
Brandon J. Yik
University of South Florida

Follow this and additional works at: https://digitalcommons.usf.edu/etd

Part of the Chemistry Commons, and the Science and Mathematics Education Commons

Scholar Commons Citation

This Dissertation is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact digitalcommons@usf.edu.

by

Brandon J. Yik

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy
Department of Chemistry
College of Arts and Sciences
University of South Florida

Major Professor: Jeffrey R. Raker, Ph.D.
James W. Leahy, Ph.D.
Scott E. Lewis, Ph.D.
Eun Sook Kim, Ph.D.

Date of Approval: October 4, 2022

Keywords: Lewis acid–base, Machine learning, Nucleophiles, Multilevel modeling, Faculty beliefs

Copyright © 2022, Brandon J. Yik
Dedication

“It doesn’t matter what you are, it only matters what you do. It’s your choice.”

– Sam Winchester, Supernatural

This work is dedicated to those that helped me grow to become the person I am today and those that helped me along this journey. I would not be here without the immeasurable support from family and friends. This dissertation is especially dedicated to all of my former instructors: individuals that not only educate, but care, inspire, mentor, support, listen, adapt, sacrifice, and so much more. Thank you for making the choice to go into this profession and doing great things. Works herein are steppingstones to better understand how to help you best help your students. “We got work to do.”
Acknowledgments

There are many people I would like to thank and acknowledge. Without their help and support, none of this work would be possible. I would like to especially thank my family for their unconditional love and support throughout my life and education: Von, Ying, Steven, and Suzy. None of this would have been possible without you.

Thank you to the friends I have made during my time in Florida for their friendship, fun, laughs, and support: Amber, John, Andy, Mike, Andrew, Bartosz, Jamie, Ted, AJ, Ayesha, Guizella, and Ying. Thank you to my lifelong friends for your continued support; in particular, thank you: Sushmi, Swati, Ehab, Sunny, Thomas, and Tosin. Thank you to all the pets for their emotional support and joy: Aveline, Isabella, Mango, Mittenz, Mochi, and Ruby.

Thank you to my previous advisors and mentors who stimulated my interest in chemistry and science and who supported my endeavors. Thank you to my previous research advisors: Dr. Theodore Goodson III and Dr. Henry La Pierre. Thank you to the graduate students and postdoctoral fellows who helped mentor me: Dr. Joseph Furgal, Dr. Sung Hei Yau, and Dr. Phi Doan. I want to extend a special thank you to the M-STEM Academies for affording opportunities that led me into this journey; specifically: Dr. Deborah Goldberg, Dr. Lynn Carpenter, and Shannon Zuniga.

Thank you to my previous instructional mentors and colleagues who encouraged me to pursue my interest in teaching and education. Thank you for providing opportunities to teach with you, talk with you, and learn from you: Dr. Kathleen Nolta, Dr. Amanda Stephens, Dr. Christy O’Mahony, Dr. Kimberly Schurmeier, Dr. Hui Zhu, Dr. Carrie Shepler, Dr. Michael Evans, and in particular, Dr. John Tillotson.
Thank you to all my chemistry education research colleagues, past and present, for an amazing and priceless time at USF. Thank you for all the friendship, advice, and feedback. Thank you to the Raker group members: Dr. Justin Pratt, Dr. Megan Connor, Dr. Aaron Clark, and Stephanie Frost. In particular, I would like to extend a special thank you to Dr. Amber Dood for laying the groundwork for part of my dissertation work, for guidance in machine learning, for your advice, and for being a good friend. Thank you, all, for being instrumental parts of my growth.

Thank you to all the students and instructors who participated in the many surveys required to collect data for work in this dissertation. Thank you to the instructors that allowed me collect data in their courses: Dr. Daniel Cruz-Ramírez de Arellano, Dr. Kimberly Fields, and Dr. Frankie Costanza. An extended thank you to Dr. Daniel Cruz-Ramírez de Arellano and Dr. Jeffrey Raker for allowing me to a teaching assistant in their courses; I have learned a lot about the teaching of organic chemistry from working with them.

Thank you to my co-authors for their continued support throughout my graduate career: Dr. Amber Dood, Dr. Daniel Cruz-Ramírez de Arellano, Dr. Kimberly Fields, Stephanie Frost, Dr. Naneh Apkarian, Dr. Marilyne Stains, Dr. Charles Henderson, Dr. Melissa Dancy, and Dr. Estrella Johnson. This dissertation would not be complete without your contributions.

Thank you to the academic, administrative, and business staff in the Department of Chemistry for making my life as a graduate student a little bit easier: Kaitlyn Kroner, Paige Romero, Karissa Dailey, Ryan Jahn, and Joshue Martinez. Your work is always appreciated.

Thank you to those who helped fund my graduate education. Thank you to the National Science Foundation for support through a research assistantship. Thank you to the Office of Graduate Studies for support through a Dissertation Completion Fellowship. Thank you to the Department of Chemistry for support through a teaching assistantship, the Theodore and Venette Askounes Ashford Doctoral Fellowship in Chemistry, and a Martin Travel Endowment award. Thank you to the Askounes Ashford and Martin families for the endowments that established these fellowships and awards, and for your commitments to graduate education.
Thank you to my committee who supported me throughout my graduate career. Dr. Scott Lewis, thank you for always bringing a new perspective to my work and challenging me to think more critically about methods, claims, and limitations of my work. Dr. Eun Sook Kim, thank you for being a patient educator and providing helpful guidance in multilevel modeling. Dr. James Leahy, thank you for being interested in my work and pushing the bounds so that I communicate more effectively with other scientists and administrators.

Finally, I would like to say thank you to major professor, Dr. Jeffrey Raker. Thank you for our continuous advice and encouragement. Thank you for always being available and there for me when I need it. Thank you for being an amazing advisor; you taught me how to be a scholar, but also showed me how to be a better person. Thank you for allowing me to explore my interests and passions. Thank you for always encouraging self-care and work-life balance, even if neither of us always exemplified did the latter… I will always be grateful to have had the opportunity to work with you; we could work on anything and have fun doing so. Thank you for teaching me the power of punctuation: this is something I will carry on in my own writing! Jeff, I will sincerely miss working with you. Thank you for being an integral part of my journey.
# Table of Contents

List of Tables..............................................................................................................................................vii

List of Figures................................................................................................................................................ix

List of Equations .............................................................................................................................................x

Abstract ..........................................................................................................................................................xi

Chapter 1 Introduction .................................................................................................................................1
  1.1 Students’ understanding of reaction mechanisms ..................................................................................1
    1.1.1 Overview of the work ................................................................................................................2
  1.2 Evaluating the uptake of research-based instructional strategies .....................................................6
    1.2.1 Overview of the work ................................................................................................................7
  1.3 Summary of the work ..........................................................................................................................11
  1.4 References ...........................................................................................................................................12

Chapter 2 Methods ....................................................................................................................................25
  2.1 Interrater reliability (IRR) ..................................................................................................................25
    2.1.1 Percent agreement ......................................................................................................................26
    2.1.2 Cohen’s kappa ............................................................................................................................26
    2.1.3 IRR in this work ..........................................................................................................................27
  2.2 R program ...........................................................................................................................................28
  2.3 Machine learning performance metrics ..............................................................................................31
    2.3.1 Confusion matrix ........................................................................................................................31
    2.3.2 Accuracy ......................................................................................................................................32
    2.3.3 $F_1$ score ....................................................................................................................................33
    2.3.4 Matthews correlation coefficient ...............................................................................................34
  2.4 Multilevel modeling .............................................................................................................................34
    2.4.1 Multilevel linear regression ........................................................................................................35
    2.4.2 Multilevel logistic regression .....................................................................................................36
  2.5 References ...........................................................................................................................................39

Chapter 3 Development of a machine learning-based tool to evaluate correct Lewis acid–base model use
in written responses to open-ended formative assessment items..................................................................44
  3.1 Note to Reader ......................................................................................................................................44
  3.2 Abstract ...............................................................................................................................................44
  3.3 Introduction .........................................................................................................................................45
    3.3.1 Student understanding of acid–base models ..............................................................................48
    3.3.2 Role of formative assessments in developing acid–base model use by students .....................51
    3.3.3 Argument for a generalized predictive model ..........................................................................52
    3.3.4 Research question ......................................................................................................................53
Chapter 4

3.4 Methods ............................................................................................................... 53
  3.4.1 Constructed response items........................................................................... 53
  3.4.2 Data collection ............................................................................................... 54
  3.4.3 Development of classification scheme ......................................................... 55
  3.4.4 Development of a machine learning model .................................................... 57

3.5 Results and discussion ......................................................................................... 63
  3.5.1 Performance metrics ..................................................................................... 64
  3.5.2 Evaluation of machine learning algorithm ....................................................... 66

3.6 Limitations ............................................................................................................. 71
  3.6.1 Sample homogeneity ..................................................................................... 71
  3.6.2 Predictive model accuracy .............................................................................. 72
  3.6.3 Machine learning methodology ....................................................................... 72
  3.6.4 Model shortcomings ....................................................................................... 73
    3.6.4.1 False positives .......................................................................................... 73
    3.6.4.2 False negatives ....................................................................................... 74
    3.6.4.3 Mixed acid–base model use ..................................................................... 75
    3.6.4.4 Gaming the system ............................................................................... 76

3.7 Implications ........................................................................................................... 77
  3.7.1 Implications for instructors .......................................................................... 77
  3.7.2 Implications for researchers ......................................................................... 79

3.8 Conclusion ............................................................................................................ 81

3.9 Conflicts of interest .............................................................................................. 81

3.10 Acknowledgments ............................................................................................... 81

3.11 References .......................................................................................................... 82

Chapter 4 Generalized rubric for level of explanation sophistication for nucleophiles in organic chemistry reaction mechanisms ................................................................................. 98
  4.1 Note to Reader ..................................................................................................... 98
  4.2 Abstract ............................................................................................................... 99

4.3 Introduction ......................................................................................................... 99
  4.3.1 Understanding of organic chemistry reaction mechanisms ......................... 102
  4.3.2 Understanding of nucleophiles ...................................................................... 104
  4.3.3 Moving toward meaningful assessments ......................................................... 105
  4.3.4 Rubric for evaluating understanding of nucleophiles ...................................... 108
  4.3.5 Research goal .............................................................................................. 112

4.4 Methods .............................................................................................................. 113
  4.4.1 Constructed response items ........................................................................... 113
  4.4.2 Data collection ............................................................................................... 113

4.5 Results and discussion ......................................................................................... 116
  4.5.1 Refining and operationalizing the rubric levels .............................................. 116
  4.5.2 Rubric use by Organic Chemistry 1 and Organic Chemistry 2 courses ........ 118
  4.5.3 Rubric use by nucleophile type ................................................................. 119
  4.5.4 Rubric use by reaction family ....................................................................... 120
  4.5.5 Rubric use by prompt variations ................................................................. 120

4.6 Case study – Using the rubric in teaching practice ............................................. 123
  4.6.1 Reflective teaching practices ................................................................. 127
  4.6.2 Implications for teaching practice .............................................................. 129

4.7 Limitations .......................................................................................................... 130
  4.7.1 Instructor implementation ................................................................. 130
  4.7.2 Sample homogeneity .............................................................................. 132
Chapter 5 Evaluating the impact of malleable factors on percent time lecturing in
gateway chemistry, mathematics, and physics courses .......................................................... 150
5.1 Note to Reader .................................................................................................................. 150
5.2 Abstract .......................................................................................................................... 150
5.3 Introduction ..................................................................................................................... 151
  5.3.1 Conceptual framework ............................................................................................... 154
    5.3.1.1 Contextual factors ............................................................................................... 157
      5.3.1.1.1 Department characteristics ........................................................................ 157
      5.3.1.1.2 Department appointment expectations ...................................................... 157
      5.3.1.1.3 Classroom contextual factors ....................................................................... 159
    5.3.1.2 Personal factors .................................................................................................. 161
      5.3.1.3 Teacher thinking factors .................................................................................. 163
      5.3.1.4 Framework conceptualization .......................................................................... 165
  5.3.2 Research question ....................................................................................................... 166
5.4 Methods .......................................................................................................................... 166
  5.4.1 Survey development ................................................................................................. 166
  5.4.2 Participants ............................................................................................................... 167
  5.4.3 Data collection .......................................................................................................... 168
  5.4.4 Multilevel models ..................................................................................................... 168
  5.4.5 Explanation of variables ......................................................................................... 170
5.5 Results ............................................................................................................................. 171
  5.5.1 Department characteristics factors ........................................................................ 174
  5.5.2 Department appointment expectations .................................................................... 175
  5.5.3 Classroom contextual factors ................................................................................ 176
  5.5.4 Personal factors ...................................................................................................... 177
  5.5.5 Teacher thinking factors ....................................................................................... 177
5.6 Discussion ......................................................................................................................... 178
  5.6.1 Classroom spaces .................................................................................................... 178
  5.6.2 Shared decision on instructional methods ............................................................... 180
  5.6.3 Experience as a student in a course using research-based
      instructional strategies ................................................................................................. 181
  5.6.4 Participation in teaching-related professional development .................................. 182
  5.6.5 Scholarship of teaching and learning or discipline-based education
      research ......................................................................................................................... 184
  5.6.6 Holding a growth mindset ...................................................................................... 185
  5.6.7 Non-significant factors .......................................................................................... 186
  5.6.8 Limitations ............................................................................................................. 187
5.7 Conclusions ...................................................................................................................... 188
5.8 Abbreviations .................................................................................................................. 189
5.9 Acknowledgments .......................................................................................................... 190
5.10 Authors’ contributions .................................................................................................. 190
5.11 Funding .......................................................................................................................... 190
5.12 Competing interests ...................................................................................................... 190
5.13 References................................................................................................................. 190
Chapter 6 Association of malleable factors with adoption of research-based instructional
strategies in introductory chemistry, mathematics, and physics................................. 212
  6.1 Note to Reader........................................................................................................... 212
  6.2 Abstract...................................................................................................................... 212
  6.3 Introduction .............................................................................................................. 213
  6.4 Conceptual frameworks........................................................................................... 215
  6.5 Methods .................................................................................................................... 219
    6.5.1 Respondents ....................................................................................................... 219
    6.5.2 Data collection .................................................................................................. 220
    6.5.3 Research-based instructional strategies adoption scale .................................. 221
    6.5.4 Multilevel models ............................................................................................. 223
  6.6 Results ....................................................................................................................... 224
    6.6.1 Multilevel models ............................................................................................ 224
    6.6.2 Association of stages of RBIS adoption with percent lecturing .................... 227
  6.7 Discussion ................................................................................................................. 230
    6.7.1 Importance of classroom spaces ....................................................................... 230
    6.7.2 Perceived value of assessment of teaching performance .............................. 231
    6.7.3 Experience as a student in a course using research-based
        instructional strategies ......................................................................................... 233
    6.7.4 Engagement in teaching-related professional development ......................... 234
    6.7.5 Participation in the scholarship of teaching and learning ............................. 236
    6.7.6 Holding a growth mindset .............................................................................. 237
    6.7.7 (Dis)satisfaction with student learning ....................................................... 239
    6.7.8 Association of malleable factors with RBIS tryout, adoption, and
        percent lecturing ................................................................................................. 240
  6.8 Limitations ............................................................................................................... 240
  6.9 Conclusions .............................................................................................................. 241
  6.10 Conflict of interest ................................................................................................. 243
  6.11 Author contributions ............................................................................................. 243
  6.12 Funding .................................................................................................................... 243
  6.13 Acknowledgments .................................................................................................. 243
  6.14 References............................................................................................................. 243

