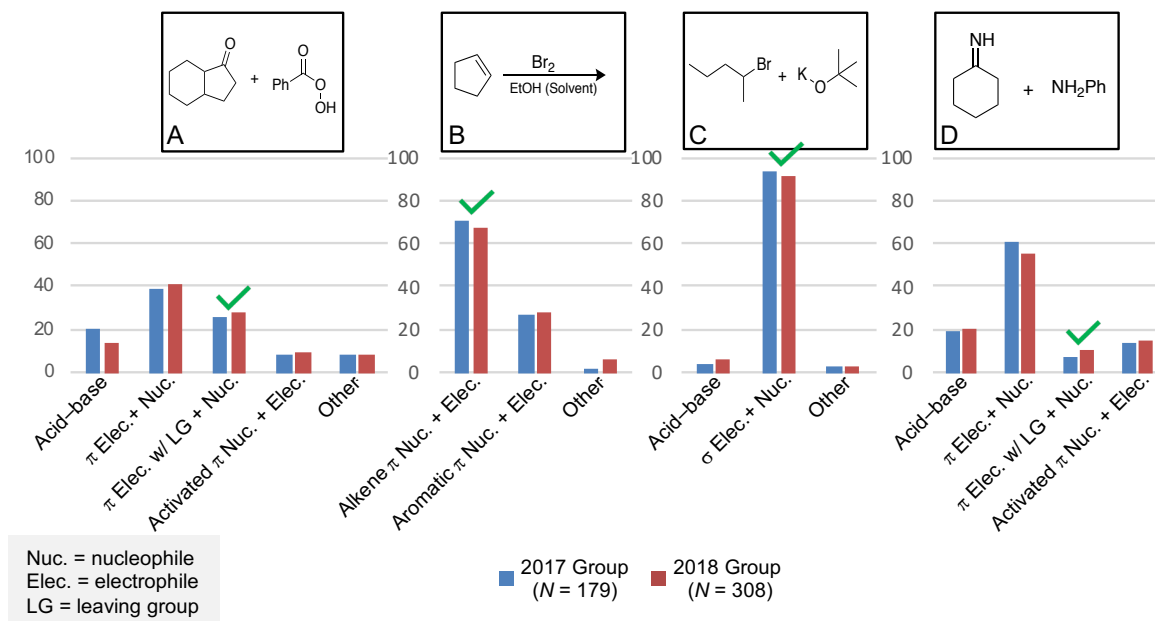


FIGURE 25. STABILITY OF TRENDS OBSERVED IN A FINAL EXAMINATION CATEGORIZATION TASK. A CHECKMARK IN THE FIGURE DENOTES CORRECT MECHANISTIC CATEGORY. NO SIGNIFICANT DIFFERENCES, $P < .05$ BASED ON CHI-SQUARED.

Limitations:

The goal of this study was to investigate the relationship between how students categorize organic chemistry reactions in a redesigned curriculum, using this relationship to establish the validity and reliability evidence of data collected with the online categorization instrument. While strong evidence of a relationship between the online closed categorization task at the end of the term and final examination performance was observed, we are unable to determine if a high achieving grade is directly related with a student understanding these underlying mechanistic patterns and principles of reactivity. In addition to the limitations previously described in Chapter two, this new quantitative approach to investigating the expertise demonstrated in the open sort with respect to scoring the levels of interpretation provides less of depth around the knowledge and categories used by participants. Through the use of a multiple administrations we aimed to enhance the reproducibility and generalizability of the findings collected using the categorization tasks, but due to the small sample size in the 2018 reproduction group we are unable to see the full

breadth of patterns which may have emerged from the new sample and can only provide a limited context of students enrolled within the "Mechanistic patterns and principles" curriculum. The limited context additionally provides limitations to the generalizability of this instrument regarding its ability to capture expertise in a traditional setting.

Conclusions:

This work extends on the prior work in the context of the curriculum evaluation discussed in Chapter two, investigating the connections of multiple different categorization tasks including open and closed sorts, as well as varied stakes (i.e., high vs. low stakes) in a "Mechanistic patterns and principles" curriculum, as well as the relationship to academic performance in organic chemistry. The findings from this work increase the validity of the conclusions drawn from the online categorization tasks, illustrating a clear relationship between the expertise demonstrated in the online categorization and academic performance in Organic Chemistry II. Through using the low stakes online categorization task in conjunctions with high stakes categorization problems, and academic performance, we aimed to capture multiple measurements of a students' expertise in organic chemistry. A correlation analysis of the relationship between the participants' level of interpretation (/100) and match with expert (%) provide a clearer understanding of the relationship between the connections students choose to see and their ability to think like an expert. These findings demonstrated a growing relationship between choice and ability as they developed their expertise in the domain ($r_{\text{early}} = .27$, $r_{\text{late}} = .53$). To predict academic performance, the closed sort provided a strongly correlated metric with upcoming major assessments ($r = .53 - .70$), while the open sort possessed small to medium correlations with academic performance ($r = .28 - .38$). Additionally, the strong correlation with the academic performance of major assessments, i.e., a metric of students' expertise in organic chemistry, not only increases the validity evidence for the findings but suggests the closed sort may be a powerful educational resource for instructors to

identify potentially at-risk students. The findings from the categorization tasks were shown to be reproducible broadly by investigating the levels of expertise used in the online categorization from year to year, showing no difference in the mean values used, and a strong correlation in the ability demonstrated by participants matched on final exam performance ($r = .67$). Analysis from an item-based perspective illustrated reproducible trends in the data with no significant difference from year to year. Collectively, these findings aid to the reliability of the low stakes online categorization task for consistently assessing how different samples of students in the redesigned “Mechanistic pattern and principles” curriculum are interpreting these reactions.

Implications for teaching and learning:

Findings from this work highlight the different roles of understanding both how a participant naturally chooses to sort and their cued ability, providing instructors in Organic Chemistry courses with a simple instrument to evaluate changes in the levels of expertise of students' as progress they throughout the curriculum. Additionally, the closed sort proves to be a practical formative, consequence-free assessment for instructors to potentially predict the academic performance of students before they enter a major assessment. The use of this instrument would allow instructors to identify high priority, at-risk students who may be academically weaker earlier, allowing for additional time to provide supports for success. While the high-stakes categorization tasks were not found to have a relationship with academic performance, the role of explicitly emphasizing these deep underlying features throughout the course remains imperative.

Implications for research:

The present work investigated the formation of a relationship between the participants' choice and ability but provides little to no insight to the role of the formative guided in-class activities and high stakes tasks included throughout the course to emphasize expert-like thinking. Further research would be required to identify how their inclusion throughout the course can be

used to better emphasize these deep underlying features and if these contribute to the strong correlation between academic performance and cued ability. Additionally, to address one of the largest limitations of this study, (*i.e.*, the limited context), future work should aim to use a multi-institution implementation of the online categorization task to increase the generalizability of these findings. While the utility of these online categorization tasks was found to be effective in the transformed organic chemistry curriculum, the current study design provides no evidence regarding how students in the traditional curriculum would perform and if these instruments remain to be an effective predictor of academic performance across different assessment contexts in organic chemistry education.

Chapter Four: Developing and validating a self-efficacy beliefs instrument for undergraduate student in organic chemistry

Introduction:

While a strong understanding of the content knowledge acts as one of many goals of a curriculum, the inclusion of non-content goals like attitude and self-beliefs (*e.g.*, self-efficacy beliefs) are pivotal features to ensuring a successful curriculum (Reed & Holme, 2014; Schmid, Youl, George, & Read, 2012). Self-efficacy beliefs are described as an individual's perceived beliefs in their capability to produce a given attainment (Bandura, 1997). Previously, establishing instruments with strong validity and reliability evidence to assess students' conceptual knowledge has played an essential role in the overall theme of this thesis with the goal of improving learning. Specifically, this thesis work demonstrated the positive influence of the redesigned curriculum on cognitive factors such as expertise formation. Prior work in the context of the curriculum evaluation has demonstrated increased abilities in the redesigned curriculum pertaining to correctly answering familiar and unfamiliar mechanism related questions, when compare to a traditional curriculum (Webber & Flynn, 2018). In the field of Chemistry Education Research, there have been several instruments that look beyond the scope of content knowledge. These instruments have been developed to measure aspects of attitude towards chemistry (Bauer, 2008; Xu & Lewis, 2011), cognitive expectations (Grove & Bretz, 2007), beliefs compared to experts (Barbera, Adams, Wieman, & Perkins, 2008), metacognitive awareness (Welsh, 2015) and self-efficacy (Dalgety et al., 2003; Uzuntiryaki & Aydin, 2009; Villafañe, Xu, & Raker, 2016; Zusho et al., 2003). In the context of the redesigned curriculum, improved self-efficacy beliefs act as one of the essential intended outcomes for learners.

While earlier findings in this thesis supported students possessing greater *abilities* as they progressed through the curriculum, the work described is unable to determine whether this

correlates to an improvement in their self-efficacy *beliefs about their abilities* in organic chemistry. Moving forward, it is important to ensure that learners are not only developing these abilities and expertise in the domain but are forming strong self-efficacy beliefs around their abilities. Within the framework of the curriculum evaluation, an instrument which is appropriate designed to investigate undergraduate students' self-efficacy beliefs in organic chemistry will allow for future researcher to evaluate 1) the alignment between a student's demonstrated abilities and their self-efficacy beliefs in their abilities, 2) changes to students' self-efficacy beliefs about Organic Chemistry due to an educational intervention such comparing the redesigned "Mechanistic patterns and principles" curriculum to a traditional curriculum, and 3) how populations' (*i.e.*, genders, race/ethnicities) self-efficacy beliefs around organic chemistry are influenced by their respective curriculum. While instruments for evaluating self-efficacy beliefs in chemistry have been constructed, they have commonly been tailored for a general chemistry audience (Dalgety et al., 2003; Uzuntiryaki & Aydin, 2009) or lack an emphasis on the key principles of reactivity in organic chemistry (Villafañe et al., 2016). Presently, this work aims to construct a valid and reliable instrument which addresses the shortcoming of prior works, to allow for future researcher to subsequently evaluate undergraduate students' self-efficacy beliefs towards organic chemistry.

Theoretical framework

Early work applied self-efficacy beliefs in clinical studies for the treatment of phobias (Bandura, 1977), researchers have since expanded that scope towards investigating the role of self-efficacy beliefs many fields such healthcare (Farrell et al., 2018; Hall, Bernhardt, & Dodd, 2015; Leary, 1985), business (Drnovšek, Wincent, & Cardon, 2010; Khedhaouria, Gurău, & Torrès, 2015; Piperopoulos & Dimov, 2015; Wilson, Kickul, & Marlino, 2007) and education. In the context of education, self-efficacy beliefs have been shown to influence factors such as learning, motivation, emotion, and achievement (Bong & Clark, 1999; Bong & Skaalvik, 2003; Schunk &

Usher, 2012; T. Williams & Williams, 2010; Zimmerman, Bandura, & Martinez-Pons, 1992). An individual's self-efficacy can influence decisions made in relation to career choices (Betz & Hackett, 1983; Hackett & Betz, 1981; Patall, 2012), effort exerted on an activity, and their resilience towards obstacles (Joët, Usher, & Bressoux, 2011; Moos & Azevedo, 2009).

Sources of self-efficacy beliefs

Self-efficacy beliefs are regarded as malleable and changing with respect to an individual's experiences (Bandura, 2015). Bandura (1997) theorized four sources contributing to an individual's self-efficacy beliefs: mastery experiences, vicarious experiences, verbal persuasion, and emotional/physiological states (Figure 26).

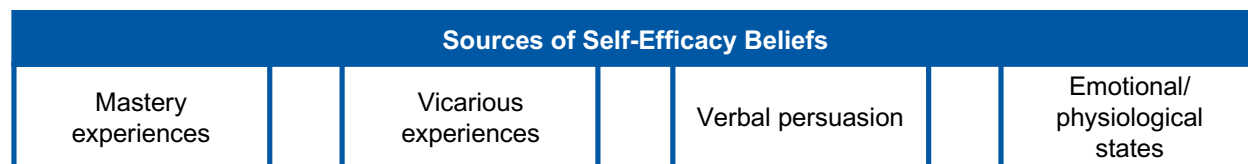


FIGURE 26. FOUR CONTRIBUTING SOURCES OF SELF-EFFICACY BELIEFS

Mastery of experience is described by an individual's experiences, where past performances act as a tangible indicator of their capabilities (Bandura, 1977). Of the four sources, mastery experiences are thought to possess the most influence on an individual's self-efficacy beliefs. For example, an individual who has viewed their experience in high school science as successful may have increased self-efficacy beliefs towards science as they transition into university science. While a person's experiences in a domain, both success and failures, may contribute towards their self-efficacy this is dependent on the individual's interpretation, meaning someone who is not easily discouraged by failure may not have their self-efficacy beliefs towards science change by a negative experience (Schunk & DiBenedetto, 2016).

The second source of self-efficacy is derived from vicarious experiences provided by social models. This source of self-efficacy is strengthened through an individual seeing people in similar contexts to themselves achieve specific goals; simply put, if they can do it, so can I (Bandura,

1997). While the influence of vicarious experiences can be essential in strengthening self-efficacy, they consequently also can produce negative impacts on self-efficacy beliefs. For example, seeing a peer whom you know is putting in a great deal of effort fail to achieve a specific task can hinder an individual's own beliefs about their personal perceived ability. Vicarious experiences can also be undermined by a negative performance experience, supporting the prominent influence mastery experiences possess (Schunk & DiBenedetto, 2016).

The remaining sources of self-efficacy belief stem from social persuasion and emotional/physiological states. Self-efficacy beliefs are developed from social persuasions are enhanced if received from a credible and/or knowledge person such as a professor in the domain of interest who can influence their beliefs that the goal is attainable, compared to if the persuasion was received from a peer (Schunk & DiBenedetto, 2016). Bandura (1997) notes that while social persuasion can have a positive influence, it much easier for social persuasion to have negative influence instilling a notion that a person lacks the capabilities. In the context of learners, this source of self-efficacy beliefs can be encouraged through positive feedback from instructors (Schunk & Usher, 2012). The final source is captured by emotional and physiological states. Emotional and physiological states such as if an individual possesses anxiety towards an academic task, this can influence an individual's perceived self-efficacy beliefs and their judgement of their capabilities. Again, the individual's interpretation of these states plays a large role in the influence on self-efficacy beliefs (Schunk & Usher, 2012).

Related constructs

Self-efficacy beliefs are conceptually similar to other constructs but possess key distinctions from these related constructs such as ability, self-confidence, outcome expectations, and self-concept. This section aims to highlight the similarities and dissimilarities between the constructs to provide a clearer distinction of self-efficacy beliefs as a construct.

As previously described, self-efficacy describes one's beliefs about their ability to produce given attainment(Bandura, 1997). Self-efficacy beliefs do not capture an individual's actual ability with respect to the given task. In the context of education, prior literature has shown there is a relationship between students with high self-efficacy beliefs in math and academic ability, but this relationship might not always be present (Collins, 1982; Pajares & Kranzler, 1995).

The term "self-confidence" is described as a general judgement of an individual's capability but commonly lacks the specificity and direction to a task that is captured within self-efficacy beliefs (National Research Council, 1994). Bandura (1997) describes self-confidence as a catchword, while self-efficacy is a theory-based construct. While a generally more confident person is often thought to be more self-efficacious, as was with ability, this assumption may not always prove valid (Schunk & DiBenedetto, 2016). Some disregard Bandura's distinction and use self-confidence in a similar manner to self-efficacy as an activity-specific construct as it is thought to be a more familiar term (National Research Council, 1994).

Bandura (2006) differentiates outcome expectations from self-efficacy beliefs, clarifying outcomes expectations are judgements about the outcome of a specific action, such as “I will receive a grade of A in this course” (Uzuntiryaki & Aydin, 2009). Rather, self-efficacy beliefs describe a perceived judgement of their ability to complete a task. Self-efficacy beliefs can contribute towards a person’s outcome expectations, but as seen with the previously constructs, might not always be an automatic relationship. Schunk & DiBenedetto (2016) describe the following example: “An academically self-efficacious student may not apply to particular universities whose entrance requirements are rigorous and whose acceptance rates are low”, as this individual may perceive the expected outcome to accompany other undesirable effects.

Though self-efficacy is related to self-concept, self-concept describes a more global construct resulting from a general perception of one's self based on experiences and their environment (Huang, 2011; Schunk & Usher, 2012). Self-concept, unlike self-efficacy, is not in relation to a specific outcome. Whereas an item within a self-efficacy scale might state "I'm confident I can understand the basic concepts taught in this course" (Pintrich, Smith, Garcia, & McKeachie, 1991), a self-concept scale is commonly thought to be broader and often involves comparisons to other's work (Schunk & DiBenedetto, 2016), for example, "Are you good at chemistry compared to others in your class?". One of the key distinctions between two constructs stems from self-efficacy being a future-oriented, malleable, goal-referenced construct, whereas self-concept relies on past experiences and is stable over time (Bong & Clark, 1999; Gibbons & Raker, 2019; Schunk & Usher, 2012). Self-concept has been heavily investigated in education research for its role in predicting academic performance and relationship to self-efficacy beliefs. Choi (2005) investigated the relationship between both constructs and academic performance in college students, comparing degrees of specificity (*i.e.*, general self-efficacy, academic self-efficacy, academic self-concept, course-specific self-efficacy and course-specific self-concept). This research demonstrated a strong relationship between the constructs of specific self-efficacy and self-concept ($r = .81, p < .01$). Additionally, this work highlighted both academic and course-specific self-concept as a predictor of academic performance, while also noting that only course-specific self-efficacy acts as a predictor of academic performance. Throughout the literature, there has been inconsistency regarding which constructs acts as the best predictor of academic performance, with some supporting self-concept (Gibbons & Raker, 2019; Skaalvik & Skaalvik, 2004) and some self-efficacy (Bong, Cho, Ahn, & Kim, 2012; Ferla, Valcke, & Cai, 2009).

However, findings from the work by Choi (2005) support the emphasis on a course specific self-efficacy rather than broader academic or general self-efficacy.

Guidelines for constructing a self-efficacy instrument

When developing an instrument to evaluate self-efficacy beliefs, the ideal instrument will possess domain specificity, capturing the multifaceted nature of the construct, and gradations of challenge. As self-efficacy beliefs are task-specific, omnibus and general instruments capturing general self-efficacy lose much of their predictive power, as observed by Choi (2005). Bandura, (2006) states, "scales of perceived self-efficacy must be tailored to the particular domain of functioning that is the object of interest.". Through the use of these omnibus instruments to assess self-efficacy beliefs, it is transformed into a personality trait rather than a task-specific judgement (Bandura, 1997; Pajares, 1996). Bandura has advised when developing instruments to assess self-efficacy to ensure they possess specificity and correspondence with criterial tasks in the domain (Bandura, 1986).

As a part of these requirements, self-efficacy beliefs scale need to be tailored to the target population of interest and avoid factors with little relation to the domain. For example, in our context, we aim to assess self-efficacy beliefs relating to organic chemistry. In order to adhere to the guidelines above, the instrument should ask students to evaluate their beliefs around their ability to perform essential tasks within the domain of organic chemistry, but tasks should not be beyond the scope of the undergraduate student population. Not surprisingly, it is ill-advised to include items unrelated to the domain, for example, self-efficacy relating to writing essays may have little conceptual relation to the desired construct of self-efficacy relating to Organic Chemistry, reducing the ability to capture the intended construct. Lastly, Bandura (2006) describes the importance of capturing the multifaceted nature of the construct, as an individual's perceived self-efficacy is better captured by the multiple related facets than just a single facet. Let consider

the example of self-efficacy to regulate study habits: self-efficacies to regulate study time, adverse mental states and understand the material can come into play to regulate the main domain and should be accounted for (Bandura, 2006, 2015).