Chapter 7 Conclusion ....................................................................................................... 272
  7.1 Student explanations of reaction mechanisms ....................................................... 273
    7.1.1 Summary of results ........................................................................................ 273
    7.1.2 Implications for instruction ............................................................................. 274
    7.1.3 Implications for research ................................................................................ 276
  7.2 Uptake of research-based instructional strategies ............................................... 279
    7.2.1 Summary of results ........................................................................................ 279
    7.2.2 Implications for research .............................................................................. 279
    7.2.3 Implications for policy .................................................................................. 281
  7.3 Overall implications ............................................................................................... 282
  7.4 Summary ................................................................................................................. 284
  7.5 References.............................................................................................................. 285

Appendix A Supporting Information for Chapter 3 ......................................................... 297
  A.1 Constructed response items used in the training, cross-validation, and split
      validation sets ....................................................................................................... 297
  A.2 Constructed response items used in the external validation set .......................... 300
**List of Tables**

Table 2.1  Example calculation of percent agreement .......................................................... 26  
Table 3.1  Representative examples of correct use of the Lewis acid–base model ............... 56  
Table 3.2  Distribution of human-classified correct and incorrect use/non-use of the Lewis acid–base model .............................................................. 58  
Table 3.3  Predictive model results on the cross-validation set ........................................... 67  
Table 3.4  Predictive model results on the stratified split-validation set ............................... 68  
Table 3.5  Predictive model results on the remaining split-validation set ............................. 69  
Table 3.6  Predictive model results on the external validation set ....................................... 71  
Table 3.7  Distribution of human-classified correct and incorrect use/non-use of the Lewis acid–base model for the external validation set ......................... 71  
Table 4.1  Rubric for levels of explanation sophistication for nucleophiles ....................... 111  
Table 4.2  Reaction mechanisms used ............................................................................. 114  
Table 4.3  Constructed response item prompt variations .................................................. 115  
Table 4.4  Distribution of levels of explanation sophistication for nucleophiles ............... 118  
Table 4.5  Example excerpts from responses for the levels of explanation sophistication for nucleophiles ................................................................. 121  
Table 4.6  Distribution of levels of explanation sophistication by nucleophile type .......... 121  
Table 4.7  Distribution of levels of explanation sophistication by reaction family ............ 122  
Table 4.8  Distribution of levels of explanation sophistication by reactions involving proton transfers ........................................................................................................ 122  
Table 4.9  Distribution of levels of explanation sophistication by prompt type .................. 123  
Table 4.10  Distribution of levels of explanation sophistication for Case Study Time  
Point 1: Reactions of alkenes ......................................................................................... 125
Table 4.11 Distribution of levels of explanation sophistication for Case Study Time Point 2: Substitution reactions ................................................................. 125

Table 4.12 Distribution of levels of explanation sophistication for Case Study Time Point 2: Reactions with alkenes and alkynes ........................................ 125

Table 4.13 Distribution of levels of explanation sophistication for Case Study Time Point 3: Substitution reactions with alcohols ........................................ 127

Table 4.14 Distribution of levels of explanation sophistication for Case Study Time Point 3: Reactions of aldehydes and ketones to form alcohols ............. 127

Table 5.1 Factors used in the final multilevel model and their coding ......................... 172

Table 5.2 Factors in explaining percent lecturing in introductory chemistry, mathematics, and physics courses .............................................................. 174

Table 6.1 Malleable factors included within this study situated within the TCSR framework and hypotheses about how factors impact the adoption of RBIS with relevant citations ..................................................... 217

Table 6.2 RBIS Adoption Scale with associated CACAO stage of adoption .................. 222

Table 6.3 Distribution of stages of adoption ................................................................... 225

Table 6.4 Factors associated with RBIS Tryout and Adoption ...................................... 228

Table 6.5 Association of RBIS stages of adoption with percent lecturing .................... 229
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Example SVM plot for two predictors</td>
<td>30</td>
</tr>
<tr>
<td>2.2</td>
<td>Example SVM plot for three predictors</td>
<td>30</td>
</tr>
<tr>
<td>2.3</td>
<td>Confusion matrix</td>
<td>32</td>
</tr>
<tr>
<td>3.1</td>
<td>Example of constructed-response items</td>
<td>54</td>
</tr>
<tr>
<td>3.2</td>
<td>Process of model development and evaluation</td>
<td>59</td>
</tr>
<tr>
<td>3.3</td>
<td>Confusion matrix showing the outcomes of predicted vs. actual classifications</td>
<td>64</td>
</tr>
<tr>
<td>4.1</td>
<td>Example reaction mechanisms</td>
<td>115</td>
</tr>
<tr>
<td>4.2</td>
<td>Timeline of the case study: course content, examinations, and assessment time points</td>
<td>124</td>
</tr>
<tr>
<td>5.1</td>
<td>Conceptualization of the Teacher-Centered Systemic Reform (TCSR) model for change in higher education with control variables (discipline and highest degree awarded) and malleable factors included in this study</td>
<td>156</td>
</tr>
<tr>
<td>5.2</td>
<td>Comparative association of variables with percent time lecturing: (A) Department characteristics, personal factors, and teacher thinking. (B) Classroom contextual factors and department appointment expectations</td>
<td>175</td>
</tr>
<tr>
<td>6.1</td>
<td>Association of instructor-level malleable factors between (A) RBIS Tryout and Percent Lecturing Models and (B) RBIS Adoption and Percent Lecturing Models</td>
<td>232</td>
</tr>
</tbody>
</table>
## List of Equations

<table>
<thead>
<tr>
<th>Equation</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equation 2.1</td>
<td>Accuracy</td>
<td>32</td>
</tr>
<tr>
<td>Equation 2.2</td>
<td>Precision</td>
<td>33</td>
</tr>
<tr>
<td>Equation 2.3</td>
<td>Recall</td>
<td>33</td>
</tr>
<tr>
<td>Equation 2.4</td>
<td>( F_1 ) score</td>
<td>33</td>
</tr>
<tr>
<td>Equation 2.5</td>
<td>Matthews correlation coefficient</td>
<td>34</td>
</tr>
<tr>
<td>Equation 2.6</td>
<td>Multilevel model for percent lecturing</td>
<td>36</td>
</tr>
<tr>
<td>Equation 2.7</td>
<td>Multilevel model for stages of RBIS adoption</td>
<td>37</td>
</tr>
<tr>
<td>Equation 3.1</td>
<td>Accuracy</td>
<td>65</td>
</tr>
<tr>
<td>Equation 3.2</td>
<td>( F_1 ) score</td>
<td>65</td>
</tr>
<tr>
<td>Equation 3.3</td>
<td>Precision</td>
<td>65</td>
</tr>
<tr>
<td>Equation 3.4</td>
<td>Recall</td>
<td>65</td>
</tr>
<tr>
<td>Equation 3.5</td>
<td>Matthews correlation coefficient</td>
<td>66</td>
</tr>
<tr>
<td>Equation 5.1</td>
<td>Multilevel model for percent lecturing</td>
<td>170</td>
</tr>
</tbody>
</table>
Abstract

The work in this dissertation is presented in two parts. The first part (Chapters 3 and 4) details work relating to students' explanations of reaction mechanisms in organic chemistry. The second part (Chapters 5 and 6) details work relating to the evaluating the uptake of research-based instructional practices in introductory chemistry, mathematics, and physics courses.

To evaluate learning of organic chemistry reactions, instructors must ask students to construct written explanations of reaction mechanisms. Written assessments should focus on what is happening and why it is happening to promote deeper student understanding. However, for instructors to gain insight into students' understanding, the time and effort to evaluate the explanations is prohibitive. Further, such evaluation is often not standardized or grounded in the research literature. Lexical analysis and machine learning algorithms can be used to score students' written responses to open-ended constructed response items to alleviate grading burdens. Specifically, lexical analysis and machine learning techniques can be used produce predictive models that aid with evaluating the quality of students' explanations. Rubrics are an additional tool that can help students and instructors formatively assess written explanations. In this work, a generalized predictive model using machine learning techniques was developed to evaluate students' understanding of acid–base reactions and properties of acids and bases through the correct use of the Lewis acid–base model (Chapter 3). Research on students' understanding of nucleophiles grounded the development of a generalized rubric to evaluate students' level of written explanation sophistication for nucleophiles (Chapter 4).

Evaluating instructional practices in chemistry, mathematics, and physics, necessitates multidisciplinary, large-scale studies of the factors that influence teaching pedagogies. Data from a multi-institution, large-scale survey of postsecondary introductory-level chemistry, mathematics,
and physics instructors in the United States was used to model two outcomes: percent lecturing (Chapter 5) and stages of research-based instructional strategy adoption (Chapter 6). These outcomes were modeled using multilevel modeling with contextual, personal, and teacher thinking factors as the explanatory variables. Multilevel models produced explanatory models of instructional practices, providing change agents with productive avenues for improving instructors’ pedagogies.
Chapter 1

Introduction

Discipline-based education research (DBER) is “an empirical approach to investigating teaching and learning that is informed by an expert understanding of [science, technology, engineering, and mathematics (STEM)] disciplinary knowledge and practice” (NRC, 2012, p. 9). According to the National Academies’ DBER report (NRC, 2012, p. 9), some of the long-term goals of DBER are to “understand the nature and development of expertise in a discipline,” “help to identify and measure … instructional approaches that advance students,” and “contribute to the knowledge base in a way that can guide the transformation of DBER findings to classroom practice.” The work reported in this dissertation aims to advance these goals by investigating the students’ learning (Chapters 3 and 4) and instructors’ teaching (Chapter 5 and 6). These investigations are encompassed within two research themes: (i) students’ understanding of organic chemistry reaction mechanisms and (ii) evaluating the uptake of research-based instructional strategies in introductory chemistry, mathematics, and physics.

1.1 Students’ understanding of reaction mechanisms

The electron pushing formalism was first developed by Kermack and Robinson (1922) to describe the movement of electrons throughout the process of a reaction. Chemists use the electron pushing formalism along with reaction mechanisms to describe, explain, and predict chemical reactivity (Goodwin, 2003, 2008, 2012), and instructors use reaction mechanisms to connect course content and explain reactivity behind organic chemistry reactions (Morrison and Boyd, 1959; Wheeler and Wheeler, 1982; Bhattacharyya, 2013). The importance of reaction mechanisms is noted by the inclusion of mechanisms in undergraduate organic chemistry
textbooks (e.g., Solomons et al., 2016; Klein, 2021), and instructors continually regard reaction mechanisms as a foundational topic in organic chemistry (Bhattacharyya, 2013; Bhattacharyya and Harris, 2018; Nedungadi and Brown, 2021).

The research literature has demonstrated that students struggle with understanding the electron pushing formalism and reaction mechanisms (e.g., Bhattacharyya and Bodner, 2005; Kraft et al., 2010; Strickland et al., 2010; Bhattacharyya, 2014; Flynn and Featherstone, 2017; Caspari et al., 2018a, 2018b; Bodé et al., 2019). Students lack deep understanding of reaction mechanisms and adorn the mechanistic pictures with arrows (Grove et al., 2012), use rote memorization and surface-level approaches to construct mechanisms (Ferguson and Bodner, 2008; Cooper and Stowe, 2018). Students’ understanding of reaction mechanisms do not meet instructors’ expectations.

Written assessments that use purposeful constructed-response prompts can promote understanding of chemical concepts (Birenbaum and Tatsuoka, 1987; Cooper, 2015; Stowe and Cooper, 2017; Underwood et al., 2018). Resultingly, many researchers have called for the need to ask students to explain why (e.g., Caspari et al., 2018a; Bodé et al., 2019; Crandell et al., 2019, 2020; Dood et al., 2020). Open-ended written assessments should prompt students what is happening in reaction mechanism and why it happens to promote deeper student learning and also allow instructors to obtain deeper insight into their students’ understanding (Bell and Cowie, 2001; Cooper, 2015; Cooper et al., 2016; Underwood et al., 2018).

1.1.1 Overview of the work

The work reported in Chapters 3 and 4 originates from a study asking general chemistry students to classify a reaction, i.e., an aqueous acid–base proton transfer, and to describe what is happening and why it is happening (Cooper et al., 2016). Students described this particular reaction (between hydrochloric acid and water), using the Brønsted–Lowry and Lewis acid–base models, and students’ explanations ranged from descriptive (i.e., the what), to mechanistic (i.e.,
the *what* and *how*), and causal (i.e., the *what* and *why*). In a replication study, in a handwritten homework assignment, students were also asked to describe what is happening and why it is happening for the same constructed response item (i.e., the reaction between hydrochloric acid and water) by other researchers (Dood et al., 2018) at the University of South Florida (i.e., the study site for Chapters 3 and 4). Whereas Cooper et al. (2016) categorized students’ classifications of the reaction as mutually-exclusive acid–base model use, Dood et al. (2018) categorized students’ classifications more broadly including mixed-model use in addition to mutually-exclusive Arrhenius, Brønsted–Lowry, and Lewis acid–base model use. Additionally, Dood et al. (2018) reported an association between students’ use of the Lewis acid–base model in their explanations and students’ score on an exam that included acid–based related items and on students’ subscore on the acid–base related items.

The time to score handwritten responses in a homework assignment, especially in the context of instruction, is impractical. Previous work (e.g., Haudek et al., 2012; Prevost et al., 2016) demonstrates that lexical analysis is a viable option to score student responses. To circumvent problems with handwritten assignments (e.g., illegible handwriting), Dood et al. (2018) collected electronically written responses through online surveys for the development of an automated text analysis predictive model using SPSS Modeler (IBM Corp., 2017). A binomial logistic regression predictive model was developed that assigned a binary category to a response (i.e., 1 for Lewis acid–base model use and 0 for Lewis acid–base model nonuse) and resulted in 87% accuracy for the training data (i.e., data used to develop the model; $n = 534$) and 82% accuracy for the validation data (i.e., data not used to develop the model; $n = 218$); these prediction accuracies exceed the 70% accuracy recommended for predictive coding in formative assessment use (Haudek et al., 2012; Nehm and Haertig, 2012; Prevost et al., 2016).

Chapter 3 builds upon the predictive model development work reported by Dood et al. (2018; 2019). The predictive model reported by Dood et al. (2018) was a foundational step in alleviating time and effort issues related to scoring written formative assessment responses in
organic chemistry courses; however, this model is limited by the single assessment item from which the predictive model was developed (i.e., one-step aqueous acid–base proton transfer reaction). Other predictive models have been developed in chemistry (Haudek et al., 2012; Noyes et al., 2020), biology (Haudek et al., 2012; Moharreri et al., 2014; Prevost et al., 2016; Carter and Prevost, 2018; Sieke et al., 2019; Uhl et al., 2021), and statistics (Kaplan et al., 2014); however, these predictive models still remain limited to be used with a single assessment item.

The work reported in Chapter 3 advances the field by reporting a generalizable predictive model in lieu of a constructed response-specific model. The developed predictive model advances generalizability by not just asking the what and why scaffold for reaction mechanisms, but also for properties of molecules (i.e., why a molecule is an acid, base, or amphoteric?). Instead of scoring response based on use or non-use of the Lewis acid–base model, responses in this study were scored more conservatively based on correct use or incorrect use/non-use of the Lewis acid–base model. This model was also developed using a larger corpus of data: 15 different constructed response items and 2,095 responses. Additionally, instead of SPSS Modeler, which has limitations due to the software cost and ease of the dissemination of research findings (Dood et al., 2018), R software (R Core Team, 2019) was used because it is freely available and would be easier to disseminate research findings. The model was validated with four distinct sets of data (n = 500–6,425) with percent accuracies 84.50–92.74% and demonstrated that the development of generalizable predictive models is possible.

The predictive model detailed in Chapter 3 is a proof-of-concept for the development of generalized predictive models. Reaction mechanisms are comprised of pieces and parts that make up the mechanistic steps, such as proton transfer, nucleophilic attack, and loss of a leaving group. The development of predictive models for the different components of a reaction mechanism would give more detailed insight into students’ understanding of reaction mechanisms compared to a holistic score. However, the model described in Chapter 3 is limited by the binary outcome (i.e., correct use or incorrect use/non-use of the Lewis acid–base model) and would not
provide greater insight into students’ understanding. Therefore, Chapter 4 outlines a rubric for the levels of explanation sophistication of nucleophiles that can ultimately be used to develop predictive models for the different mechanistic components. The rubric can further assist instructors in hand-grading responses pending development of predictive models. This four-level rubric will provide more information than a binary correct use or incorrect/non-use score. It was developed through the synthesis of multiple frameworks describing students’ understanding of reaction mechanisms (Sevian and Talanquer, 2014; Caspari et al., 2018a; Bodé et al., 2019; Dood et al., 2020).

The rubric for levels of explanation of sophistication features the levels: absent, descriptive, foundational, and complex. In the absent level, responses are non-normative or do not contain an explanation of the mechanistic step involving the nucleophile. In the descriptive level, responses simply describe the nucleophile engaging in bond-forming processes. In the foundational level, nucleophilic behavior is described at a surface, electronic level (i.e., using electrons). And, in the complex level, nucleophilic behavior is described a deeper, electronic level using implicit features (e.g., partial charges, electronegativity, and electron density).

To evaluate the rubric for levels of explanation sophistication for nucleophiles, 19,936 responses were collected using 85 unique reaction mechanisms and two prompt types. The rubric was evaluated by course type (i.e., Organic Chemistry 1 and Organic Chemistry 2), nucleophile type (i.e., lone pair, sigma, pi, and combinations), reaction family type (i.e., reactions involving a carbocation, aromatic reaction, nucleophilic additions, reduction reactions, bimolecular substitution reactions, and a combination of nucleophilic addition and bimolecular substitution), involving proton transfer (i.e., presence or absence of a proton transfer reaction), and by prompt type (i.e., the original what and why scaffold and a more cued scaffold). The rubric was characterized with a set of design elements for a quality rubric (see Dawson, 2017) and results suggest that the rubric is generalizable: in a year-long organic chemistry course, across multiple prompt types, with different nucleophile types, and in a variety of reaction families and types. A
case study is also presented in Chapter 4 that details how an instructor can use the rubric to inform reflective practices and actionable items.

1.2 Evaluating the uptake of research-based instructional strategies

It is well documented that the use of active learning techniques is associated with enhanced student outcomes (e.g., achievement, passing rates, transfer rates, and graduation rates) in STEM (Springer et al., 1999; Lorenzo et al., 2006; Haak et al., 2011; Ruiz-Primo et al., 2011; Freeman et al., 2014; Wang et al., 2017; Theobald et al., 2020; Riedl et al., 2021). According to Bonwell and Eison (1991) and Freeman et al. (2014), the consensus definition of active learning is:

“[engaging] students in the process of learning through activities and/or discussion in class, as opposed to passively listening to an expert. It emphasizes higher-order thinking and often involves group work.”

Evidence-based instructional practices (EBIPs; see Stains and Vickrey, 2017) or research-based instructional strategies (RBIS; see Dean et al., 2012) are similarly used labels for teaching practices that have research and evidence to support students’ academic performance and outcomes. EBIPs and RBIS may be included under the umbrella term, “active learning” (e.g., Schroeder et al., 2007; Freeman et al., 2014; Theobald et al., 2020). In implementing these RBIS or EBIPs (referred to as “RBIS” hereafter) in introductory STEM courses, the hopes are not only to enhance student outcomes, but also to increase the number of students completing and STEM degrees; introductory STEM courses have been identified as causes for students to leave STEM disciplines (Seymour and Hewitt, 1997; Koch, 2017; Seymour and Hunter, 2019).

While there are many notable benefits of using active learning, traditional lecturing remains as the predominant mode of instruction in postsecondary STEM courses (Henderson
and Dancy, 2009; Borrego et al., 2010; Durham et al., 2018; Stains et al., 2018; López et al., 2022). Studies report various individual and contextual factors, or barriers, that influence STEM faculty members’ teaching, including (broadly speaking) the department (e.g., Austin, 1994, 1996; Fairweather, 2008; Henderson and Dancy, 2009), institutional reward systems (e.g., Braxton et al., 2002; Leslie, 2002; Fairweather, 2005; Schuster and Finkelstein, 2006; Fairweather, 2008), students (e.g., Hativa, 1995; Silverthorn, 2006; Cummings, 2008), and faculty members’ beliefs (e.g., Blackburn and Lawrence, 1995; Leslie, 2002; Gess-Newsome et al., 2003; Schuster and Finkelstein, 2006; Henderson et al., 2011). Additionally, the DBER report highlights that:

A reliable baseline understanding of faculty instructional practices in the sciences and engineering also is needed; this research should mitigate the limitations of self-report data to the extent possible and provide insights into variations by discipline, institutional type, and course type (NRC, 2012, p. 184).

It is necessary to investigate what support, guidance, and knowledge faculty need to efficaciously implement research-based practices (Henderson and Dancy, 2009) and how these factors influence the adoption of these practices. However, research on the importance of factors influencing instructional decisions relative to other factors is limited (see Quinn-Patton, 2010). Thus, multidisciplinary studies that evaluate the importance of these factors in relation to one another is not only absent from the research literature, but is required to gain insight into how to best support faculty (AAAS, 2019).

1.2.1 Overview of the work

Work presented in this dissertation (Chapters 5 and 6) builds upon large-scale, survey-based studies in postsecondary chemistry (Gibbons et al., 2018; Stains et al., 2018), mathematics (Johnson et al., 2018; Apkarian et al., 2019), and physics (Henderson and Dancy, 2009; Walter
et al., 2016; Walter et al., 2021) designed to investigate the instructional practices in these disciplines. Studies note that (i) lecture is the predominant mode of instruction, (ii) instructors have some knowledge of RBIS, but (iii) often try RBIS and discontinue their use, and (iv) barriers to the implementation of RBIS mostly relate to the context (Henderson and Dancy, 2009; Lund and Stains, 2015; Gibbons et al., 2018; Stains et al., 2018; Johnson et al., 2019; Apkarian et al., 2021).

Contextual, personal, and belief factors that influence instructors' pedagogical decisions have been described by a number of researchers (e.g., Gess-Newsome et al., 2003; Dancy and Henderson, 2010; Andrews and Lemons, 2015; Lund and Stains, 2015; Sturtevant and Wheeler, 2019). Woodbury and Gess-Newsome (2002) comprehensively reviewed the literature to understand how reform initiatives change classrooms, developed the Teacher-Centered Systemic Reform (TCSR) model, and later modified the TCSR model for higher education contexts (Gess-Newsome et al., 2003). The TCSR framework conceptualizes that teachers’ thinking and practices are the center for change that happens in classrooms at institutions (Gess-Newsome et al., 2003). In Chapters 5 and 6, the TCSR framework is used to situate 17 factors into contextual factors, personal factors, and teacher thinking factors that have been reported to influence enacted pedagogies. In these studies, contextual factors include discipline, highest degree awarded by an institution, teaching load, tenure status, student evaluations of teaching, assessment of teaching performance, class size, classroom setup, the interaction effect between class size and classroom setup, and shared decision making; personal factors include previous RBIS experience as a student, engagement in the scholarship of teaching and learning (SOTL) or DBER, and participation in teaching-focused coursework, workshops, and new-faculty experiences; and teacher thinking factors include holding a growth mindset and satisfaction with student learning.

The DBER report (NRC, 2012) that there has been a limited number of discipline-specific national surveys of faculty, which have included chemistry (Gibbons et al., 2018), engineering (Borrego et al., 2010), geosciences (Macdonald et al., 2005), mathematics (Fukawa-Connelly et al., 2016; Johnson et al., 2018; Johnson et al., 2019), and physics (Henderson and Dancy, 2009);
cross-disciplinary studies are needed. While self-report data may face possible reliability threats and conclusions based on self-report data should be interpreted with caution, studies have provided evidence suggesting observational data of teaching practices align well with self-report data in surveys (Durham et al., 2018; Gibbons et al., 2018); in an ideal study, however, observational data and self-report data would complement one another, but this is impractical. Additionally, in order for a multidisciplinary study to have the statistical power to empirically test the relative association of factors, surveys with self-reported data are necessary.

A national survey was conducted in Spring 2019 that included respondents from multiple disciplines (i.e., chemistry, mathematics, and physics) and institution types (i.e., associates, bachelor’s, and graduate degree granting). These respondents were instructors of introductory chemistry (i.e., general chemistry), mathematics (i.e., single-variable calculus), and physics (i.e., introductory quantitative physics). Five main components made up the survey: (i) course context, (ii) instructional practices, (iii) awareness and usage of active learning instructional techniques, (iv) perceptions, beliefs, and attitudes related to students, learning, and departmental context, and (v) personal demographics and experience. In these studies (Chapters 5 and 6), analysis of survey data had one main goal: to evaluate the uptake of RBIS in introductory chemistry, mathematics, and physics. To reach this goal, two approaches using multilevel modeling (Raudenbush and Bryk, 2002; Snijders and Bosker, 2012) were taken to empirically test the association of malleable factors relative to one another.

The first approach was to model the relative association of these malleable factors with percent time lecturing. In this study (Chapter 5), percent time lecturing is a proxy for time spent not using active learning; that is, students are not passively listening to the instructor and are engaged with one another while RBIS are implemented. Multilevel linear regression was used for the continuous variable (i.e., a variable that can take an infinite set of values), percent time lecturing. Data suggests that several evaluated malleable contextual, personal, and teacher thinking factors are statistically significantly associated with percent time lecturing when
controlling for all other factors used in the study: class size, classroom setup, shared decision making, previous RBIS experience as a student, engagement in SOTL, taking teaching-focused coursework, participation in teaching-focused workshops and new faculty experiences, and holding a growth mindset. Policy recommendations include (i) constructing classroom spaces that support and promote active learning, (ii) coordinating large enrollment courses with multiple course sections, (iii) offer and encourage participating in professional development programs, and (iv) encourage, recognize, and value SOTL work.

The second approach was to model the relative association of these factors with stages of RBIS adoption (i.e., awareness, tryout, and adoption; Dormant, 2011; Landrum et al., 2017). In this study (Chapter 6), the stages of RBIS adoption are a measure of where instructors are at in the continuum of adopting change (i.e., adoption of RBIS). One multilevel logistic regression was used to model RBIS tryout (versus RBIS awareness). Several malleable factors are associated with RBIS tryout: class size, the interaction effect between class size and classroom setup, previous RBIS experience as a student, participation in teaching-focused workshops, and dissatisfaction with student learning. Recommendations for change agents (i.e., people trying to enact change) promoting the tryout of RBIS among instructors include (i) identifying dissatisfied instructors, (ii) design workshops that providing training and resources to implement RBIS, and encourage instructors that teach large classes to teach in classrooms that allow for group work.

Another multilevel logistic regression was used to model RBIS adoption (versus RBIS tryout). Several malleable factors are associated with RBIS adoption: perceived influential importance of assessment of teaching performance in annual review, promotion, or tenure; class size; classroom setup; previous RIBS experience as a student; engagement in SOTL; participation in teaching-related new faculty experiences; holding a growth mindset; and satisfaction with student learning. Recommendations for change agents to get more instructors to adopt RBIS include (i) demonstrating the need to access student learning with evidence by (ii)
engaging instructors in SOTL, (iii) encouraging instructors to teach in rooms that allow for group work, and (iv) helping fostering growth mindset beliefs.

Analysis of data also led to recommendations for institutional leaders and policy makers: (i) value and reward teaching efforts, (ii) build classroom spaces that allow for group work, (iii) encourage and incentivize participation in all forms of teaching-related professional development, and (iv) showcase the benefits of using RBIS.