While an instrument that possesses optimal domain specificity is invaluable, the right level of specificity required for that instrument is determined by the intended context. Identifying the optimal level of specificity has been described as the greatest challenge in designing self-efficacy instruments (Tschannen-Moran & Hoy, 2001). Items that are too specific may lose their ability to be transferred to other contexts and can present more as a problem to solve rather than a judgement of their ability to perform a task, whereas instruments that are too general may decontextualize the task-specific nature of self-efficacy beliefs (Andrew, 1998; Bandura, 2006; Uzuntiryaki & Aydin, 2009). Tschannen-Moran & Hoy (2001) provided the following example in their work investigating teacher efficacy of an item that has become too specific “I am confident I can teach simple subtraction to middle-income second graders in a rural setting who do not have specific learning disabilities, as long as my class is smaller than 22 students and good manipulatives are available”. The importance of self-efficacy beliefs rooted within the conceptual domain of interest is essential to properly assess the intended construct and improve the predictive capability of the instrument. Education research has often disregarded this warning, using general and omnibus self-efficacy instruments (Pajares, 1996). Lastly, an effective self-efficacy scale captures gradations of challenge or obstacles to successful performance (Bandura, 2006). If there are no obstacles or gradation of challenge than it may appear as though everyone possesses high or poor self-efficacy beliefs. Gradation of challenge also relates to the target population, as an instrument designed for undergraduate students in Organic Chemistry might possess different levels of challenges than one designed for high school students learning organic chemistry.

Measuring self-efficacy beliefs in STEM education

Since the early work in self-efficacy beliefs relating to clinical psychology treatments (Bandura, 1977; Condiotte & Lichtenstein, 1981), investigating the role of self-efficacy beliefs in education research is becoming more and more prominent. To better evaluate self-efficacy beliefs across different areas of STEM, instruments have been developed in multiple areas of discipline-based education research including mathematics (Lent, Lopez, & Bieschke, 1991), physics (Çalışkan, Selçuk, & Erol, 2007; Shaw, 2004), biology (Baldwin, Ebert-May, & Burns, 1999), and chemistry (Dalgety et al., 2003; Ferrell & Barbera, 2015; Uzuntiryaki & Aydin, 2009; Villafañe et al., 2016; Zusho et al., 2003).

In biology education, the College Biology Self-efficacy Instrument (CBSEI) for nonmajors was developed as a means to evaluate increases in biological literacy resulting from teaching and learning strategies (Baldwin et al., 1999). To assess the validity of the instrument, *i.e.*, the accuracy of the instrument to measure the intended construct, multiple approaches were included, such as an exploratory factor analysis with 200 students to elicit the internal structure of the instrument. Exploratory factor analysis, commonly referred as EFA, is a statistical procedure that examines the relationship between the items in the instrument, assessing the degree to which items are related to a underlying construct (or latent variable) that cannot be directly measured, in this case, biology self-efficacy. This approach aids in the identifications clusters of items that are highly correlated together, which in turn establish "factors", reducing multiple item instrument down to specific dimensions (Arjoon et al., 2013; Field, 2013; Gorsuch, 2015; Henson & Roberts, 2006). Additionally, response process validity was investigated using a focus group containing members of the target population to identifying areas of weakness within the instrument. This preliminary investigation resulted in an instrument with 23 items rated on a five-point Likert-scale (1 = strongly agree to, 5 = strongly disagree). A confirmatory factor analysis revealed three dimensions (or

factors) in the instrument relating to biological literacy: 1) methods of biology; 2) generalization to other biology/science courses and analyzing data; and 3) application of biological concepts and skills. A confirmatory factor analysis (or CFA) work in a similar fashion to an EFA, but fits the model to a pre-determined number of factors established through the EFA or underlying theoretical principles (Henson & Roberts, 2006; Osborne, Osborne, Costello, & Kellow, 2011).

Additionally, the authors investigated the relationship between students' self-efficacy in biology, as measured through their instrument, with an independent measure of biological literacy, the National Association of Biology Teachers/ National Science Teachers Association High School Biology Examination (Baldwin et al., 1999). This Examination consisted of two components, 1) content questions revolved around skills such as recalling information and transferring information to a novel situation, 2) process questions revolved around the interpreting and concluding data. This work demonstrated a weak correlation between the subscales and the independent measure, with Pearson *R*-values ranging from .18 to .27. The authors suggest this indicates that self-efficacy beliefs in biological literacy can be distinguished from the skills assessed by the examination. Although one of the earlier disciplines specific self-efficacy instruments, the use of the CBSEI has continued in education research (Ainscough et al., 2016; Gormally, Brickman, Hallar, & Armstrong, 2009) and due to the rigorous development process, it has informed the design of other instruments in anatomy and physiology self-efficacy (Witt-Rose, 2003), as well as chemistry self-efficacy (Dalgety et al., 2003).

Within the domain of chemistry, several instrument with individual strengths and weakness have been developed for evaluating self-efficacy beliefs. The remainder of this section will discussion various chemistry focused self-efficacy beliefs instruments, and their relevant validity and reliability evidence collected during their development as summarized in Table 3. Instruments

such as the Chemistry Attitudes and Experience Questions (CAEQ) (Dalgety et al., 2003) and the Chemistry-adapted version of the Motivated Strategies and Learning Questionnaire (Chem-MSLQ) (Zusho et al., 2003) act as broader instruments with only a subscale dedicated towards self-efficacy beliefs, whereas instruments such as the College Chemistry Self-Efficacy Scale (CCSS) (Uzuntiryaki & Aydin, 2009) and Organic Chemistry Self-Efficacy Scale (OCSE) (Villafañe et al., 2016) act as specific instruments tailored only to measure self-efficacy beliefs.

TABLE 3. SUMMARY OF RELEVANT SELF-EFFICACY BELIEFS INSTRUMENT, INCLUDING VALIDITY AND RELIABILITY EVIDENCE COLLECTED. MARK (✓) DENOTES EVIDENCE WAS COLLECTED.

		CBSEI	CAEQ	Chem-MLSQ	CCSS	OCSE
General details	Number of items	23	69 17 around SEB	66 7 around SEB	21	3–12*
	Number of factors	3	1	1	3	1
	Target population	Undergraduate non-major biology students	Undergraduate students in chemistry (major and nonmajor)	Introductory Chemistry students	Undergraduate students in chemistry	Organic Chemistry Students
	SEB specific instrument	✓			✓	✓
	Face	✓	✓	✓	✓	✓
Validity evidence collected	Test content		✓		✓	
	Response Process	✓	✓			
	Internal Structure	✓	✓		✓	✓
	Relation to other variables	✓	✓	✓	✓	
Reliability evidence	Internal consistency	✓	✓		✓	✓
	Test re-test					

*Number of items included varied based on content in curriculum covered.

SEB = Self-efficacy beliefs

CBSEI = College Biology Self-Efficacy Instrument, CAEQ = Chemistry Attitude and Experience Questionnaire, Chem-MSLQ = Chemistry adapted Motivated Learning Strategies Questionnaire, CCSS = College Chemistry Self-efficacy Scale, OCSE = Organic Chemistry Self-Efficacy.

The CAEQ (Dalgety et al., 2003) instrument captured multiple facets of motivation containing three subscales made up of 69 items. The CAEQ was developed with the intent of

investigating factors influencing student enrollment choices in chemistry at the tertiary level. In addition to the 17-item subscale dedicated to self-efficacy beliefs, this instrument possesses subscales relating to attitude toward chemistry and chemistry learning experiences. Similarly to the College Biology Self-efficacy Instrument (Baldwin et al., 1999), the self-efficacy beliefs portion of the CAEQ was intended to contain multiple dimensions, with four individual factors including self-efficacy including learning chemistry theory self-efficacy, applying chemistry theory self-efficacy, learning chemistry skills self-efficacy, and applying science skills self-efficacy. This four-factor structure was confirmed in a pilot study, but further investigation into the internal structure validity evidence of the construct collapsed this structure into a single factor representing chemistry self-efficacy, with no meaningful substructures. Items within the self-efficacy beliefs portion of this instrument have been critiqued for being too specific, for example, "Achieving a passing grade in a chemical hazards course", making it an instrument that may not be completely generalizable (Uzuntiryaki & Aydin, 2009). With this instrument developed, the authors evaluated the differences in self-efficacy beliefs between those planning on enrolling in second-year chemistry and those who were not. That work demonstrated that students with the intent to enroll in the additional chemistry course had higher chemistry self-efficacy beliefs than those who did not.

Similarly to the CAEQ, Zusho et al., (2003) put forward a chemistry specific instrument to measure multiple facets of students motivations processes, including self-efficacy, value beliefs, goal orientation, and affect in term of interest and anxiety. Additionally, their work aimed to create an instrument that can predict students' performance in chemistry using motivation processes, as well as including cognitive processes (*i.e.*, cognitive strategies and self-regulation strategies). The self-efficacy component was constructed by adapting the items from the Motivated Strategies for

Learning Questionnaire (MLSQ) instruments (Pintrich et al., 1991). The MLSQ is a commonly utilized instrument intended to evaluate college students' motivational orientations and cognitive use of learning strategies but acts as a general measure of self-efficacy for learning and performance rather than a discipline or course-specific measure. For example, an item within the self-efficacy subscale states, "I expect to do very well in this class.". The self-efficacy subscale of the MSLQ has previously been demonstrated promise for predicting academic performance (Gibbons & Raker, 2019; Pintrich et al., 1991). While the MLSQ has been highly cited and validated with various samples, the work by Zusho et al., (2003) neglects to provide any additional validity evidence for their adapted version.

More recent work by Uzuntiryaki & Aydin (2009) developed an instrument tailored solely for investigating self-efficacy beliefs in chemistry. The College Chemistry Self-Efficacy Scale (CCSS) instrument containing 21 items aims to address some of the shortcomings within the CAEQ while still catering to a general undergraduate chemistry population. In addition to critiques of specificity, the authors note the lack of items in the CAEQ relating to laboratory skills; neglecting the role laboratory activities on meaningful learning as supported by previous literature (Hofstein & Lunetta, 2004). In the preliminary stages of the instrument construction, the authors had intended for four factors to be investigated within their instrument including: self-efficacy for knowledge/comprehension-level skills, self-efficacy for higher-order skills, self-efficacy for psychomotor skills, and self-efficacy for everyday applications. Through the factor analysis, the first two dimensions were collapsed into one factor "cognitive skills". Items within this dimension related to "Students' beliefs in their ability to deal with intellectual operations in chemistry" (Uzuntiryaki & Aydin, 2009). The items relating to psychomotor skills consisted of students' beliefs to deal with physical skills for example, "How well can you construct laboratory

apparatus?”. The last dimension relating to everyday applications will be further explained in a later section.

As a means to further validate the CCSS, Uzuntiryaki & Aydin (2009) investigated how the three unique dimensions of their self-efficacy instrument can be used to differentiate between major (N=151) and non-major (N=198) students in a chemistry domain. While they observed a main effect in their multivariate analysis demonstrating students in a chemistry-based major possess greater self-efficacy than non-majors, follow-up univariate analysis demonstrated only a significant difference was only observed in relation to the self-efficacy for everyday applications dimension. Additionally, Uzuntiryaki & Aydin (2009) investigated the relationship between their dimensions and academic performance as defined by the end of semester course grades. This correlation analysis revealed significant positive correlations for all dimensions, with cognitive skills demonstrating a moderate correlation ($r = .34, p < .05$) and the remaining dimensions possessing small but still significant correlation as defined by J. Cohen (1988). The observed relationship between the self-efficacy capture by the CCSS and academic performance has been further explored and confirmed in a refined version of the CCSS instrument (Ferrell, Phillips, & Barbera, 2016).

Villafañe et al. (2016) developed a self-efficacy instrument specific for organic chemistry to evaluate a reciprocal causation model of self-efficacy beliefs and academic performance (Figure 27). In the hypothesized reciprocal causation model, these two measures are interrelated where "prior self-efficacy influences future self-efficacy, prior performance influences future performance, self-efficacy influences performance, and performance influences self-efficacy" (Villafañe et al., 2016). The organic chemistry self-efficacy (OCSE) instrument containing 12 items total and was administered to students five times throughout their first semester Organic

Chemistry course (Table 4). Academic performance measures were recorded five times as well using instructor prepared exams as well as a final examination from the Examinations Institute of the American Chemical Society Division of Chemical Education. At each administration, items were added to the instrument as the content was covered in the course. Items developed in the instrument do not appear to be validated by cognitive interviews or an expert panel review, but the internal factor structure was investigated using confirmatory factor analysis at each administration. Although the number of items at each administration varied, the instrument demonstrated a consistent one-factor fit of the given questionnaire using confirmatory factor analysis. While not as exhaustive of an instrument as those previously described, this work demonstrated the use of a domain-specific instrument for evaluating the relationship between self-efficacy in organic chemistry and academic performance, where self-efficacy increases with the influence of positive experiences, while additionally demonstrating the need for multiple administrations to better capture the overall picture of this relationship.

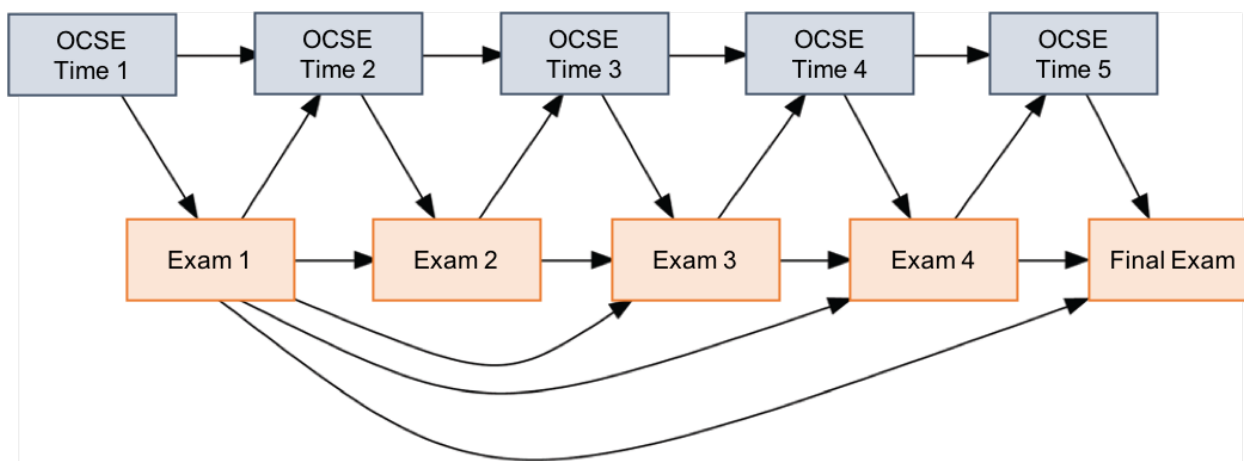


FIGURE 27. RECIPROCAL CAUSATION MODEL WITH SNOWBALL EFFECT.

Figure republished with permission of Royal Society of Chemistry – Chemistry Education Research and Practice, from *Self-efficacy and academic performance in first semester organic chemistry: testing a model of reciprocal causation*, Villafañe, S. M., Xu, X., & Raker, J. R., 17(4), 2019; permission conveyed through Copyright Clearance Center, Inc.

TABLE 4. ORGANIC CHEMISTRY SELF-EFFICACY (OCSE) ITEMS, ADAPTED FROM VILLAFANE ET AL. (2016)

Item – Prompt: How well can you...
1. Draw resonance structures for a given molecule or ion?

-
2. Determine the relative boiling points for a series of compounds?
 3. Identify functional groups present from an Infrared (IR) spectrum?
 4. Determine the most stable chair conformation of a substituted cyclohexane?
 5. Convert between a bond line structure and a Newman projection?
 6. Construct an energy diagram for the sigma bond rotation of a molecule?
 7. Convert between a drawn structure and structure built from a model kit?
 8. Assign (R) or (S) configuration to a chiral center?
 9. Convert between a bond line structure and a Fischer projection?
 10. Predict the product of a reaction from starting materials and reagents?
 11. Determine the relative stability of a series of ions?
 12. Determine the relative stability of a series of radicals?
-

Exploring self-efficacy beliefs in STEM

Self-efficacy beliefs in STEM have been an area of particular interest due to their predictive utility for academic performance, with a large body of literature exploring this relationship (Dalgety et al., 2003; Gibbons & Raker, 2019; Lynch & Trujillo, 2011; Vogel & Human-Vogel, 2016; T. Williams & Williams, 2010; Witt-Rose, 2003; Zusho et al., 2003), but also play an important role in investigating change due to an intervention such as changes to instructional context (Rao Vishnumolakala et al., 2018; Sawtelle, Brewe, Goertzen, & Kramer, 2012; Vishnumolakala, Southam, Treagust, Mocerino, & Qureshi, 2017; Winkelmann et al., 2015). Schmid, Youl, George, & Read (2012) suggested that a well-designed preparatory course plays an important role in improving self-efficacy beliefs, but the self-efficacy beliefs students hold are not consistent across all populations of students (Villafañe, Garcia, & Lewis, 2014; Zeldin, Britner, & Pajares, 2008). Previous work has demonstrated that throughout diverse areas of science education, self-efficacy beliefs have been shown to play an integral role in a person's selection of careers (Zeldin et al., 2008). That work used interviews to propose that men and women possess different sources for their self-efficacy beliefs, where men rely on mastery of experience and women rely on social persuasions and vicarious experiences. These gender-based differences indicate that it is

important to consider all of the four sources of self-efficacy when developed educational interventions.

Dalgety & Coll (2006) explored the changes in students' self-efficacy as they progressed throughout their first-year chemistry course using the CAEQ. These changes were investigated using a mixed method approach, blending quantitative use of the CAEQ triangulating self-efficacy with qualitative data from interviews with 19 students. This research revealed that chemistry students had high self-efficacy beliefs at the end of the year compared to the two early administrations at the beginning and middle of the year, but the authors attributed this finding partially to the potential for the drop-out of students with low self-efficacy beliefs. This work also demonstrated differences in self-efficacy based on gender for certain items within the instrument, in which males were more confident than females on items such as "Choosing an appropriate formula to solve a chemistry problem" and "Knowing how to convert the data obtained in a chemistry experiment to get a result" throughout the year. The authors summarize these differences stating males are more confident in advanced level skills and become more confident in tasks that involve both theoretical and practical knowledge, additionally noting, the gap between the self-efficacy based on gender was smaller towards the end of the year compared to the beginning of the year.

Across STEM, gender-based differences in self-efficacy beliefs have been heavily investigated (Lent et al., 1991; Lindstrøm & Sharma, 2011; Sawtelle, 2011; Shaw, 2004; Villafañe et al., 2014; Wachsmuth, Runyon, Drake, & Dolan, 2017; Witt-Rose, 2003). Building on the work of Dalgety & Coll (2006), Villafañe, Garcia, & Lewis (2014) investigated how self-efficacy beliefs change throughout an introductory chemistry course, specifically exploring differences based on sex, and race/ethnicity. This work used a modified version of the self-efficacy sub-scale of the CAEQ,

reducing the 17 items down to 5 items relating to chemistry knowledge, excluded items such as those relating laboratory skills, and additional modified the instrument to a 5-point scale. This instrument was delivered to students five times throughout the semester, with the first administration occurring on the first day of class and the subsequent administrations occurring in the class immediately before an exam. This approach was used to avoid having students base their self-efficacy beliefs on the exam results. This study demonstrated an overall positive shift in self-efficacy beliefs, but across different populations (*i.e.*, based on sex or race/ethnicity) this trend was not always true. A negative trend in self-efficacy was observed for Black and Hispanic males, meaning at the beginning of the course, they possessed higher self-efficacy. The authors suggest they were perhaps initially overconfident, but as they continued towards the end of the semester, they became more realistic. Additionally, this work supported the gap between gender observed by Dalgety & Coll (2006), again observing males possessing higher self-efficacy but having differences based in sex becoming less noticeable by the end of the semester. Both the previously described works act as an important call to action to understand how different populations in a classroom are influenced in order to create an optimal learning environment for all students, with equity and inclusion at the forefront, as well as the need for appropriate measures to capture these differences.