1.3 Summary of the work

In addition to this introductory chapter, six additional chapters are presented. Chapter 2 describes the statistical methods used in the studies presented in this dissertation; this chapter also details the how the R program works. Chapters 3 and 4 are studies that involve student explanations of organic chemistry reaction mechanisms. Chapter 3 details the development of a generalizable machine learning-based tool used to classify students’ written explanations of acid–base reaction mechanisms and properties of molecules (i.e., why a molecule can act as an acid, base, or both) as either correct use or incorrect/non-use of the Lewis acid–base model. Chapter 4 details the development and application of a generalizable rubric to evaluate level of explanation sophistication for students’ understanding of nucleophiles. Chapters 5 and 6 are studies that involve evaluating the uptake of instructional practices in introductory chemistry, mathematics, and physics courses. Chapter 5 evaluates the association of a set of 17 malleable factors with the amount of time an instructor spends lecturing as a proxy for the implementation of active learning strategies. Chapter 6 evaluates the association of these same malleable factors with the stages of research-based instructional strategy adoption. Lastly, Chapter 7 summaries the research presented in Chapters 3–6 and offers broad implications and future directions for both chemical education practitioners and researchers, as well as change strategy recommendations and outlook on sustained adoption of research-based strategies.
1.4 References


Chapter 2

Methods

This chapter details the methods used in the works presented in this dissertation (Chapters 3–6). Topics addressed in this chapter include interrater reliability and the development of an R program using machine learning techniques. Machine learning performance metrics (e.g., accuracy, confusion matrix, $F_1$ score, and the Matthews correlation coefficient) will also be described. Finally, multilevel modeling was used, and multilevel linear regression and multilevel logistic regression will be detailed.

2.1 Interrater reliability (IRR)

The work presented in Chapters 3 and 4 encompass the theme of students’ understanding of reaction mechanisms. These works heavily rely on the qualitative classification, or coding, of students’ written responses to constructed response items. Codebooks, or in these cases — rubrics, are developed that describe each classification (i.e., code) with a description (i.e., definition) and an exemplar (Creswell, 2013). One method to provide evidence for trustworthiness of the data is to establish interrater reliability (IRR); two or more raters classify (i.e., code) the same set of responses using the codebook and compare these classifications (Lange, 2011). Raters can consist of two or more humans, or in the work presented in Chapter 3, between one human and one predictive model. Several statistical methods can be used for calculating IRR, including percent agreement and Cohen’s kappa (Cohen, 1960).
2.1.1 Percent agreement

Percent agreement is one measure for calculating IRR. Percent agreement can be calculated by generating a matrix of observations (i.e., responses) with each of the raters’ classifications (Table 2.1). The difference between these classifications can be calculated for each item, and the number of no differences (i.e., zeros) is divided by the total number of observations to give percent agreement (McHugh, 2012); in the example shown in Table 2.1, the percent agreement would be 50%. Percent agreement is the percentage of observations raters agree on classifications and ranges from 0 to 100% (Lombard et al., 2002).

Table 2.1. Example calculation of percent agreement

<table>
<thead>
<tr>
<th>Observation</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Number of Zeros</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent Agreement</td>
<td>50%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Percent agreement is quick and straightforward to calculate and is interpretable. Two raters are given in this example because percent agreement calculations in Chapters 3 and 4 are between two raters; however, percent agreement can accommodate any number of raters (McHugh, 2012). In Chapters 3 and 4, percent agreement is used as one measure of IRR between two human raters, and also between one human rater and one computer rater.

2.1.2 Cohen’s kappa

Cohen’s kappa (κ) is another measure for calculating IRR. Cohen’s κ measures the percent agreement between two or more raters and corrects for any agreement that may have occurred by chance (Cohen, 1960). This statistic can be used with a range of data (e.g.,
categorical, mutually exclusive categorical, paired observations, and symmetric cross-tabulations) and raters (e.g., intraraters and interraters; Cohen, 1960). Kappa values range from −1 to +1, where 0 represents the agreement that can be expected from random chance and +1 represents the perfect agreement between raters (Cohen, 1960); κ values can fall below 0, but are unlikely in practice (Marston, 2010).

Cohen (1960) suggests that κ should be interpreted: ≤ 0 (no agreement), 0.01–0.20 (none to slight agreement), 0.21–0.40 (fair agreement), 0.41–0.60 (moderate agreement), 0.61–0.80 (substantial agreement), and 0.81–1.00 (almost perfect agreement). However, some (e.g., McHugh, 2012) suggest that Cohen’s κ values should interpreted more conservatively (i.e., less lenient) depending on the context and implications of the IRR.

### 2.1.3 IRR in this work

The published studies in Chapters 3 and 4 use percent agreement as a measure of accuracy between two human raters, and in Chapter 3, also between a human rater and a computer rater. In Chapter 3, percent agreement ranges from 83% (before discussion of classifications) to 98% (after discussion of classifications). In Chapter 4, percent agreement ranges from 53% (before discussion of classifications) to 96% (after discussion of classifications).

The published study in Chapter 3 also uses Cohen’s κ as a measure of agreement. Cohen’s κ values are reported for validation of predictive models developed using machine learning techniques. For validations sets, Cohen’s κ values range 0.67–0.80 indicating substantial agreement as suggested by Cohen (1960). Each of these validation sets are parsed by prompt type; Cohen’s κ values range 0.45–0.88 indicating a range of moderate agreement to almost perfect agreement. Cohen’s κ values that correspond to moderate agreement (i.e., 0.41–0.60) indicate that there are more disagreements than agreements and is reflected in the machine learning model having difficulty parsing true negatives from false positives (see Chapter 2.3.1).
2.2 R program

An R program (R Core Team, 2019) was developed using automated text analysis and machine learning techniques to assign a binary classification to written responses to constructed response items. There are four components of this development process: (i) collecting and preparing data, (ii) data preprocessing and feature extraction, (iii) model training, and (iv) model evaluation.

In collecting and preparing data, Qualtrics (2020) was used to deploy surveys containing constructed response items. These surveys were given to students in the first or second semester of a year-long organic chemistry course before a term or final examination. After the time period where the survey is active, data are downloaded from Qualtrics (2020), deidentified, and classified by a human rater. The R program will then read a .csv file that contains a list of responses and scores as assigned by a human rater; in Chapter 3, this score is either 0 (i.e., incorrect or non-use of the Lewis acid–base model) or 1 (i.e., correct use of the Lewis acid–base model).

In data preprocessing, the text is prepared for model training. The training data (i.e., corpus) first undergoes a series of commands that will convert the case of characters to lowercase (Kwartler, 2017), and remove alphanumeric and special characters, and punctuation. English stop words, or commonly used words with little meaning (e.g., “I,” “ours,” and “yours”), are then removed. A custom stop word dictionary was created that included words that did not relate to the use of acid–base chemistry models or words that are not used to describe organic reaction mechanisms. Stop words from this custom dictionary were also removed. Another custom dictionary of patterns and replacements was also created. Patterns are misspelled words and replacements are the proper spellings that will substitute the pattern. For example, misspelled words such as “nuclaeophilic” are replaced with the correct spelling, “nucleophilic.” For word normalization, a process called lemmatization was used; in lemmatization, inflected words (i.e., patterns) are normalized to a root word, and in the case of this program, the singular verb (i.e., replacements). For example, “attack,” “attacked,” and “attacking” all become “attacks.” Then,
replacements are substituted for patterns. Lastly, leading, trailing, and excess white spaces were removed yielding a set of 257 words that relate to use of acid–base models or reaction mechanisms. These series of steps were completed with commands from the ‘tm’ (Feinerer et al., 2008) and ‘qdap’ (Rinker, 2020) packages in R (R Core Team, 2019).

In feature extraction, the 257 words that resulted after data preprocessing are known as features. These features are known as unigrams, or a set of single words. The unigrams are pared into a document-term matrix, where the matrix represents the pattern of the presence or absence of the features in a response; this matrix is n responses by 257 features. The value in each matrix cell is equivalent to the term frequency, or the number of times the term t appears in a response (Kwartler, 2017).

For model training, support vector machines (SVM) with a linear basis function kernel (Cortes and Vapnik, 1995) was used as the algorithm to classify responses for correct use or incorrect/non-use of the Lewis acid–base model. To simplify, the data has two outcomes (i.e., correct use or incorrect/non-use) and these outcomes can be determined by using two predictor variables. In this example, plotting the data (Figure 2.1) with the two predictor variables against one another gives a two-dimensional plot (i.e., a box) where one outcome (i.e., orange dots) can be separated from the other outcome (i.e., green dots) using a line; note: this is a perfect case scenario where the data is perfectly separated and thus predicted. If a third predictor variable is added in and used to separate and predict the outcome, the plot is now three-dimensional (i.e., a cube; Figure 2.2). In this example, a plane can be used to separate the data; the example shown in Figure 2.2 respresents a perfect case scenario where the data is perfectly separated. For the model in this study, there are 257 predictor variables, or features, that are used to separate the outcomes; a 257-dimension case is difficult to visualize, but the outcomes can be separated with a hyperplane.
In SVM in this model, data are mapped in multidimensional space and the cost penalty parameter is optimized; the parameter regulates how much the algorithm should avoid misclassification of the data by optimizing the hyperplane to classify the data (Cortes and Vapnik, 1995; Joachims, 2002; Gaspar et al., 2012). This hyperplane attempts to maximize the distance between support vectors (i.e., data points nearest to the hyperplane) between the two classes of data. Model training was completed with commands from the ‘caret’ (Kuhn, 2008) package in R (R Core Team, 2019).

In model validation, the trained model is used to predict scores of the responses and are compared with scores from a human rater. Three validation methods are performed: cross-
validation, split-validation, and external validation. First, in cross-validation, the training corpus is broken up into $k$ groups, with one group used to train the model and $k - 1$ group(s) used to validate the model (Ramasubramanian and Singh, 2019); in Chapter 3, this was done using a two-fold, twice-repeated cross-validation method where $k = 2$ that was repeated twice. Second, in split-validation, the responses not used in the training corpus were parsed into a stratified set, which composed of a random sample of 100 responses from each of the five prompt types, and a remaining set, which composed of the remainder of the training corpus data. The model was evaluated with the stratified and remaining split-validation sets. Third, in external validation, the model was evaluated on a set of responses that were not included in the training corpus (Vabala et al., 2019) and that were also collected using a new set of prompts.

2.3 Machine learning performance metrics

Machine learning performance metrics include, but are not limited to Cohen’s $\kappa$ (see Chapter 2.1.1), confusion matrix (including true positive, true negative, false positive, and false negative), accuracy, $F_1$ score, and the Matthews correlation coefficient. These metrics are used to evaluate predictive model results when compared to a human rater.

2.3.1 Confusion matrix

The pairing of a human-rated score and a computer-rated score can result in four possible outcomes: both human- and computer-rated scores are correct Lewis use given that the human-rated score is correct Lewis use (i.e., true positive, TP), both human- and computer-rated scores are incorrect Lewis use/non-use given that the human-rated score is incorrect Lewis use/non-use (i.e., true negative, TN), human-rated score as incorrect Lewis use/non-use and computer-rated score as correct Lewis use given that the human-rated score is incorrect Lewis use/non-use (i.e., false positive, FP), and human-rated score as correct Lewis use and computer-rated score as incorrect Lewis use/non-use given that the human-rated score is correct Lewis use (i.e., false
negative, FN). The combination of these four outcomes is given in a confusion matrix (Figure 2.3) which demonstrates how each of the outcomes are determined (Chicco and Jurman, 2020).

In Chapter 3, each of the four classification outcomes are given for each of the validation sets and prompt subsets. For validation sets, TP range 88.07–96.87%, FN range 3.13–11.93%, TN range 74.74–80.04%, and FP range 19.96–25.26%. For validation prompt subsets, TP range 85.71–97.73%, FN range 2.14–14.29%, TN range 41.67–91.25%, and FP range 8.75–58.33%.

**Figure 2.3. Confusion matrix**

### 2.3.2 Accuracy

Accuracy (Equation 2.1) is defined as the sum of true positive and true negative over the sample size and represents the ratio between the correctly predicted scores and all the scores in the data (Chicco and Jurman, 2020). Percent accuracy (i.e., accuracy times 100) ranges between 0% and 100%, with 100% indicating perfect classification (Chicco and Jurman, 2020).

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{(Eq. 2.1)}
\]

In Chapter 3, percent accuracies are given for each of the validation sets and prompt subsets. Percent accuracies for the validation sets range 84.50–92.74% and percent accuracies for validation prompt subsets range from 75.00–94.41%.
2.3.3 F₁ score

Two metrics, precision (Equation 2.2) and recall (Equation 2.3), are common metrics that involve two confusion matrix categories (Chicco et al., 2021). Precision, also called the positive prediction value, represents the ratio between true positive scores (i.e., responses classified as correct Lewis use by both the human and computer) and all scores that were predicted positive (i.e., all responses classified as correct Lewis use by the computer), and ranges from 0 to 1 where 1 is the best value (Ramasubramanian and Singh, 2019). Recall, also called the true positive rate or sensitivity, represents the ratio between true positive scores (i.e., responses classified as correct Lewis use by both the human and computer raters) and all scores that were actually positive (i.e., all responses classified as correct Lewis use by a human rater), and ranges from 0 to 1 where 1 is the best value (Ramasubramanian and Singh, 2019).

\[
\text{Precision} = \frac{TP}{TP + FP} \quad \text{(Eq. 2.2)}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad \text{(Eq. 2.3)}
\]

The F₁ score (Equation 2.4) attempts to combine precision and recall into a single metric, and is the harmonic mean of precision and recall, and ranges from 0 to 1 with an F₁ of 1 indicating perfect classification (Kwartler, 2017; Chicco and Jurman, 2020). Three issues arise with F₁: (i) F₁ score is independent from TN, (ii) F₁ is not symmetric for class swapping (i.e., swapping TP and FP along with swapping TN and FN), (iii) and F₁ does not consider the proportion in each confusion matrix class (Chicco and Jurman, 2020).

\[
F₁ = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad \text{(Eq. 2.4)}
\]
In Chapter 3, $F_1$ is given for each of the validation sets and prompt subsets. The $F_1$ score for the validation sets range 0.88–0.95 and the $F_1$ score for validation prompt subsets range from 0.78–0.97.

### 2.3.4 Matthews correlation coefficient

The Matthews correlation coefficient (MCC, Equation 2.5) is an alternative metric to $F_1$ and aims to correct the limitations of $F_1$ since MCC is unaffected by unbalanced data sets and accounts for TN (Matthews, 1975; Baldi et al., 2000; Chicco and Jurman, 2020). MCC values range −1 to +1 with a MCC of zero indicating classification from random chance and an MCC of +1 indicating perfect classification, and is interpreted similarly to Cohen’s kappa. MCC has been demonstrated to be a more reliable metric in evaluating a two-class confusion matrix over accuracy and $F_1$ (Chicco and Jurman, 2020; Chicco et al., 2021).

\[
MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}
\]  

(Eq. 2.5)

In Chapter 3, the MCC is given for each of the validation sets and prompt subsets. The MCC for the validation sets range 0.67–0.78 and the MCC for validation prompt subsets range from 0.46–0.88.

### 2.4 Multilevel modeling

The studies presented in Chapters 5 and 6 encompass the theme of evaluating the uptake of instructional practices. These works rely on the application of multilevel modeling techniques. Multilevel linear regression is used in Chapter 5 and multilevel logistic regression is used in Chapter 6.
2.4.1 Multilevel linear regression

Linear regression is a statistical method used to model a continuous dependent variable using one or more predictor variables. In simple linear regression, one predictor variable is used and in multiple linear regression, multiple predictor variables are used; the relationship between the predictor variables and the dependent variable is modeled using linear functions (Montgomery et al., 2021). In linear regression, predictor variables in logistic regression can be continuous, categorial, or combinations of continuous and categorical variables (Montgomery et al., 2021).

Four assumptions are associated with linear regression models: linearity, homoscedasticity, independence, and normality (Montgomery et al., 2021). However, in the study described in Chapter 5, observations are not independent. This study included 2,382 instructors (level 1) nested in 1,405 departments (level 2) at 749 institutions (level 3). Instructors at the same department and at the same institution violate the assumption of independence, and therefore, a multilevel modeling approach should be taken (Raudenbush and Bryk, 2002; Snijders and Bosker, 2012). With the nesting structure of multilevel models, observations are interdependent (Sommet and Morselli, 2017).

In Chapter 5, a three-level multilevel linear regression is used. In this study, the dependent variable is the percent time spent lecturing in introductory chemistry, mathematics, and physics courses. This dependent, continuous variable (i.e., a variable that can take an infinite set of values) is modeled with 17 malleable predictor variables. For the multilevel linear regression used in Chapter 5, the random intercepts model can be written as in Equation 2.6; random slopes are not modeled. Here, \( i, j, \) and \( k \) index levels 1, 2, and 3, respectively. \( \beta_n \) represent the level 2 predictor coefficients and \( \pi_n \) represent the level 1 predictor coefficients (Raudenbush and Bryk, 2002; Snijders and Bosker, 2012). A multilevel regression coefficient is the estimate of the change in the dependent variable per unit increase in the value of the predictor variable when all other variables are accounted for and held constant (Raudenbush and Bryk, 2002; Snijders and Bosker, 2012).
\[
LECTURE_{ijk} = \gamma_{000} + \beta_{001} \text{CHEM}_{jk} + \beta_{002} \text{PHYS}_{jk} \\
+ \beta_{003} \text{BACH}_{jk} + \beta_{004} \text{GRAD}_{jk} + \pi_{100} \text{SIZE}_{ijk} \\
+ \pi_{200} \text{ROOM}_{ijk} + \pi_{300} \text{SIZE}_{ijk} \times \text{ROOM}_{ijk} \\
+ \pi_{400} \text{DECISION}_{ijk} + \pi_{500} \text{LOAD}_{ijk} \\
+ \pi_{600} \text{TENURED}_{ijk} + \pi_{700} \text{TENURETRACK}_{ijk} \quad \text{(Eq. 2.6)} \\
+ \pi_{800} \text{SET}_{ijk} + \pi_{900} \text{ATP}_{ijk} + \pi_{1000} \text{RBIS}_{ijk} \\
+ \pi_{1100} 8.56 \text{SOTL}_{ijk} + \pi_{1200} \text{TFC}_{ijk} \\
+ \pi_{1300} \text{WKSP}_{ijk} + \pi_{1400} \text{NFE}_{ijk} \\
+ \pi_{1500} \text{GROWTH}_{ijk} + \pi_{1600} \text{SATISFACTION}_{ijk} \\
+ r_{ij} + \mu_{jk} + \mu_{k}
\]

### 2.4.2 Multilevel logistic regression

Logistic regression is a statistical method used to model group membership using a set of predictor variables. Like linear regression, predictor variables in logistic regression can be continuous, categorial, or combinations of continuous and categorical variables (Sheskin, 2011). Dependent variables can be binomial (i.e., only two outcomes; Cox, 1958) or multinomial (i.e., three or more outcomes; Engel, 1988). A clear difference between linear regression and logistic regression is that in linear regression, the dependent variable is continuous (i.e., is not bounded), whereas in logistic regression, the dependent variable is binary (i.e., either 0 or 1; DeMaris, 1995). Additionally, logistic regression gives the conditional probability that a dependent variable equals one at a given value of a predictor variable, and distinguishes from linear regression which gives the predicted mean value of a dependent variable given a value of a predictor variable (Sommet and Morselli, 2017). Logistic regression has been previously used in chemistry education research (e.g., Legg et al., 2001; Emenike et al., 2013; Tang et al., 2014; Dood et al., 2018; Hinds and Shultz, 2018; Srinivasan et al., 2018; Dood et al., 2020; Young and Lewis, 2022); however, the use of multilevel logistic regression has been limited (e.g., Harris et al., 2020).

Multilevel logistic regression additionally accounts for nested data structures (Raudenbush and Bryk, 2002; Snijders and Bosker, 2012). For example, in the study reported in Chapter 6, there are 2,303 instructors that come from 1,371 departments at 741 institutions. That is, instructors (level 1) are nested within departments (level 2) that are nested within institutions.
(level 3). With this data structure, a standard logistic regression analysis would not suffice as this data structure violates the assumption of independence (Raudenbush and Bryk, 2002; Snijders and Bosker, 2012). Therefore, observations are interdependent and multilevel logistic regression seeks to within-level effects from between-level effects (Sommet and Morselli, 2017).

In Chapter 6, two-level multilevel logistic regressions are used. Suppose that a binary response, \( y_{ijk} \), is observed for an instructor \( i \) in department \( j \) in institution \( k \). The probability of the response equal to one is defined as \( p_{ijk} = P(y_{ijk} = 1) \), and \( p_{ijk} \) is modeled using a logit link function with the assumption that \( y_{ijk} \) follows a Bernoulli distribution (Guo and Zhao, 2000). For the multilevel logistic regression used in Chapter 6, the random intercepts model can be written as in Equation 2.7; random slopes are not modeled. Here, \( i, j, \) and \( k \) index levels 1, 2, and 3, respectively. \( \beta_n \) represent the level 2 predictor coefficients and \( \pi_n \) represent the level 1 predictor coefficients (Raudenbush and Bryk, 2002; Snijders and Bosker, 2012). A regression coefficient is the estimate in the increase in log odds of the dependent variable per unit increase in the value of the predictor variable when all other variables are accounted for and held constant (Szumilas, 2010).

\[
\ln \left[ \frac{p_{ijk}}{1 - p_{ijk}} \right] = \beta_{001} \text{CHEM}_{jk} + \beta_{002} \text{PHYS}_{jk} + \beta_{003} \text{BACH}_{jk} + \beta_{004} \text{GRAD}_{jk} + \pi_{100} \text{SIZE}_{ijk} + \pi_{200} \text{ROOM}_{ijk} + \pi_{300} \text{SIZE}_{ijk} \times \text{ROOM}_{ijk} + \pi_{400} \text{DECISION}_{ijk} + \pi_{500} \text{LOAD}_{ijk} + \pi_{600} \text{TENURED}_{ijk} + \pi_{700} \text{TENURETRACK}_{ijk} + \pi_{800} \text{SET}_{ijk} + \pi_{900} \text{ATP}_{ijk} + \pi_{1000} \text{RBIS}_{ijk} + \pi_{1100} \text{8.56SOTL}_{ijk} + \pi_{1200} \text{TFC}_{ijk} + \pi_{1300} \text{WKSP}_{ijk} + \pi_{1400} \text{NFE}_{ijk} + \pi_{1500} \text{GROWTH}_{ijk} + \pi_{1600} \text{SATISFACTION}_{ijk} \quad (\text{Eq. 2.7})
\]

An odds ratio is the probability of the dependent variable occurring over the probability of the dependent variable not occurring (Szumilas, 2010). Odds ratios are measures of the odds that the dependent variable will occur if a predictor variable increases by one unit (or is present
in the case of binary variables) compared to the odds of the dependent variable occurring if the predictor variable does not increase by one unit (or is not present in the case of binary variables) when all other variables are accounted for and held constant; therefore, an odds ratio is based on one unit change in the predictor variable (Szumilas, 2010; Ranganathan et al., 2017). The odds ratio is equal to the exponential of the regression coefficient (Ranganathan et al., 2017). For an odds ratio greater than one, the odds of the dependent variable occurring are higher when the predictor variable is increased by one unit (or present for a binary variable), and for an odds ratio less than one, the odds of the dependent variable occurring are lower when the predictor variable is increased by one unit (or present for a binary variable). For example, in the Tryout Model, RBIS awareness was coded as 0 and RBIS tryout was coded as 1. The odds ratio for RBIS use as a student is 6.34; the interpretation of this odds ratio can be as follows: when all other variables are held constant, the odds of RBIS tryout for instructors that were students in courses taught using RBIS are 6.34 times higher than for instructors that have not taken any courses using RBIS.

In Chapter 6, two multilevel binomial logistic regression model is used to evaluate the association of a set of 17 malleable factors with the stages of research-based instructional strategy (RBIS) adoption (Dormant, 2011; Landrum et al., 2017). The RBIS Adoption Scale was used to measure the stage of RBIS adoption for introductory chemistry, mathematics, and physics instructors in the United States; there are three stages of adoption described in this study: awareness, tryout, and adoption. In the Tryout Model, RBIS awareness was coded as 0 and RBIS tryout was coded as 1. In the Adoption Model, RBIS tryout was coded as 0 and RBIS adoption was coded as 1. For both models, odds ratios were used to evaluate the magnitude of association of the malleable factors with RBIS tryout and RBIS adoption.
2.5 References


Chapter 3

Development of a machine learning-based tool to evaluate correct Lewis acid–base model use in written responses to open-ended formative assessment items

3.1 Note to Reader

This chapter is a published manuscript in Chemistry Education Research and Practice. This chapter is reproduced from:


with permission from the Royal Society of Chemistry. Further copyright information can be found in Appendix E.1.

This work has co-authors. Amber J. Dood was a former graduate student at the University of South Florida who initially collected data used in this study and provided assistance in the development of the machine learning model. Daniel Cruz-Ramírez de Arellano and Kimberly B. Fields are organic chemistry instructors at the University of South Florida who allowed for data collection in their classes. Jeffrey R. Raker is the principal investigator for this project.

3.2 Abstract

Acid–base chemistry is a key reaction motif taught in postsecondary organic chemistry courses. More specifically, concepts from the Lewis acid–base model are broadly applicable to understanding mechanistic ideas such as electron density, nucleophilicity, and electrophilicity;
thus, the Lewis model is fundamental to explaining an array of reaction mechanisms taught in organic chemistry. Herein, we report the development of a generalized predictive model using machine learning techniques to assess students’ written responses for the correct use of the Lewis acid–base model for a variety ($N = 26$) of open-ended formative assessment items. These items follow a general framework of prompts that ask: why a compound can act as (i) an acid, (ii) a base, or (iii) both an acid and a base (i.e., amphoteric)? Or, what is happening and why for aqueous proton-transfer reactions and reactions that can only be explained using the Lewis model. Our predictive scoring model was constructed from a large collection of responses ($N = 8520$) using a machine learning technique, i.e., support vector machine, and subsequently evaluated using a variety of validation procedures resulting in overall 84.5–88.9% accuracies. The predictive model underwent further scrutiny with a set of responses ($N = 2162$) from different prompts not used in model construction along with a new prompt type: non-aqueous proton-transfer reactions. Model validation with these data achieved 92.7% accuracy. Our results suggest that machine learning techniques can be used to construct generalized predictive models for the evaluation of acid–base reaction mechanisms and their properties. Links to open-access files are provided that allow instructors to conduct their own analyses on written, open-ended formative assessment items to evaluate correct Lewis model use.

3.3 Introduction

Acidity and basicity are foundational concepts in the post-secondary organic chemistry curriculum (Friesen, 2008; Cartrette and Mayo, 2011; McClary and Talanquer, 2011; Bhattacharyya, 2013; Stoyanovich et al., 2015; Brown et al., 2018; Nedungadi and Brown, 2021). Acid–base reactions are among the introductory reaction types (sometimes referred to as reaction mechanism motifs) covered (Stoyanovich et al., 2015) and are routine components of more complex reaction mechanisms (Friesen, 2008; Stoyanovich et al., 2015). Concepts surrounding acid–base chemistry are universal in rationalizing organic reaction mechanisms and therefore it
is critical that students have a strong foundation in acid–base theories (Graulich, 2015). Thus, how acid–base chemistry knowledge is taught and evaluated influences learning.

Current assessment items and tools are limited in measuring understanding of acids and bases. Concept inventories and multiple-choice-based assessments exist to measure such understanding. For example, the ACID-I concept inventory is a multi-choice assessment that evaluates student conceptions about acid strength (McClary and Bretz, 2012). Other examples of concept inventories include a test measuring high school students’ understanding of acids and bases (Cetin-Dindar and Geban, 2011), the Acid–Base Reactions Concept Inventory (ABCI) used to measure understanding of acid–base reactions in high school students through postsecondary organic chemistry (Jensen, 2013), and the Measuring Concept progressions in Acid–Base chemistry (MCAB) instrument intended to address concepts covered in general chemistry (Romine et al., 2016). However, a known shortcoming of multiple-choice assessments is that students are forced to choose an answer; this may give the false illusion that students hold certain conceptions when they could have been guessing (Birenbaum and Tatsuoka, 1987). An alternative to multi-choice-based assessments are oral examinations (including research-based think-aloud interview protocols). For example, through interviews McClary and Talanquer (2011) found that students use several different models and even mixed models when explaining acid and base strength. Such Socratic, dialogue-rooted assessments are impractical, particularly in courses with large enrollments (e.g., Roecker, 2007; Dicks et al., 2012). Measurement of acid–base concept understanding is further complicated in that observation-based studies have shown that students are able to draw acid–base mechanisms, or other mechanisms without understanding the concepts behind the representation (Bhattacharyya and Bodner, 2005; Ferguson and Bodner, 2008; Grove et al., 2012).