Research Goals:

This work seeks to construct a domain-specific self-efficacy beliefs instrument for undergraduate students rooted in the principles of reactivity in organic chemistry, rather than more general chemistry. This instrument aims to provide instructors with a valid and reliable means to capture changes in self-efficacy beliefs in organic chemistry with respect to three intended sub-dimensions: cognitive skills in organic chemistry, everyday application of organic chemistry, and experimental practices relating to organic chemistry. Creating a valid and reliable instrument to

assess self-efficacy beliefs in organic chemistry will allow for future work to evaluate the role of the curriculum on self-efficacy beliefs and what differences exist, if any, with respect to self-efficacy beliefs in organic chemistry for different underrepresented groups in Organic Chemistry.

Instrument development:

The development of the Self-Efficacy Beliefs in Organic Chemistry (SEBOC) instrument was guided by the principles of self-efficacy instrument design as described by Bandura (2006), while adhering to the standards for validity and reliability put forward by Standards of Educational and Psychological Testing (American Educational Research Association et al., 2014). The development of the SEBOC instrument consists of three phases to ensure sufficient validity, and reliability evidence was obtained, 1) Developing the initial scale, 2) Evaluating the intended dimensions within the instrument 3) Confirming the dimensions within the instrument (Figure 28). This thesis will describe only work related to phase one and phase two.

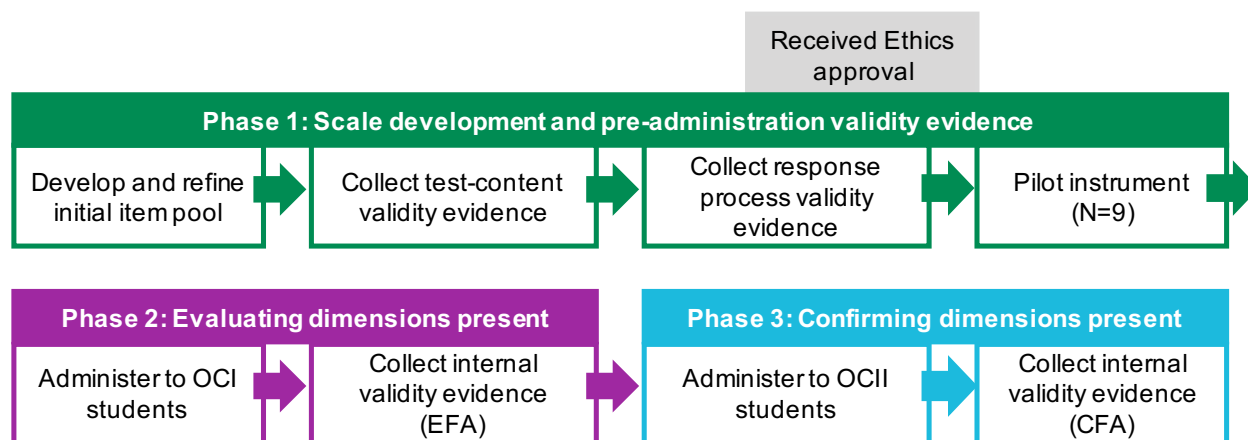


FIGURE 28. THREE PHASES OF DEVELOPMENT FOR SEBOC INSTRUMENT, OC = ORGANIC CHEMISTRY

Phase One: Initial Scale development and pilot study

During the preliminary stages of instrument construction two lenses guided the selection of the intended dimensions within the instrument: 1) learning outcomes from the “*Mechanistic patterns and principles*” curriculum 2) previous literature regarding self-efficacy beliefs in chemistry (Dalgety et al., 2003; Uzuntiryaki & Aydin, 2009; Villafañe et al., 2016; Winkelmann

et al., 2015) as well as other STEM related disciplines (Baldwin, Ebert-May, & Burns, 1999; Çalışkan et al., 2007; Shaw, 2004). The dimensions included within the SEBOC instrument aimed to target three separate, but highly interrelated, dimensions of self-efficacy beliefs in organic chemistry including 1) *cognitive skills relating to organic chemistry*, 2) *everyday applications of organic chemistry* and 3) *experimental practices relating to organic chemistry*.

The SEBOC instrument places a strong focus on the dimension of *cognitive skills relating to organic chemistry*. The 30 items included within this dimension describe domain-specific task relating to essential conceptual knowledge novices in the field typically learn at the beginning of their organic chemistry curriculum. These essential conceptual tasks serve as the foundational knowledge in the domain and underpin the entire curriculum, such as representations in organic chemistry, the symbolic language of organic chemistry, and principles of reactivity.

The second intended dimension included within the SEBOC, *self-efficacy towards everyday application of organic chemistry*, was modified from the College Chemistry Self-Efficacy scale (Uzuntiryaki & Aydin, 2009). As organic chemistry curricula evolve, so do increased discussions around the importance of situating chemistry in everyday life. Relating the relevance and usefulness of chemistry to everyday life is thought to influence students' attitudes towards chemistry (Treagust, Nieswandt, & Duit, 2000). Recent efforts illustrate the potential of a systems thinking approach to chemistry education and suggest the importance of highlighting the interconnected nature of chemistry concepts with broader systems such as biological or environmental systems (Mahaffy, Krief, Hopf, Mehta, & Matlin, 2018; Matlin, Mehta, Hopf, & Krief, 2016). As curricula shift towards incorporate system thinking, this subscale acts as an important measure to be able to capture the influence of these novel approaches.

The third intended dimension of the SEBOC instrument is rooted in *experimental practices relating to organic chemistry*. This dimension aims to extend the role of cognitive skills in organic chemistry to the application components commonly assessed outside of the theoretical components of the curriculum. This third dimension does not include self-efficacy with respect to psychomotor skills in experimental settings, but rather the self-efficacy towards the cognitive skills required in experimental settings and draws inspiration from the Methods of Biology factor from Baldwin et al. (1999), and Inquiry items from Winkelmann et al. (2015).

With the intended dimensions of the instrument established, an initial item pool was created applying the two previously described lenses, in addition to a third lens for the item development regarding the Science Practices and Crosscutting concepts defined by the *Next Generation Science Standards* (NGSS Lead States, 2013). The third lens was implemented in order to ensure the items contained in the instrument possessed a robust range of science practices that professions engage in as well as crosscutting concepts. Applying a method previously used for the evaluation of the "*Mechanistic Patterns and Principles*" curriculum (Raycroft & Flynn, 2019), the items included within instrument were characterized using the Three-Dimensional Learning Assessment Protocol (3D-LAP) (Lavery et al., 2016). As a result of this analysis, one item was modified to capture better key science practices of "Evaluating information". The item "Summarize the key points of a news report or documentary that deals with some aspects of organic chemistry" was transformed to "Critique a news report or documentary that deals with some aspects of organic chemistry." Additionally, the item "Write a relevant research question to address a problem in laboratory experiment" was added to capture the vital scientific practice of "Asking questions". The results of this analysis, as well as the 3D-LAP criterion, is included in Appendix D.

Items within the instrument were presented using a 7-point Likert scale (ranging from 1= “Not at all confident” to 7 = “Totally confident”). Traditionally, Bandura (1997, 2006) has described the standard methodology for measuring self-beliefs using a 100-point scale, with intervals of 10-units. This range aimed to ensure the instrument is sensitive and reliable. While some self-efficacy beliefs scales such as the College Chemistry Self-Efficacy (CCSS) scale (Uzuntiryaki & Aydin, 2009) have emulated this large range using a large 9-point scale, this large range is only represented by five categorical bins (*i.e.*, “very well” was captured with values 8, and 9). Recent work modifying the CCSS demonstrated an ability to maintain the intended factor structure using a condensed 5-point scale over the initial 9-point scale (Ferrell & Barbera, 2015). Additional, previous work has illustrated characteristics of data, such as distributions about the mean, skewness, or kurtosis, are not influenced by changes to scale length (*i.e.*, 5-point vs. 7-point vs. 10-point) supporting the use of a 7-point scale (Dawes, 2008).

Throughout the process of the instrument design, pre-administration validity evidence with respect to face, test-content, and response process was collected to ensure the instrument is assessing the intended construct (Arjoon et al., 2013). From the beginning of the instrument design, it is important to be incorporating aspects of validity evidence (Barbera & VandenPlas, 2011). Face validity reflects if the researcher believes the instrument is measuring the intended construct. This form of validity is often referred to as one of the weakest sources as it is a subjective measure (Barbera & VandenPlas, 2011). Face validity evidence was established through a literature review as well as informal discussions between authors and colleagues in the field of chemistry education. Face validity evidence collected led to a refinement of the initial item pool for clarity before moving forward to collecting evidence based on test content and response process validity.

Evidence based test-content acts as an indicator that dimensions, items, and instructional prompts are appropriately rooted in the intended measure conceptually (American Educational Research Association et al., 2014; Arjoon et al., 2013; Zamanzadeh et al., 2015). For the context of this research, a panel of nine experts who self-identified in domains ranging from Chemistry Education, Education, Organic Chemistry (Figure 29) were recruited to provide electronic feedback on the instrument. Specifically, a methodology for quantitative feedback was implemented aiming to elicit whether the items were clear and relevant to the intended dimensions and asked for experts to judge the alignment between experts and the intended dimensions (Zamanzadeh et al., 2015). Additionally, a free-response section was provided to allow for additional comments, questions, or areas of concern.

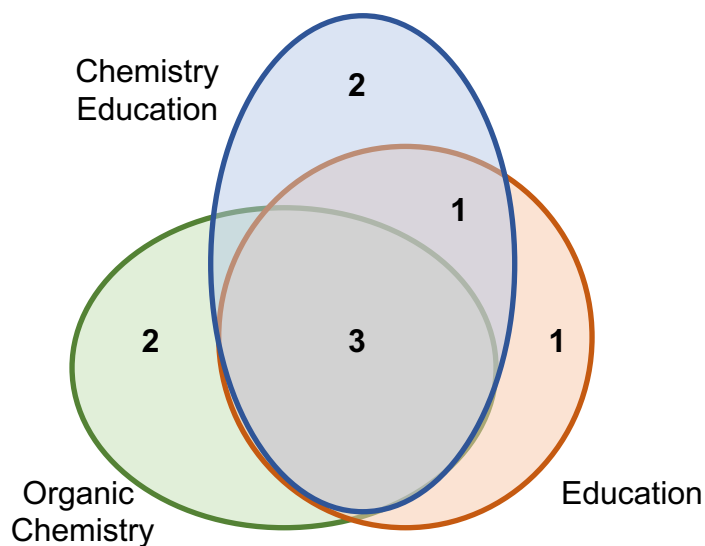


FIGURE 29. EXPERT POPULATIONS SELF-IDENTIFIED AREA(S) OF EXPERTISE

From the quantitative feedback and free-responses a total of 14-items were modified, to some degree, if experts perceived the item differently than intended, or to address potential shortcomings highlighted by the experts. For example, recurring feedback from experts asked for examples to be provided. In the item pool development, examples were used when appropriate to contextualize a term such as "chemical properties" as seen in the item "Use chemical properties

(e.g. electronegativity, atom size) to predict and justify the direction of an acid–base equilibrium”. The examples are included to guide participants, with the intent of keeping the emphasis focused on the broad task of the item and not on the use of a specific property such as electronegativity. Additionally, the expert review, the item “Predict the products of a reaction, given the starting materials” was flagged by experts for lack of an example, with concerns of being too vague for students to measure their self-efficacy beliefs towards this task. Throughout their organic chemistry curriculum, students encounter countless problems asking them this same task for a specific reaction. Rather than associating it with a specific reaction or group of reactions, the item was modified to capture better their self-efficacy beliefs about their abilities to use the underlying principles of reactivity in organic chemistry as follows “*Figure out the products of an unfamiliar reaction, given the starting materials.*”

Cognitive interviews with the students who had completed Organic Chemistry II were performed to establish response process validity, ensuring the target population perceive the items in the instrument as intended, addressing any ambiguity student perceive about items, as well as if items contained potential unseen redundancies. From the semi-structured cognitive interviews, two items were modified. For example, the item “visualize a molecule in three dimensions, given the structure” was modified as the item elicited discussion regarding the use of molecular modelling kits. While an important skill, using molecular modelling kits act more like a psychomotor skill, while the item aimed to elicit the cognitive skills required for understanding a molecular representation. To better capture this, the item was modified to “*Form a 3-dimensional mental image of a molecule*”

With the limited number of cognitive interviews, a small pilot was performed to pretest the items as recommended by Bandura (2006). This pilot consisted of Organic Chemistry II students

($N=9$) enrolled in the 2019 summer course using a traditional format to deliver the content at a large Canadian institution. Participants completed the 42-item instrument (complete instrument available as Appendix E) approximately a week before their final examination. Students were incentivized to participate with an opportunity to win one of two 50\$ Amazon gift cards. The motivation for completing the small pilot study was to highlight any items that were lacking the ability to differentiate between participants (*e.g.*, items where all participants selected the maximum efficacy). This pilot ($N=9$) study found no items lacking an ability to capture gradients of challenge, meaning no item led to a strong majority selecting one of the seven levels of confidence and did not lead to the elimination of any items.

Phase Two: Assessing the validity evidence relating to the internal structure

Instructional context and participant sample

Participants of phase two had recently completed a one-semester Organic Chemistry I course at a large bilingual Canadian university. Participants completed the 42-item instrument online hosted on SurveyMonkey at their leisure, with an average completion time of eight minutes (“SurveyMonkey,” 2019). As done in phase 1 pilot, students were incentivized to participate with an opportunity to win one of two 50\$ Amazon gift cards. The Organic Chemistry I course is a requirement for a diverse set of programs, including chemistry focused programs (*i.e.*, Chemistry, Biochemistry, Chemical Engineering), Biomedical Sciences, Health Sciences, Psychology, etc. Participants ($N=78$) were recruited from the all Organic Chemistry I course sections (approx. 1800 students) through their shared laboratory components after final course grades had been finalized. Demographic information was self-reported by 96% of participants, revealing the participant sample possessed a greater sample of females (75%) to males (25%).

Additionally, the sample possessed a strong component of students that did not classify English as their first language (36%, with 26% of this being accounted for by French as a first

language students). Participants of the study came from diverse academic programs, with 39% being enrolled in a biomedical or biopharmaceutical science program, 23% enrolled in a chemistry focused program (i.e. Chemistry, Biochemistry, or Chemical engineering), 36% stemming from other programs such biology, health sciences, psychology, etc., and lastly 7% opting to not disclose their program. The majority (50%) of participants were found to predominantly choose to identify as Caucasian, with 37% of the population identifying as Asian (including Chinese, Filipino, Korean, South Asian, Southeast Asian, and West Asian), 14% of the population identifying as Arab, 8% of the population identifying as Latin American, and 4% of the population identifying as Black, and 3% of the population identifying as indigenous. Of these students, 84% identified as a single group, with 16% choosing to identify with 2 or more groups.

Investigating the factor structure of an instrument

With the preliminary instrument construction completely, phase two and three aim to address additional elements of validity, specifically assessing the dimensionality through factor analysis. Understanding the dimensionality of a test is essential to ensure items are scored appropriately and producing valid data (Furr & Bacharach, 2014). While other tests that assess self-efficacy beliefs in chemistry were found to be unidimensional, such as the chemistry self-efficacy subscale within the CAEQ (Dalgety et al., 2003), the SEBOC instrument aims to outline three dimensions. As each of these dimensions represents distinct aspects of self-efficacy in organic chemistry, understanding the factor structure improves the mechanism in which we can use the scores. An inaccurate understanding of the dimensions of the instrument can lead to empirically and psychologically meaningless score. (Furr & Bacharach, 2014) Consider the following example in the context of the previously described, multidimensional College Chemistry Self-efficacy Scale (CCSS) (Uzuntiryaki & Aydin, 2009); suppose a student completes the CCSS

and receives an overall average self-efficacy belief in chemistry based on a cumulative score (e.g. an average score of 5/9). When interpreting a multidimensional test, each dimension contains essential information with respect to the score. When investigating only the cumulative scores, we can hypothesize two theoretical explanations: 1) the participant possesses an average self-efficacy belief score across all dimensions, or 2) the participant possesses strong self-efficacy beliefs with respect to psychomotor skills, and everyday applications but weak self-efficacy beliefs towards cognitive skills. Using factor analysis, internal structure validity evidence is collected to define these dimensions. When drawing conclusions from the scores, it is necessary to consider the individual dimensions present within the scale, rather than the simply overall values.

While the intended dimensions of an instrument can be rooted in theory, factor analysis is a statistical approach which establishes the degree to which items adhere to the hypothesized dimensionality (Arjoon et al., 2013). While weight and temperature can be directly measured, self-efficacy beliefs require indirect measures through multiple variables aiming to address the underlying construct. Factor analysis examines the relationship between the items, assessing if the items are related to the underlying construct (or latent variable) that cannot be directly measured. As previously described, this approach aids in the identifications clusters of items that are highly correlated together, which in turn establish the “Factors” (*i.e.*, the dimensions of the instrument) (Field, 2013; Ford, MacCallum, & Tait, 1986; Furr & Bacharach, 2014; Osborne et al., 2011). The factor loadings characterize the relative weight each item has towards that specific factor. While each factor shares a common variance, more commonly, the unique variance is discussed, which reflects the variance per item (Field, 2013).

When conducting an exploratory factor analysis, there are five essential steps, as described in Figure 30. (B. Williams, Onsman, & Brown, 2010). The following will discuss these factors in the context of the exploratory factor analysis for the SEBOC instrument.

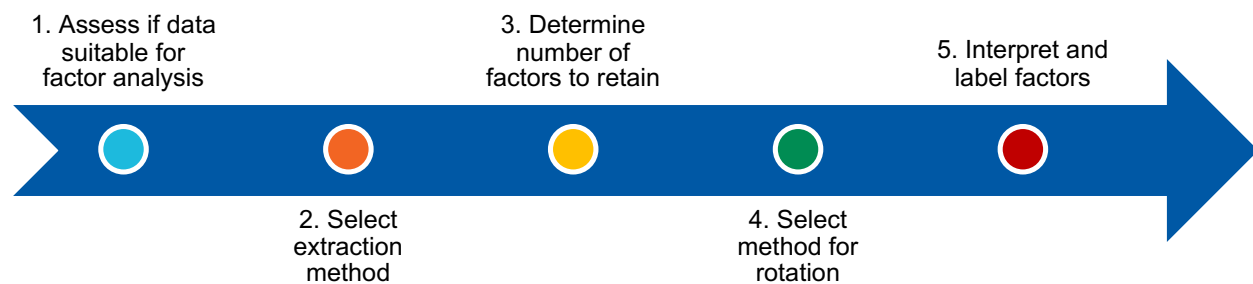


FIGURE 30. ESSENTIAL STEPS DURING EXPLORATORY FACTOR ANALYSIS

There is little agreement when determining the appropriate if a specific sample size is adequate for factor analysis (Osborne et al., 2011). Available approaches include heuristic rules of thumb stating that specific sample sizes such as a sample greater than 500 are very good (Comrey & Lee, 2013). Alternatively, specific sample to variable ratios have also been used to determine the appropriate sample size required, such as 5 participants:1 item (Comrey & Lee, 2013), or 10:1 (Nunnally, 1967). With respect to both of these approaches, Hogarty, Hines, Kromrey, Perron, & Mumford (2005) indicated that there is no minimum or specific ratio that will achieve a good factor analysis. Following many of the guidelines currently available (Comrey & Lee, 2013; Costello & Osborne, 2009; Guadagnoli & Velicer, 1988; Hair, Anderson, Tatham, & Black, 2014; Hogarty et al., 2005), the limited sample size of 78 participants recruited, and ratio of approximately 2:1 would be considered too low for factor analysis. As a result of the limited sample, the results of the exploratory factor analysis should be interpreted with extreme caution and require future analysis with a larger sample size to confirm factors identified. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is a test which can be performed before extraction producing a

value between 0 and 1 (Field, 2013). This value indicates the ratio of variance in the variables of the instrument that might be due to a common factor, (Hair et al., 2014). Values below 0.5 are referred to as Merde and deemed unacceptable for factor analysis, while values greater than 0.9 being noted as Marvellous. Alternatively, a value lower than 0.5 indicates there is a relatively large partial correlation compared to the sum of correlations which indicates a problem for factor analysis (Hutcheson, 2011b). An additional objective measure implemented to determine if the data is suitable for factor analysis is Bartlett's Test of Sphericity. This test provides us with a value indicating whether the produced correlation matrix is significantly different from an identity matrix (Field, 2013). For this test, we aim to reject the null hypothesis that the correlation matrix is the same as the identity matrix. While this is often the case as large sample sizes are more likely to produce significant correlations, the small sample size included in this research makes this less likely to occur (Field, 2013; Gorsuch, 2015; B. Williams et al., 2010).

Factor extraction aims to reduce the variables down, identifying the number of latent variables within the instruments by partitioning the shared variance from the unique variance and error variance to generate the underlying factor structure (Osborne et al., 2011). The most common mechanisms for extraction are principal component analysis (PCA) and factor analysis (Field, 2013). PCA was historically favoured as there was a lack of computing power to proceed with factor analysis (Gorsuch, 2015). An assumption in component analysis is that variables that contribute towards a component indicating the items cause or define the component. Alternatively, factor analysis assumes that the variables are a result of the underlying factor (Field, 2013). While the two differ on a theoretical perspective, they have similar functions and lead to similar results, but selecting the wrong factor extraction method can have negative consequences on the interpretability of results. Factor analysis includes only the shared variance, where PCA does

differentiate between shared and unique variance (Costello & Osborne, 2009; Gorsuch, 1988). While there exist multiple methods of factor analysis extraction, there is little consensus on why one is favourable over another (Osborne et al., 2011); therefore, throughout this work principal axis factoring (PAF) was implemented. PAF is the most frequently used methods of factor analysis extraction, and does not require any assumptions of the data's distribution or normality (Beavers et al., 2013; Osborne et al., 2011).

Commonly when determining the number of factors to retain from the extraction the Kaiser Criterion is implemented. In the Kaiser criterion, factors that possess an eigenvalue greater than, or equal to one are retained (Beavers et al., 2013). This was proposed as an eigenvalue greater than one will account for a greater amount of variance than a single variable, but this method been criticized for retaining an inaccurate number of factors (Beavers et al., 2013; Field, 2013; Hair et al., 2014; Osborne et al., 2011; B. Williams et al., 2010). Results from the Kaiser Criterion are suggested to be accurate with sample sizes above 250 and an average communality (*i.e.*, shared variance) of greater than 0.6 (Field, 2013). Another consideration with respect to communality arises if specific items have communalities below 0.4; these low values may indicate it is not be related to the construct and should be reconsidered for the instrument.

Alternative methods for determining the number of factors to retain include the graphical representation of eigenvalues on the y-axis and factors on the x-axis (referred to as a Scree Plot) and parallel analysis. To ensure the appropriate number of factors are retained, both of these approaches were implemented in conjunction with the Kaiser Criterion for a robust analysis of the potential number of factors to retain. The first factor accounts for the highest amount of variance on the scree plot, with the following factors account for decreasing amounts (Osborne et al., 2011). When investigating the number of factors to retain, the cut-off is those that occur before a point of

inflection often referred to as the “elbow”. This point of inflection indicates an area where the slope changes within the graph (Field, 2013; Hair et al., 2014). Within the scree plot, the point of inflection can be difficult to identify, and multiple points can exist, making it unclear how many factors to extract. Commonly, multiple factor analysis will be run with the different numbers of factors extracted, and the factor structures will be compared to identify which is the most conceptually and empirically logical (Hair et al., 2014). A logical factor structure will have few factor loadings below 0.3, with no factors containing less than three items and little cross-loading between factors (Field, 2013). Cross loading is observed by a single item possessing a factor loading of greater than 0.3 on multiple factors. This value of 0.3 indicating a substantive loading can vary based on the sample size, for example, the value of 0.3 is typically set for samples of 350 whereas a sample of 70 requires a loading of 0.65 (Hair et al., 2014). This work retains values greater than 0.3 rather than 0.65 for reference purposes to literature benchmark values recommends extreme caution again in interpreting all factor loadings below 0.65. If any of the aforementioned problems occur in the factor structure, it can be beneficial to drop the problematic items and rerun the analysis. The scree plot is a subjective measure for factor retention requiring the researcher to confirm where the point of inflection lies and has been suggested to be less reliable of a measure when there are less than 200 participants, indicating the results should again be interpreted with caution (Field, 2013). Parallel analysis is a less common approach for factor retention as it is not commonly included within statistical software but has been described as the best method for determining factor retention (Thompson, 2004). Parallel analysis is an objective measure that compares the actual eigenvalues extracted from the data against a set of eigenvalues produced from a randomly generated correlation matrix (Patil, Singh, Mishra, & Donavan, 2007; B.

Williams et al., 2010). In this approach, the number of factors to retain is identified by the number of actual eigenvalues than are greater than the randomly generated eigenvalues.

In order to better interpret the extracted factor structure, a technique called factor rotation is used on the initial extracted factor structure (Costello & Osborne, 2009). The initial, unrotated factor structure can suggest that items load heavily on to one or two factors with little to no items loading heavily on to the remaining factors, by rotating the factors it maximizes the loadings of items on each factor (Hair et al., 2014). Although rotation clarifies the factor structure, it cannot improve on things such as variance extracted. There exist two types of rotation, orthogonal and oblique rotations used to improve the initial factor structure. Orthogonal rotations are commonly used when factors are believed to be unrelated to each other; however, rarely are factors within social science research. Oblique rotations allow for a more realistic modelling as this mechanism for factor rotation as it accounts for any correlations that may exist, whether significantly correlated or not unrelated (Field, 2013; Hair et al., 2014). Various methods of oblique rotation exist, including Direct Oblimin and Promax (Gorsuch, 2015), but little evidence has shown one to produce superior results and the method used is often dictated by the statistical software used (Beavers et al., 2013). While oblique rotations are often more appropriate and recommended, orthogonal rotations are more commonly used. One suggested rationale for this is the orthogonal rotations provide simpler solutions compared to complex interactions within an oblique rotation (Field, 2013). When interpreting the rotated factor structure, both the strength of items and the number of items in the factors is considered. If an item is cross-loaded significantly (i.e. > 0.3) on multiple factors, it is referred to as a complex item (Hair et al., 2014). These items can be retained and aligned with the factor that addresses the latent variable, and if the item loads more strongly on a single factor than the other (i.e., a difference in factor loadings of 0.2) but if items both load

equally, or poorly than it may be a problematic item that is not measuring the intended construct and may need to be eliminated from the instrument (Hair et al., 2014; Yong & Pearce, 2016).

Reliability evidence for a single administration is often collected as an estimate of internal consistency (American Educational Research Association et al., 2014). In this manner, items that are supposedly related and/or measuring the same construct, such as the subscales of the SEBOC instrument will be correlated. While more appropriate statistical measures have been proposed for determining single administration reliability, Cronbach alpha persists as the more commonly used approach in CER and was used in this research (Komperda, Pentecost, & Barbera, 2018). For an individual subscale to be regarded as reliable, a cut-off value of 0.7 is typically implemented, but this cut-off serves a guideline rather than a strict rule of thumb (Neuendorf, 2003). Instruments that possess high Cronbach alpha values have been criticized for producing instruments with high redundancy (Briggs & Cheek, 1986).

Phase Two: Exploratory factor analysis results

The exploratory factor analysis of the Organic Chemistry I participants in Phase two revealed three problematic items to the instrument, as well as a tentative 7-factor structure. Due to the limited sample size of 78 participants, the results described are to be interpreted cautiously and are to be further expanded by a larger administration. The KMO was investigated as an alternative measure to ensure we meet sampling adequacy, with a KMO value of 0.75, deemed Middling (Hutcheson, 2011). Additionally, we observed a Bartlett's test value that rejects the null hypothesis ($p < .000$), indicating there is a significant difference between the correlation matrix and the identity matrix. In conjunction, these two tests indicate we can proceed with the factor analysis, but again due to the small sample size, we are proceeding with extreme caution.

Factors were extracted using principal axis factoring, with the Kaiser criterion of the eigenvalues for the 42-item instrument indicating to retain a total of 11 factors, accounting for 67%

of the variance and an average communality of greater of 0.67 with no items below 0.4. As previously discussed, the Kaiser criterion may not provide an accurate number of retained factors in our context, as such, multiple methods were investigated in conjunction with the Kaiser criterion approach. The Scree plot revealed multiple points of inflection indicating to retain either 2 factors, 7 factors, or 11 factors (Figure 31). Lastly, the parallel analysis comparing the randomly generated eigenvalues to the actual eigenvalues indicated to remain to a total of 7 factors (Table 5). A 7-factor structure was thus investigated, meeting the criteria for KMO (.72) and rejecting the null hypothesis for Bartlett's test ($p = .000$) indicating the correlation matrix is different from the identity matrix. Within the 7-factor structure, though, three items (items 4, 6, and 18) were found to possess communalities below 0.4 and were eliminated from the factor structure before proceeding.

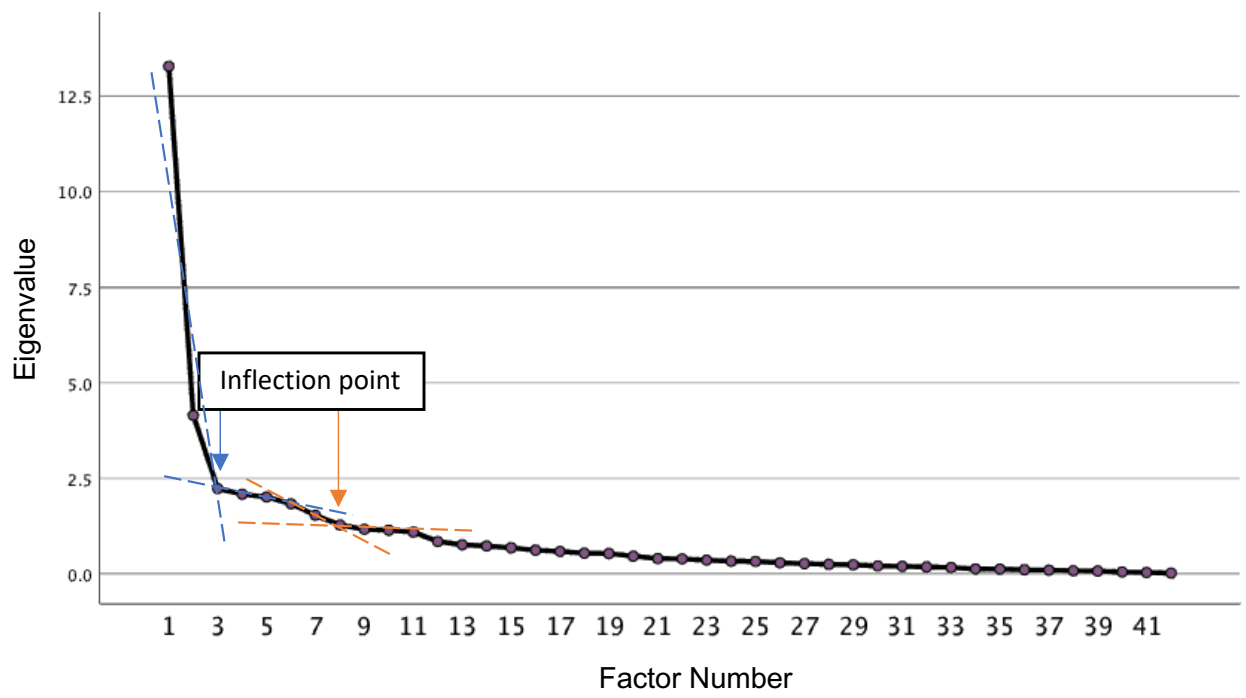


FIGURE 31. SCREE PLOT INDICATING TO RETAIN 2 OR 7 FACTORS

TABLE 5. PARALLEL ANALYSIS INDICATING TO RETAIN 7 FACTORS

Factor Number	1	2	3	4	5	6	7	8	9	10	11
Actual eigenvalue	13.275	4.147	2.231	2.087	2.014	1.834	1.535	1.281	1.169	1.143	1.097
Randomly generated eigenvalue	2.555	2.302	2.084	1.911	1.757	1.641	1.518	1.402	1.304	1.201	1.109

This modified 39 item, 7 factor structure was confirmed again through KMO values (.75) and Bartlett test ($p = .000$), account for 59% of the variance with no items containing a communality below 0.4. The intended 3 factor solution was expanded to include a total of 7 factors, within this solution, one item (item 38) possessed no significant factor loading, 12 items cross-loading on two factors, and one item (item 42) cross-loaded on three factors. Items that do not load significantly on any factor indicate problematic items to be eliminated. Additionally, cross-loading items indicates problematic items that may need to be deleted or revised for better alignment with the assigned dimension. After careful analysis of the items possessed within each factor, the cognitive factors were determined to be broken up into four subscales, while the two additional intended dimensions of everyday applications and experimental practices were moderately conserved

TABLE 6. SEBOC FACTOR LOADINGS (FACTOR LOADINGS >0.3 OMITTED). SEBOC ITEMS COLOR CODING WITH HAVE BEEN COLOR CODED WITH INTENDED DIMENSIONS, COGNITIVE, EVERYDAY APPLICATIONS AND EXPERIMENTAL PRACTICES), FACTOR LOADINGS COLOR CODED WITH NEW DIMENSIONS)

SEBOC Item	Factor						
	1	2	3	4	5	6	7
Self-efficacy beliefs towards structure and function							
8. Identify the most acidic proton in a molecule.	0.52			0.33			
9. Identify the most basic atom in a molecule.	0.56						
21. Compare two potential mechanisms for a reaction and justify which is the most likely to occur.	0.4						

24. Compare two nucleophilic or electrophilic sites and justify which is most likely to participate in the reaction.	0.62			
26. Describe how to manipulate an equilibrium to favour either products or reactants.	0.51			
27. Describe what happens at the molecular level when the position of an equilibrium is changed.	0.37		0.42	
28. Identify similarities between multiple reactions based on their underlying mechanisms.	0.4		0.34	
30. Use chemical properties (e.g. electronegativity, atom size) to predict and justify the direction of an acid–base equilibrium.	0.53			
35. Propose the stereochemical outcome for a reaction.	0.36			
Self-efficacy beliefs towards everyday applications of organic chemistry				
3. Describe how organic chemistry is involved in a global issue (e.g. societal, environment), either positively or negatively.	0.68			
10. Explain aspects of your everyday life using organic chemistry.	0.9			
20. Critique a news report or documentary that deals with some aspects of organic chemistry.	0.61			
25. Design a laboratory procedure for a reaction learned in your organic chemistry course.	0.54	0.4		
29. Explain how a profession relates to organic chemistry, if at all.	0.45		0.32	
42. Apply concepts learned in organic chemistry to other sciences.	0.35		0.3	0.35
Self-efficacy beliefs towards spectroscopy				
33. Interpret spectroscopic data (e.g., IR, NMR) to determine the structural information of a molecule.		0.8		
41. Explain how chemical properties (e.g., bond strength, electronegativity) influence spectroscopic data (e.g. IR, NMR).		0.78		
Self-efficacy beliefs towards organic mechanisms				
1. Determine the product of a single-step reaction, given the starting materials and mechanistic arrows.		0.56		
2. Draw the mechanistic arrows for a multi-step reaction, given the starting materials, intermediates, and products.		0.65		
12. Explain a concept learned in organic chemistry to a peer.		0.56		
13. Propose a retrosynthetic analysis for a target compound.		0.73		
14. Figure out the products of an unfamiliar reaction, given the starting materials.		0.75		
15. Propose a mechanism and the products for a reaction, given the starting materials.		0.83		
16. Propose a mechanism for a reaction, given the starting materials and products.		0.82		
17. Connect steps of the experimental procedure to concepts of reactivity.		0.36		0.34
40. Propose a synthesis for a target compound.	0.45	0.47		

Self-efficacy beliefs towards experimental practices in organic chemistry		
7. Write a relevant research question to address a problem in a laboratory experiment.	0.4	0.5
19. Record and tabulate raw data collected in a laboratory experiment.		0.6
22. Interpret and draw conclusions from data collected in a lab experiment.		0.66
34. Determine how experimental procedures and errors reflect data collected in a laboratory experiment.		0.78
Self-efficacy beliefs towards structural representations		
11. Visualize how molecules are interacting in space as a reaction proceeds.		0.36
32. Determine the stereoisomeric relationship (e.g., enantiomers, diastereomers, the same molecule) between two structures, if any.	-0.32	0.68
36. Propose the regiochemical outcome for a reaction.		0.34
38. Explain the relationship between the conformation and relative stability of a molecule.	0.32	0.41
39. Convert between various representations of a molecule (e.g. Lewis structure, line structure, Newman Projection).		0.57
Self-efficacy beliefs towards orbitals		
5. Assign the orbital hybridization (e.g., sp^2 , sp^3) of each atom in a molecule.		0.74
23. Identify nucleophilic or electrophilic sites in a molecule.	0.35	0.34
31. Justify the outcome of a reaction using orbitals.		0.32 0.46
37. Explain how a reaction coordinate diagram describes an associated reaction mechanism.	—	
Omitted items:		
4. Form a 3-dimensional mental image of a molecule.		
6. Predict trends in physical properties (e.g., boiling point, solubility) or an organic compound based on chemical properties (e.g. molecular size, polarity).		
18. Draw the resonance structures of a molecules.		

Factors 1–6 were found to possess high internal reliability (<0.7), with factor 7 possessing a Cronbach alpha value slightly below the acceptable cut-off of 0.7. The determined factors are outlined in Table 7 as well as their respective Cronbach alpha values.