Constructed-response assessment items that require a student to explain better measure acid–base understanding and chemistry concepts in general. Such open-ended items are vital for instructors to gain insight into students’ understanding and important for amending instruction to
improve student learning (Bell and Cowie, 2001; Fies and Marshall, 2006; MacArthur and Jones, 2008). Assessments where students are free to respond in complete thoughts to demonstrate their conceptual understanding provide deeper insight to instructors and send a message to students that deep understanding is important (Birenbaum and Tatsuoka, 1987; Scouller, 1998; Cooper et al., 2016; Stowe and Cooper, 2017; Underwood et al., 2018). However, open-response items, as with oral examinations, are not pragmatic for instructors’ use in large-enrollment courses and are not feasible for use with just-in-time-teaching (Novak et al., 1999).

Computer-assisted predictive scoring models have been built to evaluate text-based responses to open-ended items. Use of predictive scoring models reduce evaluation time, making in-class use possible (e.g., Haudek et al., 2011, 2012; Prevost et al., 2016; Dood et al., 2018, 2020a; Noyes et al., 2020). Some of these models have been designed to specifically evaluate student understanding of acid–base chemistry through written explanations; for example, Haudek et al. (2012) used a predictive model to identify levels of correct explanations of acid–base chemistry use in a biology course and Dood et al. (2018, 2019) built a predictive model to classify use of the Lewis acid–base model in responses to a proton-transfer reaction. A meta-analysis of machine learning-based science assessments has shown that these techniques are primarily employed in middle/secondary and postsecondary environments, spanning the general science domain to more specific STEM disciplines (Zhai et al., 2020). These predictive models serve a multitude of functions (e.g., assigning scores, classifying responses, identification of key concepts), use a variety of computational algorithms (e.g., regression, decision trees, Bayes, and support vector machines), and are built with an array of software (e.g., SPSS Text Analytics, Python, R, c-rater, SIDE/LightSIDE; cf., Zhai et al., 2020). Such scoring models, though, have been prompt-specific, meaning that varying the prompt by an instructor may yield the scoring models invalid; thus, new predictive scoring models must be developed for each assessment item, a process that requires hundreds of responses and multiple hours of development.
Use of the Lewis model in explaining acid–base reactions is key to mastery of organic chemistry. The goal of the work we report herein is to construct a computer-assisted predictive scoring model that detects correct use of the Lewis acid–base model in response to open-ended acid–base assessment items. This work seeks to build a single generalized predictive model that has demonstrable accuracy across an array of assessment items with the potential for instructors to use the predictive scoring model to evaluate responses to assessment items beyond those reported herein. Our results provide a foundation for more complex and nuanced predictive models built by technologies based on machine learning to evaluate understanding of reaction mechanisms beyond the foundational acid–base reaction.

3.3.1 Student understanding of acid–base models

Paik (2015) outlined the three main acid–base models taught in the introductory, postsecondary chemistry curricula: Arrhenius, Brønsted–Lowry, and Lewis. In the Arrhenius model, acids dissociate in water and increase the concentration of hydrogen ions (H\(^+\)) and bases dissociate in water and increase the concentration of hydroxide ions (OH\(^-\)). In the Brønsted–Lowry model, acids act as proton donors and bases act as proton acceptors. In the Lewis model, acids are defined as electron pair acceptors and bases are defined as electron pair donors. Students are typically introduced to the Arrhenius and Brønsted–Lowry models in secondary education (i.e., high school) chemistry courses and in postsecondary general chemistry courses. It has been noted by many that the Lewis model is glanced over, if at all presented in these courses, and emphasis is placed on the other two models (Shaffer, 2006; Drechsler and Van Driel, 2008; Cartrette and Mayo, 2011; Paik, 2015).

While the three models are interconnected, mental conceptions of acidity and basicity by learners suggest a lack of distinctness between the models: McClary and Talanquer (2011) reported that student conceptions are dependent on a compound’s surface features and that students struggle to switch between models. Studies suggest that learners struggle with acid–
base models across the postsecondary curriculum and even into the graduate chemistry curriculum (Bhattacharyya, 2006; Cartrette and Mayo, 2011; McClary and Talanquer, 2011; Bretz and McClary, 2015; Stoyanovich et al., 2015; Dood et al., 2018; Schmidt-McCormack et al., 2019). This key confusion originates partially in the ambiguous relationships between acid–base models, a lack of clear distinction between the models (Schmidt and Volke, 2003; Drechsler and Schmidt, 2005). Additionally, the sheer number of models may also cause confusion (Ültay and Çalik, 2016).

Students have difficulty switching between acid–base models when solving problems, especially when the Lewis model is the most appropriate model (Cartrette and Mayo, 2011; Tarhan and Acar Sesen, 2012). Model confusion results when students attempt to apply models in circumstances where they are not applicable; for example, Tarhan and Acar Sesen (2012) found that students had greatest difficulty with Lewis acid–base reactions, such as with the reaction between ammonia and trifluoroborane, in which less than half of the participants in their study could correctly identify the Lewis acid–base reaction. Students also struggle to incorporate concepts of the Lewis model within their current understanding of acids and bases; for example, Cartrette and Mayo (2011) observed that study participants struggled with using terminology such as nucleophilicity and electrophilicity when trying to apply those concepts to a proton-transfer reaction. However, Crandell et al. (2019) noted that when students are primed to consider “why a reaction can only be described using the Lewis model,” those students are more likely to describe the transfer of electrons when asked what is happening in a given reaction and why that reaction occurs.

Students struggle with defining, giving examples, and explaining the function of acids and bases (Bhattacharyya, 2006; Cartrette and Mayo, 2011; Tarhan and Acar Sesen, 2012; Schmidt-McCormack et al., 2019). Cartrette and Mayo (2011) found that second-semester organic chemistry students could correctly define and give examples of Brønsted–Lowry acids and bases; however, less than half of their sample could do the same for Lewis acids and bases. In a study
by Tarhan and Acar Sesen (2012), a majority of students also could not correctly classify a substance as a Lewis base. Schmidt-McCormack et al. (2019) reported that students were able to correctly identify Lewis acids, but were unable to describe how or why the compound acts as a Lewis acid. Additionally, Bhattacharyya (2006) showed that chemistry graduate students were able to provide definitions for acids and bases but were unable to apply their mental models to different situations. This suggests that despite the centrality of Lewis acids and bases in the curriculum, many students are leaving our courses with an underdeveloped understanding of acidity and basicity.

The amphoteric property of some chemical species, i.e., the species can act as both an acid and a base, poses further complications in Lewis acid–base understanding (Schmidt and Volke, 2003; Drechsler and Schmidt, 2005; Drechsler and Van Driel, 2008). The amphoteric property of water can be explained using both the Brønsted–Lowry and Lewis models; however, that explanation fails when the Arrhenius model is invoked. Students are perhaps uncomfortable with the classification of water as an acid or a base due to a reliance on the Arrhenius model (Schmidt and Volke, 2003; Drechsler and Schmidt, 2005). Schmidt (1997) found that some students hold the conception that conjugate acid–base pairs must be charged ions; students struggled with identifying a neutral conjugate acid or base as such. While such confusion is not limited to amphoteric species, charged species versus neutral species is central to the concept of amphoterism. While the Brønsted–Lowry model can be used to describe the amphoteric property, not all amphoteric compounds are proton donors and acceptors; therefore, understanding amphoterism using the Lewis acid–base model is more generalizable.

Understanding the Lewis acid–base model is vital to explaining reactivity in organic chemistry (Shaffer, 2006; Bhattacharyya, 2013; Stoyanovich et al., 2015; Cooper et al., 2016; Dood et al., 2018). Cooper et al. (2016) found that students who utilized the Lewis model to explain an aqueous, acid–base proton-transfer reaction had a higher likelihood of producing its accepted arrow-pushing mechanism. Expanding upon that work, Dood et al. (2018) found that
students who were able to explain a proton-transfer reaction using the Lewis model had higher scores on an acid–base related examination in an organic chemistry course. Crandell et al. (2019) built upon the work of Cooper et al. (2016) by adding a mechanism that can only be explained using the Lewis acid–base model, demonstrating that features of the assessment prompt influence the degree to which particular acid–base models are invoked in explanations.

### 3.3.2 Role of formative assessments in developing acid–base model use by students

Writing is a way of knowing (Reynolds et al., 2012); in the context of chemistry courses, writing is an opportunity to reflect on processes about chemical concepts and on solving problems. The process of writing by explaining primes thinking and engagement at a deeper level (Rivard, 1994; Bangert-Drowns et al., 2004; Reynolds et al., 2012). Deliberate constructed-response items, centered on writing to learn, can better promote student understanding of chemical concepts (Birenbaum and Tatsuoka, 1987; Cooper, 2015; Stowe and Cooper, 2017; Underwood et al., 2018). Construction of explanations and engagement in argument from evidence are two practices that aid in the development of conceptual understanding (National Research Council, 2012). Therefore, constructed-response items can be used to elicit reasoning through argumentation and construction of explanations. It has been suggested that deeper learning can be accomplished by asking students why a phenomenon occurs (Cooper, 2015). Cooper et al. (2016) reported that asking what is happening and why it happens for chemical reactions better elicited student explanations. This scaffold has been expanded upon by Dood et al. (2018, 2019) and Crandell et al. (2019) for acid–base chemistry; similarly, asking students what is happening and why for SN1 (Dood et al., 2020a) and SN2 (Crandell et al., 2020) reactions better elicited mechanistic understanding. While constructed-response items provide a formative means for evaluation of student understanding, they are limited in the time required to provide feedback to students.
There has been one instance, to date, of the development of a computer-assisted scoring model used to evaluate written responses to an acid–base assessment item in chemistry courses. Dood et al. (2018) reported a predictive scoring model that was then further optimized and generalized by Dood et al. (2019). While the model reported by Dood et al. (2018, 2019) is a key first step to solving the issue of practicality of speed in scoring and for use with large-enrollment courses, the work of Dood et al. still falls into the category of a predictive scoring model for a single assessment item. For evaluation of responses to open-ended assessment items, which includes predictive scoring models, to become widely developed, more generalizable models are necessary.

### 3.3.3 Argument for a generalized predictive model

Computer-assisted scoring models, based on brute-force lexical analyses (Haudek et al., 2011) and machine learning (Zhai et al., 2020) techniques are becoming more commonplace in the analysis of open-ended written assessments in post-secondary STEM course contexts: for example, chemistry (Haudek et al., 2012; Dood et al., 2018, 2020a; Noyes et al., 2020), biology (Haudek et al., 2012; Moharreri et al., 2014; Prevost et al., 2016; Carter and Prevost, 2018; Sieke et al., 2019; Uhl et al., 2021), and statistics (Kaplan et al., 2014). These predictive models are assessment item-specific; that is, each predictive model is designed and optimized to evaluate responses to a specific assessment prompt or prompt type. The time necessary to develop a single predictive model, including data collection (and sometimes additional data collection with a modified prompt), is somewhat prohibitive (Urban-Lurain et al., 2009). Rather than generalized predictive models, more limited models have been developed. For example, predictive models have been developed to measure structure–function relationships in biology with a specific assessment prompt (Carter and Prevost, 2018), rather than a generalized predictive model to measure structure–function understanding with an array of assessment prompts. When considering evaluating knowledge of Lewis acid–base understanding, we noted that while specific chemical species were needed as context from which a student would respond, the language
about those species and the interaction of multiple species is consistent. For example, when
describing an acid–base reaction, identifying areas of electron sufficiency and deficiency are key
to correctly invoking a Lewis acid–base explanation; thus, a predictive scoring model focused on
identifying instances of electron density is more important than if the name of a chemical species
was identified (e.g., a chloride ion), the latter being prompt specific.

In an effort to provide more tools to evaluate understanding of Lewis acid–base chemistry
coupled with a desire to advance work in predictive scoring models for assessments in STEM, we
sought to develop a generalized predictive model to evaluate correct use of the Lewis acid–base
model in responses to an array of assessment items.

**3.3.4 Research question**

This study was guided by one primary research question:

What level of accuracy can be achieved for a generalized predictive model developed using
machine learning techniques that predicts correct use of the Lewis acid–base model for a variety
of constructed response items?

**3.4 Methods**

This work was conducted under application Pro#00028802, “Comprehensive evaluation
of the University of South Florida’s undergraduate and graduate chemistry curricula” as reviewed
by the University of South Florida’s Institutional Review Board on December 13, 2016. Per
Institutional Review Board criteria, the activities were determined to not constitute research.

**3.4.1 Constructed response items**

The constructed response item for the proton-transfer reaction used in this study was first
reported by Cooper et al. (2016), modified for use by Dood et al. (2018), and lastly expanded to
other hydrohalic acids and water reactions by Dood et al. (2019). Data used in this study in
determining correct Lewis acid–base model use in response to aqueous proton-transfer constructed items were used previously (Dood et al., 2018, 2019) in logistic regression predictive models; herein, those data are analyzed differently and in conjunction with broader data.

In total, response to 15 constructed response items were collected and used in the training set for the reported machine learning model (see Appendix 1 for a complete list of constructed response items). The items are characterized by five types: (i) aqueous proton-transfer reaction; (ii) acid–base reaction that can only be explained using the Lewis model; and why a compound can act as (iii) an acid, (iv) a base, or (v) amphoteric (see Figure 3.1 for an example for each prompt type).

The following prompt was provided for the aqueous proton transfer and Lewis mechanism prompts:

Part A: Describe in full what you think is happening on the molecular level for this reaction. Be sure to discuss the role of each reactant and intermediate.

Part B: Using a molecular level explanation, explain why this reaction occurs. Be sure to discuss why the reactants form the products shown.

(i) Aqueous proton transfer:
Consider the mechanism for the following acid–base reaction between water and hydrochloric acid to form hydronium and chloride ion.

(ii) Lewis mechanism:
Consider the mechanism for the reaction between ammonia and trifluoroborate to form the ammonia-trifluoroborate adduct.

(iii) Why acid?:
Explain why aluminum trichloride, AlCl₃, can act as an acid.

(iv) Why base?:
Explain why pyridine, C₅H₅N, can act as a base.

(v) Why amphoteric?:
Explain why ethanol, CH₃CH₂OH, can act as both an acid and a base.

Figure 3.1. Example of constructed-response items. For a comprehensive list of all prompts used in this study, see Appendix A.1

3.4.2 Data collection

Data were collected from six semesters (Fall 2017 through Fall 2020) of the first semester and one semester (Fall 2018) of the second semester of the year-long organic chemistry course as taught by three instructors at a large, research-intensive, public university in the southeastern
Data were collected in author DCR’s first semester course in Spring 2018, Fall 2018, Spring 2019, Fall 2019, Spring 2020, and Fall 2020. Data were collected in author KBF’s first semester course in Fall 2017, Fall 2018, Fall 2019, and Fall 2020. Data were collected in author JRR’s first semester course in Fall 2017, Fall 2019, and Spring 2020, and second semester course in Fall 2018. The textbook used between Fall 2017 and Spring 2019 was Solomons, Fryhle, and Synder’s *Organic Chemistry, 12th edn* (2016); the textbook used between Fall 2019 and Fall 2020 was Klein’s *Organic Chemistry, 3rd edn* (2017).

Constructed response items were given in a survey via Qualtrics. Participants received extra credit towards their examination grade for completing the assessment. Participants completed only one survey on acidity and basicity in the term. In total, 8520 responses from 15 different constructed response items (see Appendix A.1) were collected and used in the training set of the machine learning model between Fall 2017 and Spring 2020. An additional 2162 responses from 11 constructed response items with new mechanisms and compounds were collected for additional validation of the machine learning model in Fall 2020 (see Appendix A.2 for a complete list of constructed response items).

### 3.4.3 Development of classification scheme

Responses to the assessment items were classified by correct use or incorrect use/non-use of the Lewis acid–base model based on the terminology and ideas associated with the model. Responses were classified as *correct use* if the response had ideas about electrons and actions of those electrons: the transfer of electrons or a lone pair, electrons attacking, or about electron densities and partial charges. An action verb or an implied action was necessary for a response to be classified as *correct use*; mentioning the presence of electrons or lone pairs was not enough. We also include the Ingold–Robinson approach of using reactivity vernacular to describe electrophiles as Lewis acids and nucleophiles as Lewis bases (Robinson, 1932; Ingold, 1934), such that ideas about nucleophiles attacking are classified as *correct use*. For responses to the
aqueous proton-transfer reactions that generically explained how to interpret the curved-arrow formalism, *i.e.*, without mentioning the specific compounds in the prompt, were classified as *incorrect use/non-use* as the response did not satisfy the requirements of the prompt. The classification scheme for generic explanations was also applied to the reactions that can only be explained using the Lewis model. Additionally, responses to the *why acid, why base, and why amphoteric* prompts that simply mentioned that the compound was a Lewis acid or base without describing the actions of electrons were classified as *incorrect use/non-use*.

Responses classified as *correct use* may have solely used the Lewis model or a combination of models including the Arrhenius or Brønsted–Lowry models. However, all statements had to be correct for a given response to be classified as *correct use*. Representative examples of correct use responses using the example constructed-response items given in Figure 3.1 are provided in Table 3.1.

<table>
<thead>
<tr>
<th>Prompt type</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aqueous proton transfer</td>
<td>“The lone pair on the oxygen is forming a new bond with the hydrogen from the HCl. The chlorine then takes the electron pair from the hydrogen chloride bond and becomes a chlorine anion. This reaction occurs because HCl is a strong acid and water is a weak base. This causes the Cl to want to break off. This is because chlorine is a good leaving group.”</td>
</tr>
<tr>
<td>Lewis mechanism</td>
<td>“A Lewis acid–base reaction is occurring creating the ammonia-trifluoroborane adduct. The lone pair on the nitrogen donates its pair of electrons to the BF$_3$ making the boron now have a negative charge and the nitrogen now have +1 formal charge. BF$_3$ is a very good Lewis acid as its valence shell only contains six electrons and does not have a complete octet. Nitrogen is considered the Lewis Base as it donates its pair of electrons. The base forms a covalent bond with the acid making the acid-base adduct.”</td>
</tr>
<tr>
<td>Why acid?</td>
<td>“There is an empty p-orbital on the aluminum atom, which can act as an acid and accept an electron pair.”</td>
</tr>
<tr>
<td>Why base?</td>
<td>“A base is considered an electron pair donor. The nitrogen atom in pyridine has one lone pair, in which it can donate this pair of electrons; therefore, this makes it a base.”</td>
</tr>
<tr>
<td>Why amphoteric?</td>
<td>“Ethanol can act as both an acid and a base because it can accept and donate lone pairs. When it acts as an acid, the hydroxyl proton accepts an electron pair from the strong base. When it acts as a base, it donates an electron pair to a strong acid.”</td>
</tr>
</tbody>
</table>
In the first sample of 8,520 responses, author BJY classified all responses. Then, author JRR independently classified 250 randomly selected responses ($n = 50$ of each prompt type; 3% overall). Authors BJY and JRR originally agreed on 83% ($n = 208$) of the items; after discussing disagreements, classifications were changed for 15% ($n = 38$) of the items with a 98% final agreement. Author BJY then reevaluated a random sample of 1,500 responses ($n = 300$ of each prompt type; 17.6% overall) in light of the conversation. A discussion of disagreements is presented in the Limitations section.

In the additional set of 2,162 responses, author BJY again classified all responses. Author JRR independently classified 150 randomly selected responses ($n = 10$ of each prompt type; 7% overall). Authors BJY and JRR originally agreed on 93% ($n = 139$) of the items; after discussing disagreements, classifications were changed for 5% ($n = 8$) of the items with a 98% final agreement. No reevaluation of the set of responses was conducted as author BJY did not change any classifications during the conversation of disagreements.

A summary of the human classifications for all responses by prompt type and training/validation set is given in Table 3.2.

### 3.4.4 Development of machine learning model

There are four main components of our machine learning model development process: (i) obtaining/collecting data, (ii) data preprocessing and feature extraction, (iii) model training, and (iv) model evaluation (see Figure 3.2 for an overview of the process). After data are obtained, data are preprocessed and “cleaned,” for example, so that only words with chemical meaning remain. Remaining words, also referred to as the bag of words in text analysis methodologies, represent features that are extracted into a matrix and used in training the predictive model. The predictive model is evaluated using three approaches.

Initial data ($n = 8,520$) were split into two sets: training and validation. The training set consisted of a random set of 419 responses from each of the five prompts (i.e., a total of 2,095
Table 3.2. Distribution of human-classified correct and incorrect use/non-use of the Lewis acid–base model

<table>
<thead>
<tr>
<th>Set</th>
<th>Prompt type</th>
<th>N</th>
<th>Correct use (%)</th>
<th>Incorrect use/ non-use (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training/ cross-validation set</td>
<td>Aqueous proton transfer</td>
<td>419</td>
<td>292 (70)</td>
<td>127 (30)</td>
</tr>
<tr>
<td></td>
<td>Lewis mechanism</td>
<td>419</td>
<td>352 (84)</td>
<td>67 (16)</td>
</tr>
<tr>
<td></td>
<td>Why acid?</td>
<td>419</td>
<td>243 (58)</td>
<td>176 (42)</td>
</tr>
<tr>
<td></td>
<td>Why base?</td>
<td>419</td>
<td>231 (55)</td>
<td>188 (45)</td>
</tr>
<tr>
<td></td>
<td>Why amphoteric?</td>
<td>419</td>
<td>198 (47)</td>
<td>221 (53)</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>2,095</td>
<td>1,316 (63)</td>
<td>779 (37)</td>
</tr>
<tr>
<td>Stratified split-validation set</td>
<td>Aqueous proton transfer</td>
<td>100</td>
<td>70 (70)</td>
<td>30 (30)</td>
</tr>
<tr>
<td></td>
<td>Lewis mechanism</td>
<td>100</td>
<td>88 (88)</td>
<td>12 (12)</td>
</tr>
<tr>
<td></td>
<td>Why acid?</td>
<td>100</td>
<td>51 (51)</td>
<td>49 (49)</td>
</tr>
<tr>
<td></td>
<td>Why base?</td>
<td>100</td>
<td>49 (49)</td>
<td>51 (51)</td>
</tr>
<tr>
<td></td>
<td>Why amphoteric?</td>
<td>100</td>
<td>48 (48)</td>
<td>52 (52)</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>500</td>
<td>306 (70)</td>
<td>194 (30)</td>
</tr>
<tr>
<td>Remaining split-validation set</td>
<td>Aqueous proton transfer</td>
<td>3,146</td>
<td>2,289 (73)</td>
<td>857 (27)</td>
</tr>
<tr>
<td></td>
<td>Lewis mechanism</td>
<td>100</td>
<td>88 (88)</td>
<td>12 (12)</td>
</tr>
<tr>
<td></td>
<td>Why acid?</td>
<td>990</td>
<td>547 (55)</td>
<td>443 (45)</td>
</tr>
<tr>
<td></td>
<td>Why base?</td>
<td>1,005</td>
<td>526 (52)</td>
<td>479 (48)</td>
</tr>
<tr>
<td></td>
<td>Why amphoteric?</td>
<td>1,184</td>
<td>556 (47)</td>
<td>628 (53)</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>6,425</td>
<td>4,006 (62)</td>
<td>2,419 (38)</td>
</tr>
</tbody>
</table>

responses). The minimum number of responses of each prompt type was 519; therefore, 419 responses were chosen for the training set such that the remaining 100 responses could be set aside for validation. A stratified data set for predictive model building was chosen as to prevent model building from being heavily influenced by one prompt type. We refer to the general set of responses as the data corpus. All machine learning work was completed in RStudio version 1.2.5033 (R Core Team, 2019).

The training corpus first underwent data preprocessing. Preprocessing first involved a function to convert all characters in the corpus to lowercase (Kwartler, 2017), then all non-alphanumeric, special characters, and punctuation are removed using the ‘tm’ and ‘qdap’ packages in R (Feinerer et al., 2008; Rinker, 2020). Stopwords, commonly used words in the English language that usually provide little meaning (e.g., articles), are then removed using the ‘tm’ package (Feinerer et al., 2008). Additionally, a curated dictionary of 2,413 custom stopwords was created and used to remove words; this custom dictionary was built by authors BJJ and JRR by compiling words without general meaning concerning the use of acid–base models (e.g.,
specific names of chemical species) or words that do not specifically describe reactions.

Misspelled words were defined in a list of patterns that were substituted by corresponding replacements. This processing is needed as standard text analysis libraries do not recognize technical and discipline-specific vocabulary (Urban-Lurain et al., 2009; Dood et al., 2020a) such as the chemistry words in our data corpus. For example, misspelled words such as “nuclaeophilic” are replaced with the correct spelling, “nucleophilic.” Many studies skip this step;
however, misspellings are noted as common error sources in human–computer score disagreements (e.g., Ha et al., 2011; Moharreri et al., 2014); thus, spending time to construct a database of commonly misspelled words and their many variations, as suggested by Ha and Nehm (2016), results in higher predictive model accuracy.

A process called lemmatization was used for text/word normalization. In lemmatization, inflected words are reduced such that the root word is the canonical (dictionary) form in the English language; it is usually coupled with part-of-speech tagging which is a process in which words are assigned part of speech tags associated with the language of the corpus (e.g., “was” becomes “be”). We chose singular verbs to lemmatize; for example, “attack,” “attacked,” and “attacking” all become “attacks.” Additionally, synonyms in a chemical context were grouped together; patterns with lower instance counts (e.g., “cleaves,” “disconnects,” “lyses,” “severs,” “splits,” “tears”) were replaced by the replacement (e.g., “breaks”) that was more common. A total of 1625 words to be replaced were included in the dictionary to account for misspelled words and lemmatization. This process was conducted the ‘qdap’ package in R (Rinker, 2020).

The final corpus preprocessing step was to remove leading, trailing, and excess white spaces. The ‘tm’ package was used to remove these spaces (Feinerer et al., 2008).

Following preprocessing, the next step in machine learning model building is feature extraction; this involves converting the remaining words in the data corpus (i.e., the bag of words) into independent features. The processed data corpus consists of 257 unique words called unigrams (i.e., instance of single words); Lintean et al. (2012) found that unigrams performed the best across a number of machine learning text analysis algorithms. We also considered and tested other n-grams, such as bigrams (i.e., pair of consecutive words), and a combination of both unigrams and bigrams. However, the use of unigrams was found to give the best model performance metrics (see discussion of performance metrics in Results and discussion). Our 257 unigrams were parsed into a document-term matrix, or the pattern of the presence or absence of the terms, with individual student written responses (i.e., documents) representing the rows and
unigrams (i.e., terms) representing the columns in the matrix. The document-term matrix was weighted using term frequency, i.e., calculated as the number of times term $t$ appears in the document (Kwartler, 2017). We, as well, tested other feature weightings such as term frequency-inverse document frequency which weights a feature as a product of term frequency and inverse term frequency (i.e., the log of the total number of documents divided by the number of document that term $t$ appears; Kwartler, 2017); term frequency outperformed term frequency-inverse document frequency.

Machine learning algorithms use a document-term matrix to generate a predictive model in a process called *model construction*. While many machine learning algorithms exist (cf., Ramasubramanian and Singh, 2019, for an overview of possible algorithms), with several recent studies using an ensemble of algorithms (cf., Kaplan et al., 2014; Moharreri et al., 2014; Noyes et al., 2020), we chose a single algorithm for optimizing predictive performance due to a lack of interpretability of the many predictive model outputs when an ensemble of methods is used (cf., Sagi and Rokach, 2018). In other words, it is prohibitively difficult to determine contributing error and limitations, such as false positives and negatives, of each algorithm within ensemble-based models.

In this study, we use *support vector machines* (SVM) with a linear basis function kernel (Cortes and Vapnik, 1995) algorithm for our classification. SVM is reported as robust when compared to other algorithms for text analysis classifications (Ha et al., 2011; Nehm et al., 2012; Kim et al., 2017) and in a meta-analysis, has shown to have substantial machine-human score agreement (Zhai et al., 2021). Other algorithms were also tested: regularization (ridge regression, least absolute shrinkage and selection operator (LASSO), and elastic net), Bayesian (naïve Bayes), ensemble (random forest), and other instance-based methods (SVM with radial and polynomial basis kernel functions). A baseline model (naïve Bayes) was compared with the performance of the linear SVM classifier. For a twice repeated, two-fold cross-validation (described below), the naïve Bayes classifier performed poorly (accuracy = 56.18%, Cohen’s
kappa = 0.22) compared to linear SVM (accuracy = 88.93%, Cohen’s kappa = 0.76). Linear SVM performed best of all the algorithms tested and was therefore used for model training.