TABLE 7. SEBOC INSTRUMENT RELIABILITY EVIDENCE

Factor	Number of items	Cronbach α
Self-efficacy beliefs towards...		
Structure and function	9	0.89

Everyday applications	5	0.82
Spectroscopy	2	0.82
Organic mechanisms	9	0.91
Experimental practices	4	0.75
Structural representation	5	0.76
Orbitals	3	0.69

Other areas of concern, aside from the reliability of Factor 7 exist in the above factor structure. With the ideal factor structure possessing a minimum of 3 contributing items, factor 3 fails to meet this criterion. Factor 3 revolving around spectroscopy possesses high factor loadings, but with only two items captures a limited amount variance and is thought to be weak and unstable (Hair et al., 2014). A rationale for these items appears as a unique factor can stem from the University of Ottawa's "Mechanistic patterns and principles" curricular context, where spectroscopy is not a topic covered in Organic Chemistry I. Therefore, the participants who had recently completed Organic Chemistry I may not have been able to integrate these tasks with the other tasks. These spectroscopy focused items appearing as a unique factor serves as an indicator that these tasks may not be appropriate for an introductory organic level, limiting the generalizability of the instrument due to variability in instruction contexts. As well, many factors possess complex items, that load on multiple factors with loadings below .5. These results suggest potentially problematic items that may be areas to be removed in future work. While the above results contribute internal validity evidence for the SEBOC instrument, due to the limited sample size, these results are interpreted with extreme cautions.

With the tentative factor structure present, we can examine the descriptive statistics of each factor as well as the relationship between factors to understand better how students were engaging with the instrument and the levels of self-efficacy beliefs captured. We can compare individual

factors scores using descriptive statistics, including the average, standard deviation, and range (Table 8). The descriptive statistics demonstrate students at the end of their Organic Chemistry I course, on average, possessed elevated (i.e. greater than self-efficacy 4 or "somewhat confident") for the majority of the dimensions in the instrument. The low self-efficacy found in Factor 3 relating to spectroscopy can stem from students having little exposure to the content and further supports their removal moving forward. Due to the cautious interpretation of the factor structure, inferential statistics were not applied to investigate further differences participants possess concerning each factor and are to be assessed as part of further work.

TABLE 8. SEBOC FACTORS DESCRIPTIVE STATISTICS

Factor	Avg. factor score /7	SD	Range
1. Structure and function in organic chemistry	4.8	1.2	1.8–7.0
2. Everyday applications of organic chemistry	4.3	1.0	1.7–6.0
3. Spectroscopy	3.8	1.1	1.4–6.1
4. Organic mechanisms	4.7	1.2	1.9–7.0
5. Experimental practices in organic chemistry	4.7	1.0	2.2–7.0
6. Structural representations	5.1	1.0	2.4–7.0
7. Orbitals	5.1	1.0	2.7–7.0

The correlation matrix of each factor (Table 9) provides insight into the relationship between the intended factors and the identified factors. Specifically, these findings demonstrate the moderate relationship between the cognitively rooted factors of 1 and 4 ($r = .44$), 1 and 6 ($r = .30$), and 4 and 7 ($r = .32$). Contrastingly, there no moderate correlations observed between any of the identified cognitive rooted factors and the intended everyday applications & experimental practices factors. Rather, those two factors possess the highest correlation with each other ($r = .29$). This relationship indicates a potential link between an individual's self-efficacy for everyday application and experimental practices, with both factors' asking students to apply their knowledge to a novel situation These findings provide a more detailed image of how the factor structure

determined using EFA the SEBOC instrument can be used to better capture self-efficacy beliefs in Organic Chemistry and the distinction between the cognitive factor and factors which required the application of an individual's cognitive abilities.

TABLE 9. SEBOC FACTORS CORRELATION MATRIX, CORRELATION < 0.1 OMITTED

Factor	1	2	3	4	5	6	7
1. Structure and function in organic chemistry	1						
2. Everyday applications of organic chemistry	.10	1					
3. Spectroscopy	.13	.12	1				
4. Organic mechanisms	.44		.22	1			
5. Experimental practices in organic chemistry	.17	.29	.13	.22	1		
6. Structural representations	.30		.11	.28	.17	1	
7. Orbitals	.25		.19	.32	.29	.15	1

Limitations

This work aimed to develop and validate an organic chemistry specific self-efficacy instrument, but currently only possesses limited validity evidence. The convenience sampling (L. Cohen, 2010a) used for the response process validity interviews provided a limited demographic sample of only white males. With the previous literature regarding differences in self-efficacy based on sex, we are currently unable to determine if the instrument would be interpreted consistently by all underrepresented populations in STEM. As previously described, many of the underlying assumptions required for an exploratory factor analysis were not met due to the limited sample size, meaning results should be interpreted with extreme caution. Additionally, the SEBOC instrument has not been implemented outside a "Mechanistic patterns and principles" curriculum. As described in Chapter three, while the instrument is validated for our population/context, additional validity evidence would be required before migrating the instrument to a traditional curriculum and using it as a means to evaluate between different curricular approaches.

Conclusions and Future work

This work described the approaches used when constructing the Self-Efficacy Beliefs in Organic Chemistry (SEBOC) instrument rooted in the theoretical principles of self-efficacy beliefs that possesses multiple sources of validity evidence. The 42-item instrument was initially developed to possess three inter-related subscales: cognitive skills in organic chemistry (30-items), everyday application of organic chemistry (6-items), and experimental practices relating to organic chemistry (6-items). Pre-administration validity evidence was collected with respect to test content validity using a panel of nine experts in diverse fields and response process validity through semi-structured interviews with the target population. The instrument was then refined to a 39-item 7 factor instrument dividing the cognitive skills to 5 subscales including self-efficacy beliefs towards structure and function (9 items), spectroscopy (2 items), organic mechanisms (9 items), structural representations (5 items), and orbitals (3 items) in organic chemistry, maintaining the majority of the subscales for everyday application of organic chemistry (5 items remained with 1 new item from the experimental practices) and experimental practices (4 items).

Future work aims to collect additional validity evidence supporting the internal structure of the instrument using a larger sample size. A large sample size will allow for the underlying assumptions for EFA to be met and provide stronger validity evidence in relation to the internal structure. With the factor structure solidified, a final implementation aims to confirm this factor structure. The constructed and validated SEBOC instrument will act as a tool for evaluating the influence of the "Mechanistic patterns and principles" curriculum on self-efficacy beliefs; investigating if students with increased abilities to apply fundamentals of reactivity additionally possess increased self-efficacy beliefs around organic chemistry. Further work is required to understand how the SEBOC instrument is interpreted by a diverse population and those outside the curriculum before being used to investigate how different populations (e.g., genders,

racess/ethnicity, etc.) are being benefitted equally from the transformed curricular design. Overall, this instrument will serve as an essential tool effectively evaluating the influence of organic chemistry focused educational interventions on self-efficacy beliefs at the undergraduate level.

Chapter Five: Conclusions

The studies explored throughout this thesis investigate the role of two instruments to evaluate the influence of a redesigned, "Mechanistic patterns and principles" organic chemistry curriculum. While a traditional curriculum is organized using a functional group design, the transformed organic chemistry curriculum places emphasis on the underlying mechanisms governing the reactions. This approach aims to equip students to explain the chemical basis for reaction mechanisms better, predict the reactivity of unknown reactions and transfer principles of reactivity to other domains such as biochemistry. Studies investigating the role of the transformed curriculum with respect to content related variables, *i.e.*, expertise and academic performance, were explored in Chapter two and three, but as well the less commonly investigated but important non-content related outcomes of the curriculum as seen in Chapter four. In addition to investigating these content and non-content variables, this work aims to consistently establish validity and reliability evidence for these instruments to improve the trustworthiness and generalizability. These studies have broad impact across STEM providing theoretical practices and establishing transferable instruments for educators to apply within their own curriculum evaluations

The studies described in Chapter two and three investigated student's expertise in organic chemistry, as a reflection of their organization of knowledge, using a novel online categorization task. The online categorization task included two subsequent tasks, an open sort to evaluate how learners enrolled in organic chemistry natural choose to organize their knowledge and a closed sort to evaluate their ability to categorize reaction cards according to the underlying mechanistic patterns. The online categorization task was administered using an early and late administration to investigate how participants organization of knowledge with respect to their choices and ability changed as they progressed throughout the curriculum.

The study in Chapter two investigated these changes in the longitudinal participants (N=24) categorizations, qualitatively through Gephi visualizations and quantitatively looking at the levels of interpretation used and their match with expert (%). The qualitative Gephi analysis demonstrated a shift to participants categorizations becoming more organized and similar to those of an expert. Findings from the quantitative analysis demonstrated a positive shift in expertise in how participants chose to organize their knowledge in organic chemistry, reflected by increased use of process-oriented features over static features. Additionally, the closed sort revealed an increase in their ability to categorize reactions using the underlying mechanistic patterns. This analysis also revealed areas of concern where low ability is demonstrated around complex reactions or those where the key mechanistic step is not explicit. Overall, this study demonstrates the positive influence of the "Mechanistic patterns and principles" curriculum on the formation of expertise.

In Chapter three, we describe work investigating the relationship between these two categorization tasks (*i.e.*, the open and closed sort) to probe at the alignment between the different tasks eliciting participants expertise in organic chemistry. The relationship between categorizations task was expanded by investigated the alignment between participants ability in varied stakes, (*i.e.*, low stakes online task or high stakes major assessment tasks). Findings from the early administration showcase a small correlation between choice and ability, but as we progressed towards the late administration, the correlation ($r = .53$, $p = .000$) between choice and ability establishes an improved alignment between how participant choose to categorize and the ability to view expert patterns. Through investigating the relationship of the measure of expertise in the open (*i.e.* level of interpretation, /100) and closed (Match with expert, %) with the external measure of expertise (*i.e.*, academic performance) we established criterion validity of the

instrument, as well as the promise of predictive ability of the closed sort for identifying learner in organic chemistry who might need additional supports before a major assessment. Lastly, the reproducibility of the findings from Chapter two derived from the instrument and the reproducibility of high stakes categorization tasks were evaluated. This investigation revealed the findings from Chapter two are consistent with novices in Organic Chemistry II relying on static features, and a strong reliability co-efficient ($N=20$, $r = .67$, $p = .001$) with respect to the closed sorts, Match with expert, with a reproduction group of students who have been matched on z-scored final examination grades. Lastly, the item-based approach demonstrated the reproducibility of the previous findings of a high-stakes closed categorization tasks, by having consistent trends, with little difference in sorting distributions, between the categorizations tasks from year to year.

The last study shifted from evaluating cognitive aspects of learning organic chemistry to constructing an instrument with the specificity required to evaluate student's self-efficacy beliefs in organic chemistry. Self-efficacy is defined as an individual's perceived a person's beliefs in their capability to produce a given attainment (Bandura, 1997). This 42-item instrument was designed with three intended dimensions of cognitive skills in organic chemistry (30-items), everyday application of organic chemistry (6-items), and experimental practices relating to organic chemistry. All items were investigated for their alignment with the NGSS framework, using the 3D-LAP procedure. These dimensions provide a broad assessment of different facets of students' self-efficacy in chemistry. Test content validity evidence and response process validity evidence lead to the modification and refinement of the item pool before administration in a pilot study. Internal validity evidence was collected through an exploratory factor analysis with Organic Chemistry I students ($N=78$), refining the instrument to a 39-item 7 factor with 5 subscales relating to cognitive skills in organic chemistry including self-efficacy beliefs towards structure and

function (9 items), spectroscopy (2 items), organic mechanisms (9 items), structures (5 items), and orbitals (3 items) in organic chemistry, while the intended subscales for everyday application of organic chemistry (5 items remained with 1 new item from the experimental practices) and experimental practices (4 items) remained. Future work in aims to improve the internal structure validity using a larger independent administration to produce a more powerful internal factor structure using EFA, with phase three aiming to confirm the improved factor structure using CFA.

As new curricular interventions are implemented as a means to improve chemistry education at large, it is essential to evaluate the impact on the important stakeholders (*i.e.*, learners and instructors). Subsequently, this drives the need for trustworthy instruments designed to evaluate these impacts. The work described throughout this thesis establishes the positive influence of the “Mechanistic patterns and principles” curriculum, but moreover acts as a framework for future work to conduct a comparative analysis to highlight areas of strength and improvement within the transformed Organic Chemistry curriculum.

References

- Acton, W. H., Johnson, P. J., & Goldsmith, T. E. (1994). Structural Knowledge Assessment: comparison of referent structures. *Journal of Educational Psychology*, 86(2), 303–311. <https://doi.org/10.1037/0022-0663.86.2.303>
- Adedokun, O. A., & Burgess, W. D. (2012). Analysis of paired dichotomous data: a gentle introduction to the McNemar Test in SPSS. *Journal of MultiDisciplinary Evaluation*, 8(17), 125–131. Retrieved from http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/336
- Ainscough, L., Foulis, E., Colthorpe, K., Zimbardi, K., Robertson-Dean, M., Chunduri, P., & Lluca, L. (2016). Changes in biology self-efficacy during a first-year university course. *CBE Life Sciences Education*, 15(2), 1–12. <https://doi.org/10.1187/cbe.15-04-0092>
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Testing, & Joint Committee on Standards for Educational and Psychological. (2014). *Standards for educational and psychological testing*.
- Anderson, T. L., & Bodner, G. M. (2008). What can we do about ‘Parker’? A case study of a good student who didn’t ‘get’ organic chemistry. *Chemistry Education: Research and Practice*, 9(2), 93–101. Retrieved from <http://xlink.rsc.org/?DOI=b806223b>
- Andrew, S. (1998). Self-efficacy as a predictor of academic performance in science. *Journal of Advanced Nursing*, 27(27), 596–603. Retrieved from <https://s3.amazonaws.com/academia.edu.documents/46281582/j.1365-2648.1998.00550.x20160606-9081-2xfgak.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1508395518&Signature=MqYoJ71uR1jb7Y31gKWEBfpGJl%3D&response-content-disposition=inline%3B filename%3DSel>
- Arjoon, J. A., Xu, X., & Lewis, J. E. (2013). Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence. *Journal of Chemical Education*, 90(5), 536–545. <https://doi.org/10.1021/ed3002013>
- Baldwin, J. A., Ebert-May, D., & Burns, D. J. (1999). The development of a College Biology Self-Efficacy instrument for Nonmajors. *Science Education*, 83(4), 397–408. <https://doi.org/10.1002/1098-237>
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215. <https://doi.org/10.1037/0033-295X.84.2.191>
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. W.H. Freeman and Comapny. New York: Freeman. https://doi.org/10.1007/SpringerReference_223312
- Bandura, Albert. (1986). *Social foundations of thought and action: A social cognitive approach*. *Social Foundations of Thought and Action: A social cognitive approach*. Englewood Cliffs, NJ, US: Prentice-Hall, Inc.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. W.H. Freeman and Comapny. New York: Freeman. https://doi.org/10.1007/SpringerReference_223312
- Bandura, A. (2006). Guide for Constructing Self-Efficacy Scales. In *Self Efficacy Beliefs of Adolescents* (pp. 307–337). Retrieved from <http://www.uky.edu/~eushe2/Bandura/BanduraGuide2006.pdf>
- Bandura, A. (2015). The role of self-efficacy beliefs in motivation. In E. A. Locke & G. P. Lathan (Eds.), *New Developments in Goal Setting and Task Performance* (pp. 147–157). New York, NY, US: Routledge/Taylor & Francis Group.
- Barbera, J., Adams, W. K., Wieman, C. E., & Perkins, K. K. (2008). Modifying and validating the Colorado learning attitudes about science survey for use in chemistry. *Journal of Chemical*

- Education*, 85(10), 1435–1439. <https://doi.org/10.1021/ed085p1435>
- Barbera, J., & VandenPlas, J. R. (2011). Chapter 11: All assessment materials are not created equal: The myths about instrument development, validity, and reliability. In *ACS Symposium Series* (Vol. 1074, pp. 177–193). <https://doi.org/10.1021/bk-2011-1074.ch011>
- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Third International AAAI Conference on Weblogs and Social Media* (pp. 361–362). <https://doi.org/10.1136/qshc.2004.010033>
- Bauer, C. F. (2008). Attitude towards chemistry: A semantic differential instrument for assessing curriculum impacts. *Journal of Chemical Education*, 85(10), 1440–1445. <https://doi.org/10.1021/ed085p1440>
- Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using Exploratory Factor Analysis in educational research - Practical Assessment, Research & Evaluation. *Tabachnick & Fidell*, 18(6).
- Betz, N. E., & Hackett, G. (1983). The relationship of mathematics self-efficacy expectations to the selection of science-based college majors. *Journal of Vocational Behavior*. [https://doi.org/10.1016/0001-8791\(83\)90046-5](https://doi.org/10.1016/0001-8791(83)90046-5)
- Bissonnette, S. A., Combs, E. D., Nagami, P. H., Byers, V., Fernandez, J., Le, D., ... Tanner, K. D. (2017). Using the biology card sorting task to measure changes in conceptual expertise during postsecondary biology education. *CBE Life Sciences Education*, 16(1), 1–15. <https://doi.org/10.1187/cbe.16-09-0273>
- Bodé, N. E., Caron, J., & Flynn, A. B. (2016). Evaluating students' learning gains and experiences from using nomenclature101.com. *Chemistry Education Research and Practice*, 17(4), 1156–1173. <https://doi.org/10.1039/c6rp00132g>
- Bodé, N. E., & Flynn, A. B. (2016). Strategies of successful synthesis solutions: mapping, mechanisms, and more. *Journal of Chemical Education*, 93(4), 593–604. <https://doi.org/10.1021/acs.jchemed.5b00900>
- Bong, M., Cho, C., Ahn, H. S., & Kim, H. J. (2012). Comparison of self-beliefs for predicting student motivation and achievement. *The Journal of Educational Research*, 105(5), 336–352. <https://doi.org/10.1080/00220671.2011.627401>
- Bong, M., & Clark, R. E. (1999). Comparison between self-concept and self-efficacy in academic motivation research. *Educational Psychologist*, 34(3), 139–153. https://doi.org/10.1207/s15326985ep3403_1
- Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: how different are they really? *Educational Psychology Review*, 15(1), 1–40. <https://doi.org/10.1023/A:1021302408382>
- Boster, J. S., & Johnson, J. C. (1989). Form or function: a comparison of expert and novice judgments of similarity among fish. *American Anthropologist*, 91(4), 866–889. <https://doi.org/10.1525/aa.1989.91.4.02a00040>
- Bradley, A. Z., Ulrich, S. M., Jones, M., & Jones, S. M. (2002). Teaching the sophomore organic course without a lecture. Are you crazy? *Journal of Chemical Education*, 79(4), 514. <https://doi.org/10.1021/ed079p514>
- Brandriet, A. R., & Bretz, S. L. (2014). The development of the redox concept inventory as a measure of students' symbolic and particulate Redox understandings and confidence. *Journal of Chemical Education*, 91(8), 1132–1144. <https://doi.org/10.1021/ed500051n>
- Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, 54(1), 106–148. <https://doi.org/10.1111/j.1467->

6494.1986.tb00391.x

- Çalışkan, S., Selçuk, G. S., & Erol, M. (2007). Development of physics self-efficacy scale. *AIP Conference Proceedings*, 899(2007), 483–484. <https://doi.org/10.1063/1.2733247>
- Camacho, M., & Good, R. (1989). Problem solving and chemical equilibrium: Successful versus unsuccessful performance. *Journal of Research in Science Teaching*, 26(3), 251–272. <https://doi.org/10.1002/tea.3660260306>
- Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In *The Cambridge Handbook of Expertise and Expert Performance* (pp. 21–30). New York, NY, USA: Cambridge University Press. <https://doi.org/10.1017/CBO9780511816796.002>
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2), 121–152. https://doi.org/10.1207/s15516709cog0502_2
- Chi, M. T. H., Glaser, R., Farr, M., & Chi author, M. T. H. (1988). *The nature of expertise*. (M. J. Farr & R. Glaser 1921-2012, Eds.). Hillsdale, N.J.: L. Erlbaum Associates. <https://doi.org/10.4324/9781315799681>
- Choi, N. (2005). Self-efficacy and self-concept as predictors of college students' academic performance. *Psychology in the Schools*, 42(2), 197–205. <https://doi.org/10.1002/pits.20048>
- Clayden, J. J., Greeves, N., Warren, S., & Wothers, P. (2012). *Organic Chemistry*. Oxford University Press (2nd ed.). New York, NY, USA, NY, USA: Oxford University Press. <https://doi.org/10.1086/278635>
- Cobb, P., Confrey, J., DiSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, 32(1), 9–13. <https://doi.org/10.3102/0013189X032001009>
- Cohen, J. (1988). Set correlation and contingency tables. *Applied Psychological Measurement*, 12(4), 425–434. <https://doi.org/10.1177/014662168801200410>
- Cohen, L. (2010a). *Research methods in education*. *Research Methods in Education*. <https://doi.org/10.4324/9780203224342>
- Cohen, L. (2010b). *Research Methods in Education*. *Research Methods in Education*. <https://doi.org/10.4324/9780203224342>
- Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *Journal of General Education*, 57(2), 119–130. <https://doi.org/10.1353/jge.0.0018>
- Collins, J. L. (1982). Self-efficacy and ability in achievement behavior. *Paper Presented at the Annual Meeting of the American Educational Research Association*. New York. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Self-efficacy+and+ability+in+achievement+behavior#0>
- Comrey, A. L., & Lee, H. B. (2013). *A first course in factor analysis*. Psychology press.
- Condiotte, M. M., & Lichtenstein, E. (1981). Self-efficacy and relapse in smoking cessation programs. *Journal of Consulting and Clinical Psychology*. US: American Psychological Association. <https://doi.org/10.1037/0022-006X.49.5.648>
- Cook, L. J. (2017). Using concept maps to monitor knowledge structure changes in a science classroom. *ProQuest LLC*.
- Cooper, M. M., & Klymkowsky, M. (2013). Chemistry, life, the universe, and everything: a new approach to general chemistry, and a model for curriculum reform. *Journal of Chemical Education*, 90(9), 1116–1122. <https://doi.org/10.1021/ed300456y>
- Cooper, M. M., & Stowe, R. L. (2018). Chemistry education research—from personal empiricism

- to evidence, theory, and informed practice. *Chemical Reviews*, 118(12), 6053–6087. <https://doi.org/10.1021/acs.chemrev.8b00020>
- Cooper, M. M., Stowe, R. L., Crandell, O. M., & Klymkowsky, M. W. (2019). Organic chemistry, life, the universe and everything (OCLUE): a transformed organic chemistry curriculum. *Journal of Chemical Education*, acs.jchemed.9b00401. <https://doi.org/10.1021/acs.jchemed.9b00401>
- Costello, A. B., & Osborne, J. W. (2009). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, 10(7), 1–9. <https://doi.org/10.1109/IJCBS.2009.105>
- Council, N. R. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/13165>
- Creswell, J. W. (2012). *Educational research: planning, conducting, and evaluating quantitative and qualitative research*. *Educational Research*. <https://doi.org/10.1017/CBO9781107415324.004>
- Cruz-Ramírez De Arellano, D., Towns, M. H., Cruz, D., & Towns, M. H. (2014). Students' understanding of alkyl halide reactions in undergraduate organic chemistry. *Chemistry Education Research and Practice*, 15(4), 501–515. <https://doi.org/10.1039/c3rp00089c>
- Dalgaty, J., & Coll, R. K. (2006). Exploring first-year science students' chemistry self-efficacy. *International Journal of Science and Mathematics Education*, 4, 97–116.
- Dalgaty, J., Coll, R. K., & Jones, A. (2003). Development of chemistry attitudes and experiences questionnaire (CAEQ). *Journal of Research in Science Teaching*, 40(7), 649–668. <https://doi.org/10.1002/tea.10103>
- Dawes, J. (2008). Do data characteristics change according to the number of scale points used? an experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, 50(1), 61–104. <https://doi.org/10.1177/147078530805000106>
- Domin, D. S., Al-Masum, M., & Mensah, J. (2008). Students' categorizations of organic compounds. *Chemistry Education Research and Practice*, 9(2), 114–121. <https://doi.org/10.1039/b806226a>
- Drnovšek, M., Wincent, J., & Cardon, M. S. (2010). Entrepreneurial self-efficacy and business start-up: developing a multi-dimensional definition. *International Journal of Entrepreneurial Behaviour and Research*, 16(4), 329–348. <https://doi.org/10.1108/13552551011054516>
- Farrell, B., Richardson, L., Raman-Wilms, L., de Launay, D., Alsabbagh, M. W., & Conklin, J. (2018). Self-efficacy for deprescribing: a survey for health care professionals using evidence-based deprescribing guidelines. *Research in Social and Administrative Pharmacy*, 14(1), 18–25. <https://doi.org/10.1016/j.sapharm.2017.01.003>
- Ferla, J., Valcke, M., & Cai, Y. (2009). Academic self-efficacy and academic self-concept: Reconsidering structural relationships. *Learning and Individual Differences*, 19(4), 499–505. <https://doi.org/https://doi.org/10.1016/j.lindif.2009.05.004>
- Ferrell, B., & Barbera, J. (2015). Analysis of students' self-efficacy, interest, and effort beliefs in general chemistry. *Chemistry Education Research and Practice*, 16(2), 318–337. <https://doi.org/10.1039/c4rp00152d>
- Ferrell, B., Phillips, M. M., & Barbera, J. (2016). Connecting achievement motivation to performance in general chemistry. *Chemistry Education Research and Practice*, 17(4), 1054–1066. <https://doi.org/10.1039/c6rp00148c>
- Field, A. (2013). Exploratory factor analysis. In *Discovering Statistics Using IBM SPSS Statistics*

- (4th Edition). Sage Publications Ltd.
- Fischer, C., Zhou, N., Rodriguez, F., Warschauer, M., & King, S. (2019). Improving college student success in organic chemistry: impact of an online preparatory course. *Journal of Chemical Education*, 96(5), 857–864. <https://doi.org/10.1021/acs.jchemed.8b01008>
- Flynn, A. B. (2015). Structure and evaluation of flipped chemistry courses: organic & spectroscopy, large and small, first to third year, english and french. *Chemistry Education: Research and Practice*, 16, 198–211. Retrieved from <http://pubs.rsc.org/en/Content/ArticleLanding/2014/RP/C4RP00224E>
- Flynn, A. B. (2017). Flipped Chemistry Courses: Structure, Aligning Learning Outcomes, and Evaluation. In *Online Approaches to Chemical Education* (Vol. 1261, pp. 151–164). Washington, DC, USA, DC, USA: American Chemical Society. <https://doi.org/10.1021/bk-2017-1261.ch012>
- Flynn, A. B., Caron, J., Laroche, J., Daviau-Duguay, M., Marcoux, C., & Richard, G. (2014). nomenclature101.com: a free, student-driven organic chemistry nomenclature learning tool. *Journal of Chemical Education*, 91(11), 1855–1859. <https://doi.org/10.1021/ed500353a>
- Flynn, A. B., & 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. *Chemistry Education Research and Practice*, 18(1), 64–77. <https://doi.org/10.1039/c6rp00126b>
- Flynn, A. B., & Ogilvie, W. W. (2015). Mechanisms before reactions: a mechanistic approach to the organic chemistry curriculum based on patterns of electron flow. *Journal of Chemical Education*, 92(5), 803–810. <https://doi.org/10.1021/ed500284d>
- Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: a critical review and analysis. *Personnel Psychology*, 39(2), 291–314. <https://doi.org/10.1111/j.1744-6570.1986.tb00583.x>
- Franke, T. M., Ho, T., & Christie, C. A. (2012). The Chi-Square Test: often used and more often misinterpreted. *American Journal of Evaluation*, 33(3), 448–458. <https://doi.org/10.1177/1098214011426594>
- Furr, M., & Bacharach, V. (2014). Test dimensionality and factor analysis. *Psychometrics: An Introduction*, 71–98. https://doi.org/10.1111/j.1468-5922.2010.01872_2.x
- Galloway, K. R., Leung, M. W., & Flynn, A. B. (2018). A comparison of how undergraduates, graduate students, and professors organize organic chemistry reactions. *Journal of Chemical Education*, 95(3), 355–365. <https://doi.org/10.1021/acs.jchemed.7b00743>
- Galloway, K. R., Leung, M. W., & Flynn, A. B. (2019). Patterns of reactions: a card sort task to investigate students' organization of organic chemistry reactions. *Chemistry Education Research and Practice*, 20(1), 30–52. <https://doi.org/10.1039/c8rp00120k>
- Galloway, K. R., Stoyanovich, C., & Flynn, A. B. (2017). Students' interpretations of mechanistic language in organic chemistry before learning reactions. *Chemistry Education Research and Practice*, 18(2), 353–374. <https://doi.org/10.1039/c6rp00231e>
- Gephi.org. (2011). Gephi Tutorial Layouts. *Gephi Tutorial Visualization*, 1–37. Retrieved from <https://gephi.org/tutorials/gephi-tutorial-layouts.pdf>
- Gephi.org, & Gephi. (2017). Gephi Tutorial Quick Start. Retrieved September 17, 2018, from <https://gephi.org/users/quick-start/>
- Gibbons, R. E., & Raker, J. R. (2019). Self-beliefs in organic chemistry: Evaluation of a reciprocal causation, cross-lagged model. *Journal of Research in Science Teaching*, 56(5), 598–618. <https://doi.org/10.1002/tea.21515>

- Gormally, C., Brickman, P., Hallar, B., & Armstrong, N. (2009). Effects of inquiry-based learning on students' science literacy skills and confidence. *International Journal for the Scholarship of Teaching and Learning*, 3(2). <https://doi.org/10.20429/ijstl.2009.030216>
- Gorsuch, R. L. (1988). Exploratory factor analysis. In *Handbook of multivariate experimental psychology* (pp. 231–258). Springer.
- Gorsuch, R. L. (2015). *Factor analysis*. Routledge/Taylor & Francis Group.
- Graulich, N., & Bhattacharyya, G. (2017a). Investigating students' similarity judgments in organic chemistry. *Chemistry Education Research and Practice*, 18(4), 774–784. <https://doi.org/10.1039/c7rp00055c>
- Grove, N., & Bretz, S. L. (2007). CHEMX: An instrument to assess students' cognitive expectations for learning chemistry. *Journal of Chemical Education*, 84(9), 1524–1529. <https://doi.org/10.1021/ed084p1524>
- Grove, N. P., Hershberger, J. W., & Bretz, S. L. (2008). Impact of a spiral organic curriculum on student attrition and learning. *Education in Chemistry*, 9(2), 157–162. <https://doi.org/10.1039/B806232N>
- Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103(2), 265–275. <https://doi.org/10.1037/0033-2909.103.2.265>
- Hackett, G., & Betz, N. E. (1981). A self-efficacy approach to the career development of women. *Journal of Vocational Behavior*. [https://doi.org/10.1016/0001-8791\(81\)90019-1](https://doi.org/10.1016/0001-8791(81)90019-1)
- Hair, J. F. J., Anderson, R. E., Tatham, R. L., & Black, W. C. (2014). *Multivariate data analysis* (4th Ed.). Upper Saddle River, NJ, USA: Prentice-Hall, Inc.
- Hall, A. K., Bernhardt, J. M., & Dodd, V. (2015). Older adults use of online and offline sources of health information and constructs of reliance and self-efficacy for medical decision making. *Journal of Health Communication*, 20(7), 751–758. <https://doi.org/10.1080/10810730.2015.1018603>
- Harper, M. E., Jentsch, F. G., Berry, D., Cathy Lau, H., Bowers, C., & Salas, E. (2003). TPL-KATS-card sort: A tool for assessing structural knowledge. *Behavior Research Methods, Instruments, and Computers*, 35(4), 577–584. <https://doi.org/10.3758/BF03195536>
- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66(3), 393–416. <https://doi.org/10.1177/0013164405282485>
- Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education*, 88(1), 28–54. <https://doi.org/10.1002/sce.10106>
- Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Perron, J. M., & Mumford, A. K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, 65(2), 202–226. <https://doi.org/10.1177/0013164404267287>
- Hoskinson, A. M., Maher, J. M., Bekkering, C., & Ebert-May, D. (2017). A problem-sorting task detects changes in undergraduate biological expertise over a single semester. *CBE Life Sciences Education*, 16(2), 1–12. <https://doi.org/10.1187/cbe.16-05-0175>
- Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, 49(5), 505–528. <https://doi.org/https://doi.org/10.1016/j.jsp.2011.07.001>
- Hutcheson, G. (2011). *The multivariate social scientist. The Multivariate Social Scientist*. <https://doi.org/10.4135/9780857028075>

- IBM SPSS Statistics. (2019). Retrieved from <https://www.ibm.com/analytics/spss-statistics-software>
- Irby, S. M., Phu, A. L., Borda, E. J., Haskell, T. R., Steed, N., & Meyer, Z. (2016). Use of a card sort task to assess students' ability to coordinate three levels of representation in chemistry. *Chemistry Education: Research and Practice*, 17(2), 337–352. <https://doi.org/10.1039/C5RP00150A>
- Joët, G., Usher, E. L., & Bressoux, P. (2011). Sources of self-efficacy: An investigation of elementary school students in France. *Journal of Educational Psychology*, 103(3), 649–663. <https://doi.org/10.1037/a0024048>
- Johnstone, A. H. (1991). Why is science difficult to learn? Things are seldom what they seem. *Journal of Computer Assisted Learning*, 7(2), 75–83. <https://doi.org/10.1111/j.1365-2729.1991.tb00230.x>
- Johnstone, A. H. (2006). Chemical education research in Glasgow in perspective. *Chemistry Education Research and Practice*, 7(2), 49–63. <https://doi.org/10.1039/B5RP90021B>
- Johnstone, A. H. (2010). You can't get there from here. *Journal of Chemical Education*, 87(1), 22–29. <https://doi.org/10.1021/ed800026d>
- Khedhaouria, A., Gurău, C., & Torrès, O. (2015). Creativity, self-efficacy, and small-firm performance: the mediating role of entrepreneurial orientation. *Small Business Economics*, 44(3), 485–504. <https://doi.org/10.1007/s11187-014-9608-y>
- Komperda, R., Pentecost, T. C., & Barbera, J. (2018). Moving beyond Alpha: a primer on alternative sources of single-administration reliability evidence for quantitative chemistry education research. *Journal of Chemical Education*, 95(9), 1477–1491. <https://doi.org/10.1021/acs.jchemed.8b00220>
- Kozma, R. B., & Russell, J. (1997). Multimedia and understanding: Expert and novice responses to different representations of chemical phenomena. *Journal of Research in Science Teaching*, 34(9), 949–968. [https://doi.org/10.1002/\(SICI\)1098-2736\(199711\)34:9<949::AID-TEA7>3.0.CO;2-U](https://doi.org/10.1002/(SICI)1098-2736(199711)34:9<949::AID-TEA7>3.0.CO;2-U)
- Krieter, F. E., Julius, R. W., Tanner, K. D., Bush, S. D., & Scott, G. E. (2016). Thinking like a chemist: development of a chemistry card-sorting task to probe conceptual expertise. *Journal of Chemical Education*, 93(5), 811–820. <https://doi.org/10.1021/acs.jchemed.5b00992>
- Laverty, J. T., Underwood, S. M., Matz, R. L., Posey, L. A., Carmel, J. H., Caballero, M. D., ... Cooper, M. M. (2016). Characterizing college science assessments: The three-dimensional learning assessment protocol. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0162333>
- Leary, A. (1985). Self-Efficacy and Health. *Behav. Res. Ther.*, 23(4), 437–451.
- Lent, R. W., Lopez, F. G., & Bieschke, K. J. (1991). Mathematics self-efficacy: sources and relation to science-based career choice. *Journal of Counseling Psychology*, 38(4), 424–430. <https://doi.org/10.1037/0022-0167.38.4.424>
- Lin, S. Y., & Singh, C. (2010). Categorization of quantum mechanics problems by professors and students. *European Journal of Physics*, 31(1), 57–68. <https://doi.org/10.1088/0143-0807/31/1/006>
- Lindstrøm, C., & Sharma, M. D. (2011). Self-efficacy of first year university physics students: do gender and prior formal instruction in physics matter? *International Journal of Innovation in Science and Mathematics Education (Formerly CAL-Laborate International)*, 19(2), 1–19. <https://doi.org/10.5860/crln.66.10.7525>
- Lynch, D. J., & Trujillo, H. (2011). Motivational beliefs and learning strategies in organic chemistry. *International Journal of Science and Mathematics Education*, 9(6), 1351–1365.

- <https://doi.org/10.1007/s10763-010-9264-x>
- Mahaffy, P. G., Krief, A., Hopf, H., Mehta, G., & Matlin, S. A. (2018). Reorienting chemistry education through systems thinking. *Nature Reviews Chemistry*, 2, 126. Retrieved from <https://doi.org/10.1038/s41570-018-0126>
- Markham, A. N., & Lindgren, S. (2014). From object to flow: Network sensibility, symbolic Interactionism, and social media. In M. D. Johns, S.-L. S. Chen, & L. A. Terlip (Eds.), *Studies in Symbolic Interaction* (Vol. 43, pp. 7–41). <https://doi.org/10.1108/S0163-239620140000043012>
- Mason, A., & Singh, C. (2011). Assessing expertise in introductory physics using categorization task. *Physical Review Special Topics - Physics Education Research*, 7(2), 1–17. <https://doi.org/10.1103/PhysRevSTPER.7.020110>
- Mason, A., & Singh, C. (2016). Using categorization of problems as an instructional tool to help introductory students learn physics. *Physics Education*, 51(2). <https://doi.org/10.1088/0031-9120/51/2/025009>
- Matlin, S. A., Mehta, G., Hopf, H., & Krief, A. (2016). One-world chemistry and systems thinking. *Nature Chemistry*. <https://doi.org/10.1038/nchem.2498>
- Mayer, R. E. (2012). *Information processing*. (K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra, & J. Sweller, Eds.), *APA educational psychology handbook, Vol 1: Theories, constructs, and critical issues* (Vol. 1). Washington, DC, USA: American Psychological Association. <https://doi.org/10.1037/13273-004>
- McCauley, R., Murphy, L., Westbrook, S., Haller, S., Zander, C., Fossum, T., ... Anderson, R. (2005). What do successful computer science students know? An integrative analysis using card sort measures and content analysis to evaluate graduating students' knowledge of programming concepts. *Expert Systems*, 22(3), 147–159. <https://doi.org/10.1111/j.1468-0394.2005.00306.x>
- Moos, D. C., & Azevedo, R. (2009). Learning with computer-based learning environments: a literature review of computer self-efficacy. *Review of Educational Research*, 79(2), 576–600. <https://doi.org/10.3102/0034654308326083>
- National Research Council. (1994). *Learning, Remembering, Believing: Enhancing Human Performance*. Learning, Remembering, Believing. Washington, DC: The National Academies Press. <https://doi.org/10.17226/2303>
- National Research Council. (2000a). *How people learn: brain, mind, experience, and school: expanded edition*. (J. D. Bransford, A. L. Brown, & R. R. Cocking, Eds.). Washington, DC, USA, DC, USA: National Academies Press. <https://doi.org/10.17226/9853>
- National Research Council. (2000b). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. (J. D. Bransford, A. L. Brown, & R. R. Cocking, Eds.). Washington, DC, USA: National Academies Press. <https://doi.org/10.17226/9853>
- National Research Council. (2012). *Discipline-Based Education Research*. <https://doi.org/10.17226/13362>
- Neiles, K. Y. (2014). Measuring knowledge: Tools to measure students' mental organization of chemistry information. In D. M. Bunce & R. S. Cole (Eds.), *Tools of Chemistry Education Research* (pp. 169–189). Washington, DC, USA: American Chemical Society. <https://doi.org/10.1021/bk-2014-1166.ch010>
- Neuendorf, K. (2003). Internal consistency reliability: Can Cronbach's Alpha be too high? *Journal of Personality*, 1–3.
- NGSS Lead States. (2013). *Next Generation Science Standards: For States, by States*. Washington,

- DC: The National Academies Press. <https://doi.org/10.17226/18290>
- Nunnally, J. C. (1967). Psychometric theory. *Psychometric Theory*.
- O'Dwyer, A., & Childs, P. E. (2017). Who says organic chemistry is difficult? Exploring perspectives and perceptions. *Eurasia Journal of Mathematics, Science and Technology Education*, 13(7), 3599–3620. <https://doi.org/10.12973/eurasia.2017.00748a>
- Optimal Workshop. (2015). OptimalSort Online Card Sorting Software. Retrieved September 15, 2018, from <https://www.optimalworkshop.com/optimalsort>
- Osborne, J. W., Costello, A. B., & Kellow, J. T. (2011). Best practices in Exploratory Factor Analysis. In *Best Practices in Quantitative Methods* (pp. 86–99). <https://doi.org/10.4135/9781412995627.d8>
- Pajares, F. (1996). Self-Efficacy Beliefs in Academic Settings. *Review of Educational Research*, 66(4), 543–578. <https://doi.org/10.3102/00346543066004543>
- Pajares, F., & Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in mathematical problem-solving. *Contemporary Educational Psychology*, 20(4), 426–443. <https://doi.org/10.1006/ceps.1995.1029>
- Patall, E. A. (2012). The motivational complexity of choosing: a review of theory and research. In *The Oxford Handbook of Human Motivation*. <https://doi.org/10.1093/oxfordhb/9780195399820.013.0015>
- Patil, V. H., Singh, S. N., Mishra, S., & Donovan, D. T. (2007). Parallel analysis engine to aid determining number of factors to retain [Computer software]. Available from <Http://Smishra.Faculty.Ku.Edu/Parallelengine.Htm>.
- Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ). *ERIC Institute of Educational Science*, 76.
- Piperopoulos, P., & Dimov, D. (2015). Burst bubbles or build steam? entrepreneurship education, entrepreneurial self-efficacy, and entrepreneurial intentions. *Journal of Small Business Management*, 53(4), 970–985. <https://doi.org/10.1111/jsbm.12116>
- Randles, C. A., & Overton, T. L. (2015). Expert vs. novice: approaches used by chemists when solving open-ended problems. *Chemistry Education: Research and Practice*, 16(4), 811–823. <https://doi.org/10.1039/c5rp00114e>
- Rao Vishnumolakala, V., Qureshi, S. S., Treagust, D. F., Mocerino, M., Southam, D. C., & Ojeil, J. (2018). Longitudinal impact of process-oriented guided inquiry learning on the attitudes, self-efficacy and experiences of pre-medical chemistry students. *Qscience Connect*, 1, 1–12. <https://doi.org/10.5339/connect.2018.1>
- Raycroft, M. A. R., & Flynn, A. B. (2019). What works? What's missing? An evaluation model for science curricula using five lenses of learning outcomes. *Manuscript Submitted*.
- Reed, J. J., & Holme, T. A. (2014). The role of non-content goals in the assessment of chemistry learning. *ACS Symposium Series*, 1182, 147–160. <https://doi.org/10.1021/bk-2014-1182.ch009>
- Rosenthal, J. A. (1996). Qualitative descriptors of strength of association and effect size. *Journal of Social Service Research*, 21(4), 37–59. https://doi.org/10.1300/J079v21n04_02
- Rowe, M. B. (1983). Getting chemistry off the killer course list. *Journal of Chemical Education*, 60(11), 954. <https://doi.org/10.1021/ed060p954>
- Ruiz-Primo, M. A., Shavelson, R. J., & Shultz, S. E. (1997). On the validity of concept map-base assessment interpretations: an experiment testing the assumption of hierarchical concept maps in science. *CSE Technical Report* 455, 6511(310).

- <https://doi.org/10.1017/CBO9781107415324.004>
- Sawtelle, V. (2011). *A gender study investigating physics self-efficacy*. FIU Electronic Theses and Dissertations University. <https://doi.org/10.25148/etd.FI11120705>
- Sawtelle, V., Brewe, E., Goertzen, R. M., & Kramer, L. H. (2012). Identifying events that impact self-efficacy in physics learning. *Physical Review Special Topics - Physics Education Research*, 8(2). <https://doi.org/10.1103/PhysRevSTPER.8.020111>
- Schmid, S., Youl, D. J., George, A. V., & Read, J. R. (2012). effectiveness of a short, intense bridging course for scaffolding students commencing university-level study of chemistry. *International Journal of Science Education*, 34(8), 1211–1234. <https://doi.org/10.1080/09500693.2012.663116>
- Schunk, D. (2016). *Learning Theories: An Educational Perspective* (6th ed.). New York, NY: Pearson.
- Schunk, D. H., & DiBenedetto, M. K. (2016). Self-efficacy theory in education. In *Handbook of Motivation at School* (pp. 34–54). Retrieved from <https://www.routledgehandbooks.com/doi/10.4324/9781315773384.ch3>
- Schunk, D. H., & Usher, E. L. (2012). Social cognitive theory and motivation. In *The Oxford handbook of human motivation*. (pp. 13–27). New York, NY, US: Oxford University Press.
- Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: A learning progression on chemical thinking. *Chemistry Education Research and Practice*, 15(1), 10–23. <https://doi.org/10.1039/c3rp00111c>
- Shaw, K. A. (2004). The development of a physics self-efficacy instrument for use in the introductory classroom, 137(September 2004), 137–140. <https://doi.org/10.1063/1.1807273>
- Singh, C. (2009). Categorization of problems to assess and improve proficiency as teachers and learners. *American Journal of Physics*, 77(1), 73–80. <https://doi.org/10.1119/1.2990668>
- Skaalvik, E. M., & Skaalvik, S. (2004). Self-concept and self-efficacy: A test of the internal/external frame of reference model and predictions of subsequent motivation and achievement. *Psychological Reports*, 95(3 II), 1187–1202. <https://doi.org/10.2466/pr0.95.7.1187-1202>
- Smith, J. I., Combs, E. D., Nagami, P. H., Alto, V. M., Goh, H. G., Gourdet, M. A. A., ... Tanner, K. D. (2013). Development of the biology card sorting task to measure conceptual expertise in biology. *CBE Life Sciences Education*, 12(4), 628–644. <https://doi.org/10.1187/cbe.13-05-0096>
- Smith, M. U. (1992). Expertise and the organization of knowledge: Unexpected differences among genetic counselors, faculty, and students on problem categorization tasks. *Journal of Research in Science Teaching*, 29(2), 179–205. <https://doi.org/10.1002/tea.3660290207>
- Snyder, J. L. (2000). An investigation of the knowledge structures of experts, intermediates and novices in physics. *International Journal of Science Education*, 22(9), 979–992. <https://doi.org/10.1080/095006900416866>
- Spencer, D. (2009). *Card Sorting: Designing Usable Categories*. Card Sorting: Designing Usable Categories. New York, NY, USA: Rosenfeld Media.
- Stabback, P. (2016). What Makes a Quality Curriculum? *Current Critical Issues in Curriculum and Learning*, (2), 8–12. Retrieved from <http://unesdoc.unesco.org/images/0024/002439/243975E.pdf>
- Stains, M., & Talanquer, V. (2008). Classification of chemical reactions: Stages of expertise. *Journal of Research in Science Teaching*, 45(7), 771–793. <https://doi.org/10.1002/tea.20221>
- SurveyMonkey. (2019).

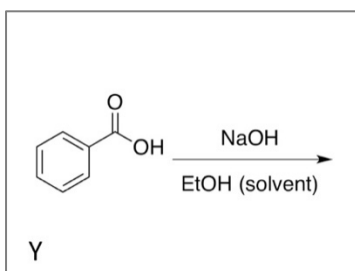
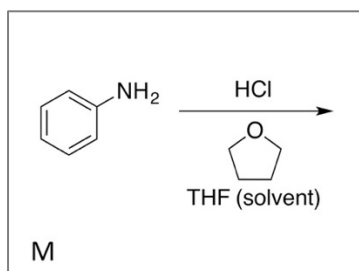
- Szu, E., Nandagopal, K., Shavelson, R. J., Lopez, E. J., Penn, J. H., Scharberg, M., & Hill, G. W. (2011). Understanding academic performance in organic chemistry. *Journal of Chemical Education*, 88(9), 1238–1242. <https://doi.org/10.1021/ed900067m>
- Talanquer, V., & Pollard, J. (2010). Let's teach how we think instead of what we know. *Chemistry Education Research and Practice*, 11(2), 74–83. <https://doi.org/10.1039/c005349j>
- Talanquer, V., & Pollard, J. (2017). Reforming a Large Foundational Course: Successes and Challenges. *Journal of Chemical Education*, 94(12), 1844–1851. <https://doi.org/10.1021/acs.jchemed.7b00397>
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC, US: American Psychological Association. <https://doi.org/10.1037/10694-000>
- Treagust, D., Nieswandt, M., & Duit, R. (2000). Sources of students difficulties in learning Chemistry. *Educación Química*, 11(2), 228. <https://doi.org/10.22201/fq.18708404e.2000.2.66458>
- Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, 17(7), 783–805. [https://doi.org/10.1016/S0742-051X\(01\)00036-1](https://doi.org/10.1016/S0742-051X(01)00036-1)
- Uzuntiryaki, E., & Aydin, Y. Ç. (2009). Development and validation of chemistry self-efficacy scale for college students. *Research in Science Education*, 39(4), 539–551. <https://doi.org/10.1007/s11165-008-9093-x>
- Villafañe, S. M., Garcia, C. A., & Lewis, J. E. (2014). Exploring diverse students' trends in chemistry self-efficacy throughout a semester of college-level preparatory chemistry. *Chemistry Education Research and Practice*, 15(2), 114–127. <https://doi.org/10.1039/c3rp00141e>
- Villafañe, S. M., Xu, X., & Raker, J. R. (2016). Self-efficacy and academic performance in first-semester organic chemistry: Testing a model of reciprocal causation. *Chemistry Education Research and Practice*, 17(4), 973–984. <https://doi.org/10.1039/c6rp00119j>
- Vishnumolakala, V. R., Southam, D. C., Treagust, D. F., Mocerino, M., & Qureshi, S. (2017). Students' attitudes, self-efficacy and experiences in a modified process-oriented guided inquiry learning undergraduate chemistry classroom. *Chemistry Education Research and Practice*, 18(2), 340–352. <https://doi.org/10.1039/c6rp00233a>
- Vogel, F. R., & Human-Vogel, S. (2016). Academic commitment and self-efficacy as predictors of academic achievement in additional materials science. *Higher Education Research and Development*, 35(6), 1298–1310. <https://doi.org/10.1080/07294360.2016.1144574>
- Wachsmuth, L. P., Runyon, C. R., Drake, J. M., & Dolan, E. L. (2017). Do biology students really hate math? Empirical insights into undergraduate life science majors' emotions about mathematics. *CBE Life Sciences Education*, 16(3), 1–10. <https://doi.org/10.1187/cbe.16-08-0248>
- Webber, D. M., & Flynn, A. B. (2018). How are students solving familiar and unfamiliar organic chemistry mechanism questions in a new curriculum? *Journal of Chemical Education*, 98(9), 1451–1467. <https://doi.org/10.1021/acs.jchemed.8b00158>
- Welsh, A. (2015). Investigating undergraduate students' metacognitive transformations in an introductory organic chemistry course. Retrieved from <https://circle.ubc.ca/handle/2429/52394>
- Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Journal of Emergency Primary Health Care (JEPHC)*, 8(3), 1–13. Retrieved from

- <http://ajp.paramedics.org/index.php/ajp/article/view/93%0Ahttps://ajp.paramedics.org/index.php/ajp/article/viewFile/93/90>
- Williams, T., & Williams, K. (2010). Self-efficacy and performance in mathematics: reciprocal determinism in 33 nations. *Journal of Educational Psychology*, 102(2), 453–466. <https://doi.org/10.1037/a0017271>
- Wilson, F., Kickul, J., & Marlino, D. (2007). Gender, entrepreneurial self-efficacy, and entrepreneurial career intentions: implications for entrepreneurship education. *Entrepreneurship Theory and Practice*, 31(3), 387–406. <https://doi.org/10.1111/j.1540-6520.2007.00179.x>
- Winkelmann, K., Baloga, M., Marcinkowski, T., Giannoulis, C., Anquandah, G., & Cohen, P. (2015). Improving students' inquiry skills and self-efficacy through research-inspired modules in the general chemistry laboratory. *Journal of Chemical Education*, 92(2), 247–255. <https://doi.org/10.1021/ed500218d>
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low- stakes assessment : problems and potential solutions. *Educational Assessment*, 10(1), 1–17. <https://doi.org/10.1207/s15326977ea1001>
- Witt-Rose, D. (2003). *Student self-efficacy in college science: An investigation of gender, age, and academic achievement*. University of Wisconsin-Stout. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:STUDENT+SELF-EFFICACY+IN+COLLEGE+SCIENCE+:+AN+INVESTIGATION+OF+GENDER+,+AGE+,+AND+ACADEMIC+ACHIEVEMENT#0>
- Xu, X., & Lewis, J. E. (2011). Refinement of a chemistry attitude measure for college students. *Journal of Chemical Education*, 88(5), 561–568. <https://doi.org/10.1021/ed900071q>
- Yong, A. G., & Pearce, S. (2016). A beginner's guide to factor analysis: focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79–94. <https://doi.org/10.20982/tqmp.09.2.p079>
- Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A.-R. (2015). Design and implementation content validity study: development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences*, 4(2), 165–178. <https://doi.org/10.15171/jcs.2015.017>
- Zeldin, A. L., Britner, S. L., & Pajares, F. (2008). A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers. *Journal of Research in Science Teaching*, 45(9), 1036–1058. <https://doi.org/10.1002/tea.20195>
- Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-Motivation for Academic Attainment: The Role of Self-Efficacy Beliefs and Personal Goal Setting. *American Educational Research Journal*, 29(3), 663–676. <https://doi.org/10.3102/00028312029003663>
- Zusho, A., Pintrich, P. R., & Coppola, B. (2003). Skill and will: The role of motivation and cognition in the learning of college chemistry. *International Journal of Science Education*, 25(9), 1081–1094. <https://doi.org/10.1080/0950069032000052207>

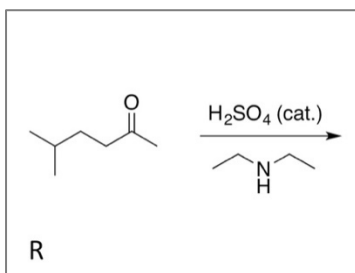
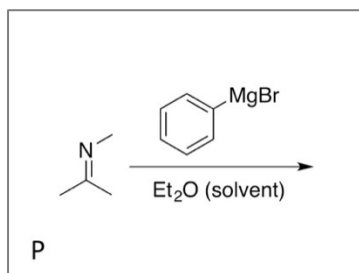
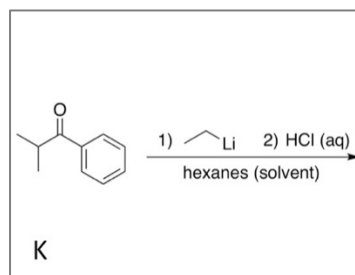
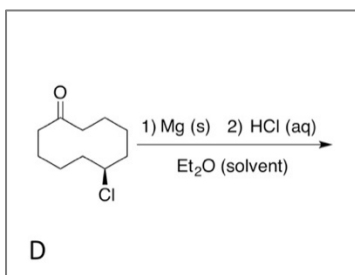
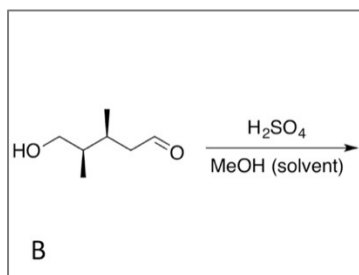
Appendices

Appendix A: Expert sort of the reaction cards based on the mechanistic patterns in the curriculum

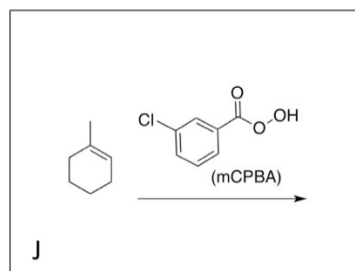
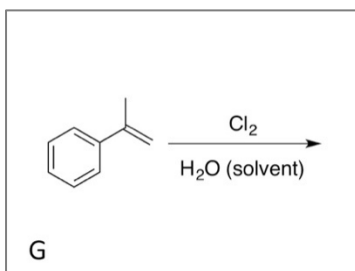
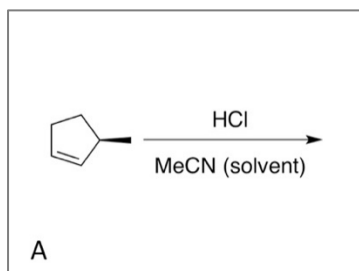
Category 1: Acid–base reactions



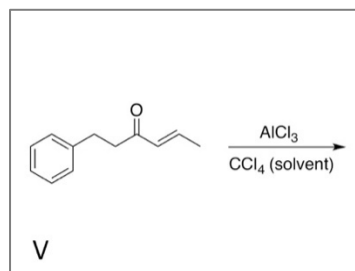
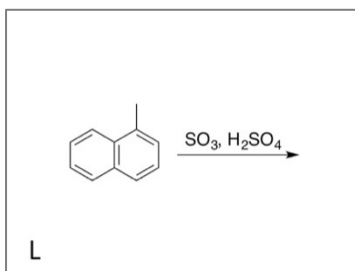
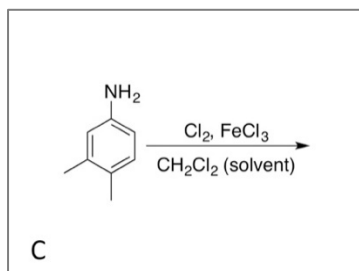
Category 2: π electrophiles + nucleophiles



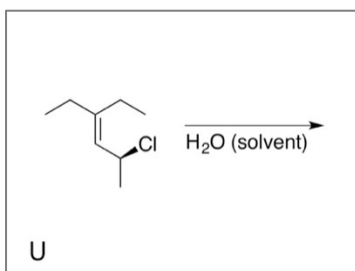
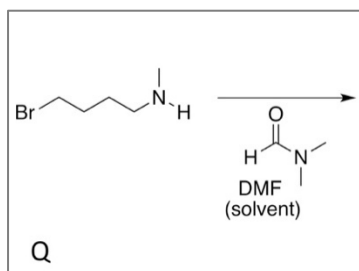
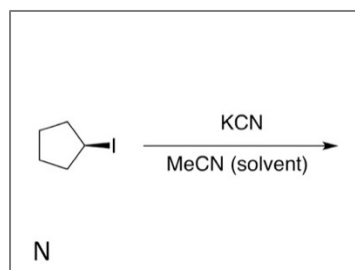
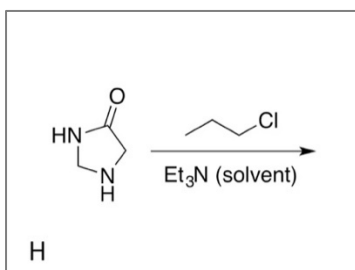
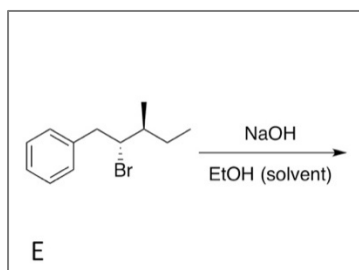
Category 3: π nucleophiles + electrophiles



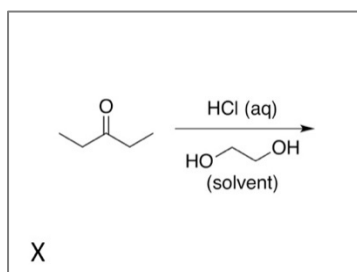
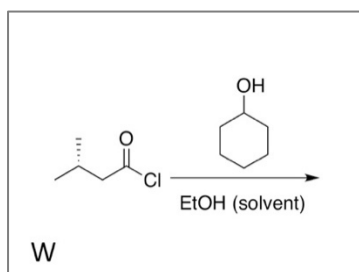
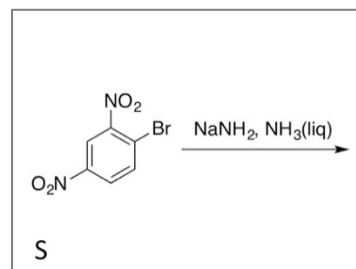
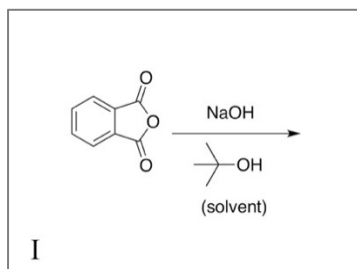
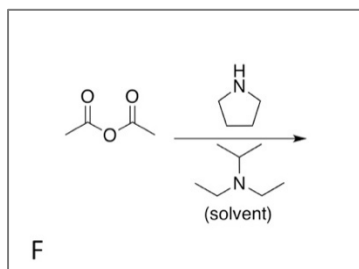
Category 4: Aromatic nucleophiles + electrophiles



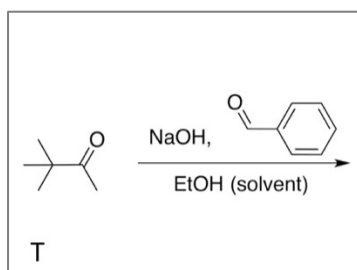
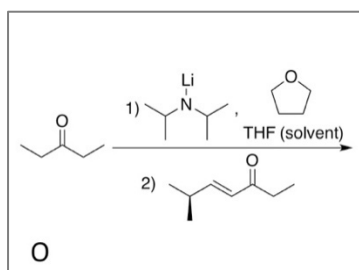
Category 5: σ electrophiles + nucleophiles/bases



Category 6: π electrophiles with a leaving group + nucleophiles



Category 7: Activated π nucleophiles + electrophiles

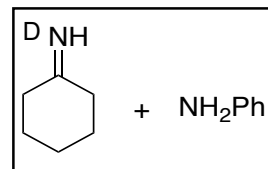
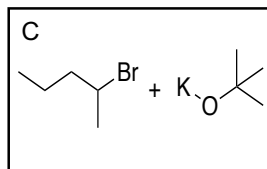
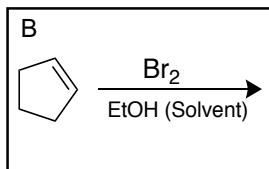
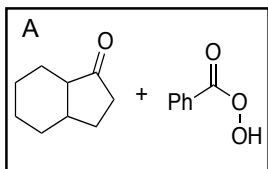


Appendix B: Response process validity interview protocol

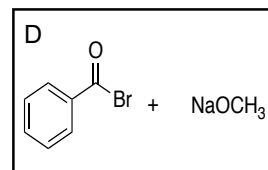
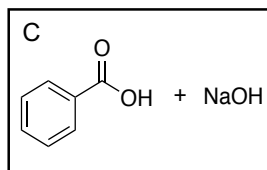
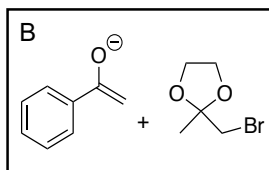
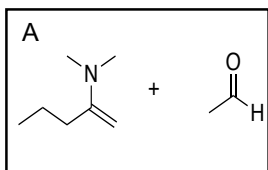
Response process validity interviews were retro-actively conducted using a semi-structured interview format ($N = 2$) containing three tasks following the original administration. The goal of this study was to identify any difficulties that may have arisen for participants who completed the sorts. The students who participated in the pilot study had completed the “Mechanistic Patterns and Principles” curriculum during the same time period as the full study but were enrolled within a different course section that did not have access to the online categorization task. This pilot study provided insight into how OCII students used the online card sort instrument and interpreted the instructional prompts and cues. The first two tasks asked participants to complete the open and closed sort by following the instructional prompts. After each task, participants were asked to comment on how the instrument impacted their ability and how the prompts or cues impacted their ability. Additionally, after completing the closed sort, participants were explicitly asked to interpret the closed sort categories, explicitly describing how participants understood the categories provided in Figure 2 of the manuscript. In a final task, participants were asked to complete a matching exercise to associate the symbolic representations of the mechanistic patterns with the complete mechanism.

Appendix C: High-stakes closed-sort like categorization task including on final examination (2017 and 2018)

Question 1. Categorize each of the following using the categories above (Figure 1)



Question 2. Circle the reaction(s) below that belong to category **G**



Appendix D: Alignment of the SEBOC items with the 3D-LAP criterion

In the protocol established by the 3D-LAP(Laverty et al., 2016), items are evaluated as to whether they meet 3-4 specific criterion per scientific practice, and 1-3 per crosscutting concept. This appendix includes the I) results of the analysis of the characterization of the nature of the items contained within the SEBOC instruments. II) An example of the 3D-LAP criterion used for the science practices and crosscutting concepts.

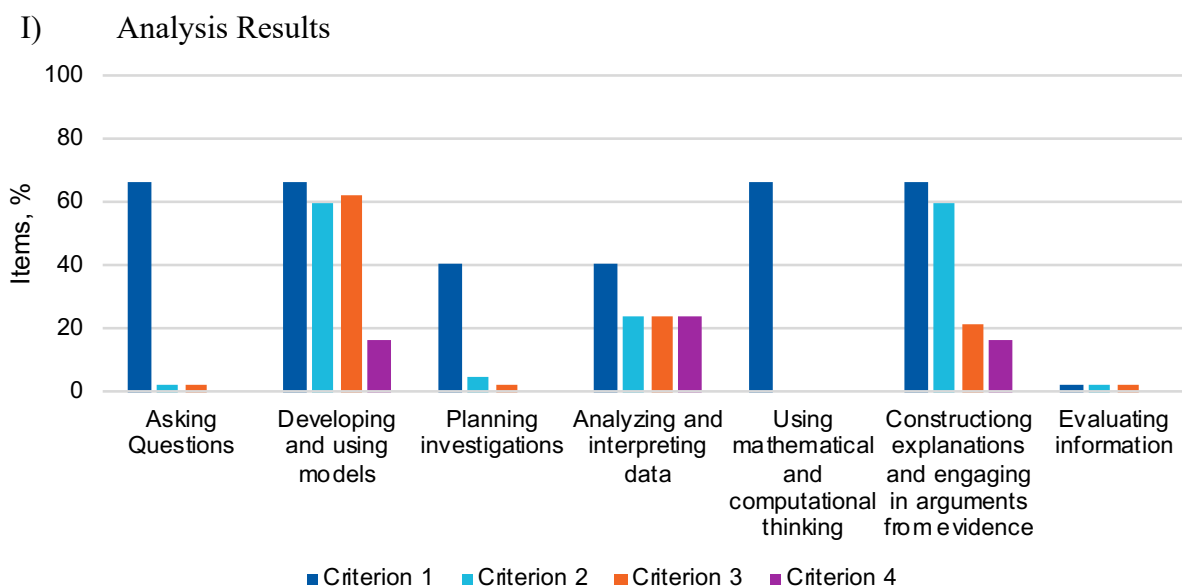


FIGURE 32. PERCENTAGE OF SELF-EFFICACY BELIEFS IN ORGANIC CHEMISTRY INSTRUMENT ITEMS EXPLICITLY ASSOCIATED WITH EACH SCIENCE PRACTICE CRITERION AS DEFINED BY THE 3D-LAP

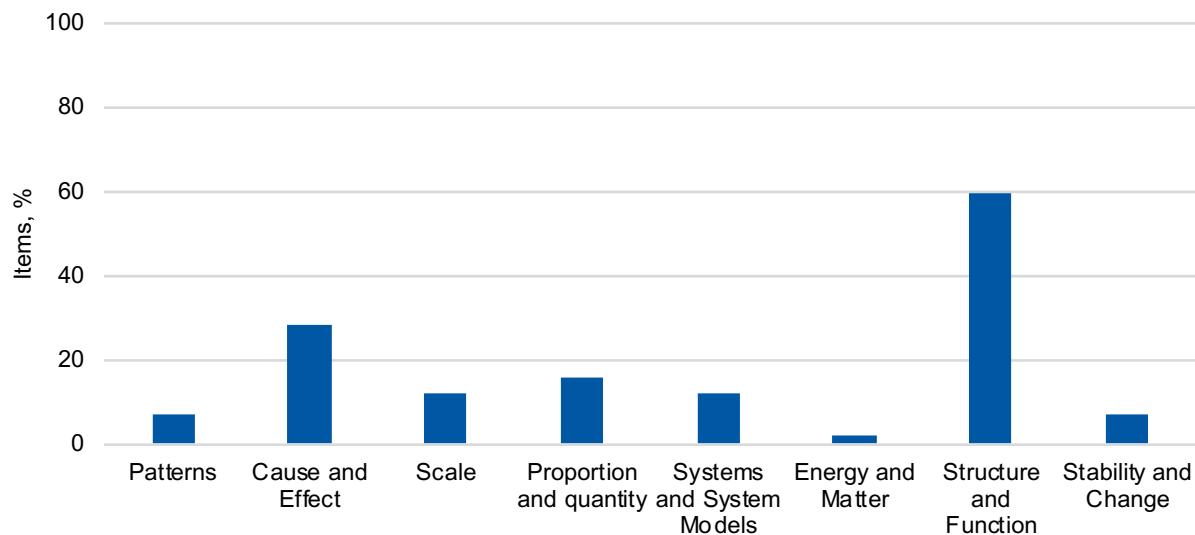


FIGURE 33. PERCENTAGE OF SELF-EFFICACY BELIEFS IN ORGANIC CHEMISTRY INSTRUMENT ITEMS EXPLICITLY ASSOCIATED WITH EACH CROSSCUTTING CONCEPT CRITERION AS DEFINED BY 3D-LAP

II) Example criterion

The following describes the four-criterion used for the science practice *Developing and using models*:

1. Question gives an event, observation, or phenomenon for the student to explain or make a prediction about.
2. Question gives a representation or asks student to construct a representation.
3. Question asks student to explain or make a prediction about the event, observation, or phenomenon.
4. Question asks student to provide the reasoning that links the representation to their explanation or prediction.

Appendix E: Self-Efficacy Beliefs in Organic Chemistry Instrument

The following describes the Self-Efficacy Beliefs in Organic Chemistry (SEBOC) Instrument. The dimensions included within the instrument and their respective items are as follows:

1. Self-efficacy towards cognitive skills (30 items: 1, 2, 4, 5, 6, 8, 9, 11, 13, 14, 15, 16, 18, 21, 23, 24, 26, 27, 28, 30, 31, 32, 33, 35, 36, 37, 38, 39, 40, 41,)
2. Self-efficacy towards everyday applications (6 items: 3, 10, 12, 20, 29, 42)
3. Self-efficacy towards experimental practices (6 items: 7, 17, 19, 22, 25, 34)

Confidence in Organic Chemistry

Directions: This survey contains 42 questions and will help us gain a better understanding of your confidence in organic chemistry. The following questionnaire asks you to rate your confidence, as of now, to perform different tasks relating to organic chemistry. There are no right or wrong answers. Please do not skip any items. Your answers are confidential.

How confident do you feel about your ability, as of now, to do the following:

* 1. Determine the product of a single-step reaction, given the starting materials and mechanistic arrows.

Not at all confident

Somewhat confident

Totally confident

☐☐☐☐☐☐☐

* 2. Draw the mechanistic arrows for a multi-step reaction, given the starting materials, intermediates, and products.

Not at all confident

Somewhat confident

Totally confident

☐☐☐☐☐☐☐

* 3. Describe how organic chemistry is involved in a global issue (e.g. societal, environment), either positively or negatively.

Not at all confident

Somewhat confident

Totally confident

☐☐☐☐☐☐☐

* 4. Form a 3-dimensional mental image of a molecule.

Not at all confident

Somewhat confident

Totally confident

☐☐☐☐☐☐☐

* 5. Assign the orbital hybridization (e.g., sp^2 , sp^3) of each atom in a molecule.

Not at all confident

Somewhat confident

Totally confident

☐☐☐☐☐☐☐

* 6. Predict trends in physical properties (e.g. boiling point, solubility) of an organic compound based on chemical properties (e.g. molecular size, polarity).

Not at all confident			Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐

* 7. Write a relevant research question to address a problem in a laboratory experiment.

Not at all confident			Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐

* 8. Identify the most acidic proton in a molecule.

Not at all confident			Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐

* 9. Identify the most basic atom in a molecule.

Not at all confident			Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐

* 10. Explain aspects of your everyday life using organic chemistry.

Not at all confident			Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐

* 11. Visualize how molecules are interacting in space as a reaction proceeds.

Not at all confident			Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐

* 12. Explain a concept learned in organic chemistry to a peer.

Not at all confident			Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐

* 13. Propose a retrosynthetic analysis for a target compound.

Not at all confident			Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐

* 14. Figure out the products of an unfamiliar reaction, given the starting materials.

Not at all
confident

Somewhat
confident

Totally confident

☐ ☐ ☐ ☐ ☐ ☐ ☐

* 15. Propose a mechanism and the products for a reaction, given the starting materials.

Not at all
confident

Somewhat
confident

Totally confident

☐ ☐ ☐ ☐ ☐ ☐ ☐

* 16. Propose a mechanism for a reaction, given the starting materials and products.

Not at all
confident

Somewhat
confident

Totally confident

☐ ☐ ☐ ☐ ☐ ☐ ☐

* 17. Connect steps of the experimental procedure to concepts of reactivity.

Not at all
confident

Somewhat
confident

Totally confident

☐ ☐ ☐ ☐ ☐ ☐ ☐

* 18. Draw the resonance structures of a molecule.

Not at all
confident

Somewhat
confident

Totally confident

☐ ☐ ☐ ☐ ☐ ☐ ☐

* 19. Record and tabulate raw data collected in a laboratory experiment.

Not at all
confident

Somewhat
confident

Totally confident

☐ ☐ ☐ ☐ ☐ ☐ ☐

* 20. Critique a news report or documentary that deals with some aspects of organic chemistry.

Not at all
confident

Somewhat
confident

Totally confident

☐ ☐ ☐ ☐ ☐ ☐ ☐

* 21. Compare two potential mechanisms for a reaction and justify which is the most likely to occur.

Not at all
confident

Somewhat
confident

Totally confident

☐ ☐ ☐ ☐ ☐ ☐ ☐

* 22. Interpret and draw conclusions from data collected in a lab experiment.

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

* 23. Identify nucleophilic or electrophilic sites in a molecule.

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

* 24. Compare two nucleophilic or electrophilic sites and justify which is most likely to participate in the reaction.

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

* 25. Design a laboratory procedure for a reaction learned in your organic chemistry course.

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

* 26. Describe how to manipulate an equilibrium to favour either products or reactants.

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

* 27. Describe what happens at the molecular level when the position of an equilibrium is changed.

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

* 28. Identify similarities between multiple reactions based on their underlying mechanisms.

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

* 29. Explain how a profession relates to organic chemistry, if at all.

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

* 30. Use chemical properties (e.g. electronegativity, atom size) to predict and justify the direction of an acid–base equilibrium.

Not at all confident				Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐	☐

* 31. Justify the outcome of a reaction using orbitals.

Not at all confident				Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐	☐

* 32. Determine the stereoisomeric relationship (e.g., enantiomers, diastereomers, the same molecule) between two structures, if any.

Not at all confident				Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐	☐

* 33. Interpret spectroscopic data (e.g., IR, NMR) to determine the structural information of a molecule.

Not at all confident				Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐	☐

* 34. Determine how experimental procedures and errors reflect data collected in a laboratory experiment.

Not at all confident				Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐	☐

* 35. Propose the stereochemical outcome for a reaction.

Not at all confident				Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐	☐

* 36. Propose the regiochemical outcome for a reaction.

Not at all confident				Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐	☐

* 37. Explain how a reaction coordinate diagram describes an associated reaction mechanism.

Not at all confident				Somewhat confident			Totally confident
☐	☐	☐	☐	☐	☐	☐	☐

* 38. Explain the relationship between the conformation and relative stability of a molecule.

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

* 39. Convert between various representations of a molecule (e.g. Lewis structure, line structure, Newman Projection).

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

* 40. Propose a synthesis for a target compound.

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

* 41. Explain how chemical properties (e.g., bond strength, electronegativity) influence spectroscopic data (e.g. IR, NMR).

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

* 42. Apply concepts learned in organic chemistry to other sciences.

Not at all
confident

Somewhat
confident

Totally confident

☐☐☐☐☐☐☐

Confidence in Organic Chemistry

43. Which of the following best describes your declared field of study?

- | | |
|---|--|
| ☐ Biochemistry | ☐ Health Science |
| ☐ Biomedical Science | ☐ Human Kinetics |
| ☐ Biology | ☐ Psychology |
| ☐ Biopharmaceutical Science | ☐ Undeclared |
| ☐ Chemistry | ☐ Prefer not to answer |
| ☐ Chemical Engineering | |
| ☐ Other (please specify) | |

44. What is your year of study?

- | | |
|--|-------------------------|
| ☐ 1 | ☐ 4 |
| ☐ 2 | ☐ 5 |
| ☐ 3 | ☐ 6 |
| ☐ Other (please specify) | |

45. My first language is

- ☐ French
- ☐ English
- ☐ Prefer not to answer
- ☐ Other (please specify)

46. What is your gender?

- ☐ Female
- ☐ Male
- ☐ Prefer not to answer
- ☐ These options do not apply to me. I identify as

47. What is your age?

- | | |
|--------------------------------|--|
| ☐ Under 18 | ☐ 24-25 |
| ☐ 18-20 | ☐ Over 25 |
| ☐ 21-23 | ☐ Prefer not to answer |

48. Which of the following groups do you identify with? (select all that apply)

- | | |
|---|--|
| ☐ Indigenous | ☐ Korean |
| ☐ Arab | ☐ Latin American |
| ☐ Black | ☐ South Asian (e.g., East Indian, Pakistani, Sri Lankan, etc.) |
| ☐ Filipino | ☐ Southeast Asian (e.g., Vietnamese, Cambodian, Malaysian, Laotian, etc.) |
| ☐ Caucasian | ☐ West Asian (e.g., Iranian, Afghan, etc.) |
| ☐ Chinese | ☐ Prefer not to answer |
| ☐ Japanese | |
| ☐ Other (please specify) | |