Support vector machines with a linear basis function kernel was used for model training in our analyses. In SVM, data are first mapped in multidimensional space (Cortes and Vapnik, 1995). In this process, for linear SVM, the $C$ or cost penalty parameter is optimized; this hyperparameter regulates how much the algorithm should avoid misclassifying the training data by looking for the optimal hyperplane to classify the data (Cortes and Vapnik, 1995; Joachims, 2002; Gaspar et al., 2012). Then, linear SVM calculates the optimal hyperplane in which it attempts to maximize the margin (i.e., greatest distance) between support vectors (i.e., data points nearest to the hyperplane) between the two classes of data (i.e., use and non-use of the Lewis model); in other words, SVM tries to find the hyperplane that best discriminates the data (Cortes and Vapnik, 1995). For our reported predictive model, hyperparameter $C = 0.0055$. The ‘caret’ package in R used was for model training (Kuhn, 2008).

To validate the predictive model, three validation methods are performed: (i) cross-validation, (ii) split-validation (also known as holdout validation), and (iii) external validation. First, cross-validation: this process involves breaking up the data into $k$ groups that are repeatedly shuffled with model construction performed on $k-1$ group(s) with the last group used for model cross-validation. We used a 2-fold cross-validation that was repeated twice, i.e., the data was equally split into half, trained on one half, then tested on the other; this process was repeated once more. While 5- to 10-fold cross-validation is considered standard (Rodríguez et al., 2010), a 2-fold cross-validation with a smaller number of repeats is considered acceptable when there are small samples as a result of the $k$-fold division (Wong and Yeh, 2020). Additionally, we did not find better performance metrics with an increase in the number of folds or repeats; a greater number of folds and repeats increases computation costs.

Second, split-validation was performed; the split validation used a stratified set and a remaining set both comprising of the remaining responses that were not used in the training set.
Note that the split-validation data is different from that of the cross-validation data as the cross-validation data was constructed from the training data. The stratified set was assembled from a random set of 100 responses from each of the five prompt types (i.e., a total of 500 responses); it is inclusive of the remaining set and was assembled from all of the remaining data that was not used for model training and construction ($N = 6,425$ total responses). We chose 100 responses from each prompt type as to mimic a typical large-enrollment organic chemistry course. Finally, external validation was performed; the machine learning model was evaluated using a data set of responses collected after the model was constructed. This set consisted of an additional 11 items that included reactions and compounds that were not used in the 15 items used in the construction of the predictive model; the external validation set consisted of 2,162 responses. The 11 prompts for external validation are reported in Appendix A.2. The external validation corpus first underwent the same data preprocessing as the training data corpus. In feature extraction, the features in the validation corpus were matched to the 257 features identified in the training corpus using a match matrix function as described by Kwartler (2017).

3.5 Results and discussion

Consistent with some studies (e.g., Cooper et al., 2016; Galloway et al., 2017; Bhattacharyya and Harris, 2018; Crandell et al., 2019; Dood et al., 2020a) and in contrast to other studies (e.g., Cartrette and Mayo, 2011; Schmidt-McCormack et al., 2019; Petterson et al., 2020; Watts et al., 2020), a majority of responses (62.4%) in our sample used language to describe the movement of electrons in mechanistic prompts. Such language is consistent with the Lewis acid–base model, and further evidence that some students do use correct language while invoking the Lewis model when explaining acidity and basicity.
3.5.1 Performance metrics

Performance of our machine learning model, i.e., congruence of human and computer scoring, is evaluated using several metrics. Each scoring prediction can be one of four outcomes: true positive (TP; human- and computer-classified as correct Lewis use), true negative (TN; human- and computer-classified as incorrect Lewis use/non-use), false positive (FP; human-classified as incorrect/non-use and computer-classified as correct Lewis use), or false negative (FN; human-classified as correct and computer-classified as incorrect Lewis use/non-use). A confusion matrix (see Figure 3.3) shows how each prediction outcome is made through the combination of the actual (human-classified) and predicted (computer-classified) scores. False positives (FP) are analogous to Type I errors and false negatives (FN) analogous to Type II errors.

<table>
<thead>
<tr>
<th>actual (human-classified)</th>
<th>incorrect use/ non-use</th>
<th>correct use</th>
</tr>
</thead>
<tbody>
<tr>
<td>predicted (computer-classified)</td>
<td>incorrect use/ non-use</td>
<td>TN (true negative)</td>
</tr>
<tr>
<td></td>
<td>correct use</td>
<td>FP (false positive)</td>
</tr>
</tbody>
</table>

Figure 3.3. Confusion matrix showing the outcomes of predicted vs. actual classifications

Three other metrics are Cohen’s kappa, percent accuracy, and the $F_1$ score. Cohen’s kappa is a statistic for interpreting interrater reliability testing that accounts for the probability that raters agree due to chance (Cohen, 1960). However, we can assume that both human and computer classifications are purposeful and informed; therefore, percent accuracy is a reliable measure (McHugh, 2012). Accuracy (Equation 3.1) is calculated as
Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (Eq. 3.1)

where the sum of the number of true positives and true negatives are divided by the sample size.

Cohen's kappa and percent accuracy are typical measures reported for models developed to evaluate written responses to assessment items. However, these metrics are most accurate with balanced data sets (i.e., same number of each classification; Kwartler, 2017). For example, if there were near equal numbers of students who correctly used and did not correctly use the Lewis acid–base model in their responses to the assessment items. Our data is heavily unbalanced for most of the individual prompt types and is skewed toward correct use of the Lewis model for the overall data set (Table 3.2). Due to this imbalance, a more accurate model performance metric is needed: the \( F_1 \) score (Kwartler, 2017). The \( F_1 \) score (Equation 3.2) is a classification model performance metric that attempts to balance precision (Equation 3.3) and recall (Equation 3.4).

\[
F_1 = 2 \left( \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \right) = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (Eq. 3.2)
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (Eq. 3.3)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (Eq. 3.4)
\]

Precision (or positive predictive value) is the number of responses correctly classified as positive (TP) out of all the responses that are computer-classified positive (TP + FP) and recall (or sensitivity) is the number of responses correctly classified as positive (TP) out of all the responses that are human-classified positive (TP + FN). Precision details how precise the model
is, or how many of the responses that the computer-classified as correct Lewis use are actually correct as classified by the computer. Recall details how sensitive the model is, or how many of the responses that were human-classified as correct Lewis are classified as correct by the computer. Mathematically, the $F_1$ score is the harmonic mean of precision and recall; false positives and false negatives are balanced. $F_1$ scores range from 0 to 1 with value of 1 indicating perfect precision and recall.

While $F_1$ is commonly used as a performance metric, one limitation is that $F_1$ is independent from TN (see Equation 3.2). In unbalanced cases, such as ours, $F_1$ can be misleading as it does not consider the proportion of each class (i.e., TP, FP, TN, and FN) in the confusion matrix (Chicco and Jurman, 2020). An alternative metric is the Matthews correlation coefficient (MCC; Equation 3.5), which is advantageous as its calculation is unaffected by unbalanced datasets (Matthews, 1975; Baldi et al., 2000; Chicco and Jurman, 2020).

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$ (Eq. 3.5)

MCC values range from $-1$ to $+1$ with a value of $+1$ indicating perfect classification and a value of zero equivalent to random guessing. While MCC has been shown to be more reliable in evaluating a two-class confusion matrix (Chicco and Jurman, 2020; Chicco et al., 2021), we report Cohen’s kappa, accuracy, $F_1$, and MCC for transparency.

### 3.5.2 Evaluation of machine learning algorithm

Evaluation of the predictive model is conducted using three validation methods: cross-validation, split-validation (stratified and remaining), and external validation. In addition to the confusion matrix classes (i.e., TP, FN, TN, and FP), we also report Cohen’s kappa, accuracy, $F_1$, and MCC for each prompt type and overall data set for the different validation methods.
Overall accuracy of our twice-repeated, 2-fold cross-validation (Table 3.3) is 88.93% with individual prompt-type accuracy ranging from 83.05% to 93.08%. Overall $F_1$ score is 0.91 with individual prompt $F_1$ scores ranging from 0.86 to 0.96; overall MCC value is 0.75 with individual prompt MCC values ranging from 0.69 to 0.81. These three metrics are in good agreement with higher accuracies having larger $F_1$ scores and MCC values. Varying prompt accuracies, $F_1$ scores, and MCC values are indicative that the predictive model performs better for certain prompt types (e.g., reaction mechanism that can be only explained with the Lewis model) over others (e.g., why is the compound a base). There are relatively low false negative rates (i.e., computer-scored as correct use of the Lewis model where a human classifier would have scored the response as incorrect use/non-use), with moderate rates for the false positives.

Table 3.3. Predictive model results on the cross-validation set

<table>
<thead>
<tr>
<th>Prompt type</th>
<th>$N$</th>
<th>$\kappa$</th>
<th>Accuracy (%)</th>
<th>$F_1$</th>
<th>MCC</th>
<th>TP (%)</th>
<th>FN (%)</th>
<th>TN (%)</th>
<th>FP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aqueous proton transfer</td>
<td>419</td>
<td>0.74</td>
<td>89.26</td>
<td>0.92</td>
<td>0.73</td>
<td>93.84</td>
<td>6.16</td>
<td>78.73</td>
<td>21.27</td>
</tr>
<tr>
<td>Lewis mechanism</td>
<td>419</td>
<td>0.72</td>
<td>93.08</td>
<td>0.96</td>
<td>0.69</td>
<td>97.73</td>
<td>2.27</td>
<td>68.66</td>
<td>31.34</td>
</tr>
<tr>
<td>Why acid?</td>
<td>419</td>
<td>0.81</td>
<td>90.93</td>
<td>0.92</td>
<td>0.81</td>
<td>93.00</td>
<td>7.00</td>
<td>88.07</td>
<td>11.93</td>
</tr>
<tr>
<td>Why base?</td>
<td>419</td>
<td>0.65</td>
<td>83.05</td>
<td>0.86</td>
<td>0.65</td>
<td>92.64</td>
<td>7.36</td>
<td>71.28</td>
<td>28.72</td>
</tr>
<tr>
<td>Why amphoteric?</td>
<td>419</td>
<td>0.77</td>
<td>88.31</td>
<td>0.92</td>
<td>0.77</td>
<td>91.92</td>
<td>8.08</td>
<td>85.07</td>
<td>14.93</td>
</tr>
<tr>
<td>Overall</td>
<td>2,095</td>
<td>0.76</td>
<td>88.93</td>
<td>0.91</td>
<td>0.75</td>
<td>94.22</td>
<td>5.78</td>
<td>79.97</td>
<td>20.03</td>
</tr>
</tbody>
</table>

It can be noted that the accuracy for positive instances (TP vs. FN) is much greater than the accuracy of negative instances (TN vs. FP). This is likely due to the imbalance in the training data set of correct Lewis use instances (63%) and incorrect Lewis use/non-use instances (37%; see Table 3.2). This means that the training set is trained more heavily on positive instances, causing there to be a discrepancy between accuracy of correctly predicting positive instances and correctly predicting negative instances. This discrepancy is even more pronounced when looking specifically at the Lewis mechanism prompt type where the model is 97.73% accurate for positive instances and 68.66% accurate for negative instances (see Table 3.3).

Overall, this 2-fold cross-validation demonstrates that there are varying metrics by prompt type; however, the predictive model holds for all of these prompt types when considered as a
whole and therefore there is evidence that the predictive model can be generalized for all these different prompt types.

A predictive model using lexical analysis and binomial logistic regression model predicting Lewis acid–base use, including correct and incorrect use/non-use, had an accuracy of 82% (Dood et al., 2018). The predictive model was further improved with new data to 86% accuracy (Dood et al., 2019); however, this model is only applicable to aqueous proton transfer reactions. Our results are more accurate, in general, to prior work from Dood et al. (2018, 2019). When our results are compared to other computer-assisted predictive scoring models, the predictive model we report is just as or more accurate than those predictive models developed for single assessment items. Thus, these initial findings and comparisons suggest that our generalized predictive model meets current/ reported accuracy standards (Zhai et al., 2020).

The stratified split-validation set (Table 4) aims to mimic the class size of a large-enrollment organic chemistry course. This validation set has comparable accuracies, $F_1$ scores, and MCC to the cross-validation set. The prompt asking “why a compound is a base” is the worst performing prompt type and the mechanistic prompt that can be only explained with the Lewis model is the best performing when considering accuracy and $F_1$. However, in this stratified split-validation set and also in the remaining split-validation set, there are only instances for the Lewis mechanism prompt type positive instances; thus, the accuracy for negative instances is lower than the cross-validation set and is reflected in the lower MCC value. We posit that if the Lewis mechanism prompt type had a larger sample size and a larger number of negative instances, we would see a smaller discrepancy here.

<table>
<thead>
<tr>
<th>Prompt type</th>
<th>N</th>
<th>$\kappa$</th>
<th>Accuracy (%)</th>
<th>$F_1$</th>
<th>MCC</th>
<th>TP (%)</th>
<th>FN (%)</th>
<th>TN (%)</th>
<th>FP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aqueous proton transfer</td>
<td>100</td>
<td>0.73</td>
<td>89.0</td>
<td>0.92</td>
<td>0.71</td>
<td>95.71</td>
<td>4.29</td>
<td>73.33</td>
<td>26.67</td>
</tr>
<tr>
<td>Lewis mechanism</td>
<td>100</td>
<td>0.45</td>
<td>90.0</td>
<td>0.94</td>
<td>0.46</td>
<td>96.59</td>
<td>3.41</td>
<td>41.67</td>
<td>58.33</td>
</tr>
<tr>
<td>Why acid?</td>
<td>100</td>
<td>0.70</td>
<td>85.0</td>
<td>0.86</td>
<td>0.70</td>
<td>88.24</td>
<td>11.76</td>
<td>81.63</td>
<td>18.37</td>
</tr>
<tr>
<td>Why base?</td>
<td>100</td>
<td>0.50</td>
<td>75.0</td>
<td>0.78</td>
<td>0.53</td>
<td>89.80</td>
<td>10.20</td>
<td>60.78</td>
<td>39.22</td>
</tr>
<tr>
<td>Why amphoteric?</td>
<td>100</td>
<td>0.80</td>
<td>90.0</td>
<td>0.90</td>
<td>0.79</td>
<td>88.24</td>
<td>11.76</td>
<td>90.38</td>
<td>9.62</td>
</tr>
<tr>
<td>Overall</td>
<td>500</td>
<td>0.69</td>
<td>85.8</td>
<td>0.89</td>
<td>0.69</td>
<td>92.81</td>
<td>7.19</td>
<td>74.74</td>
<td>25.26</td>
</tr>
</tbody>
</table>
As with the smaller sample sizes, the errors of these predictions increase, explaining the larger percentages of false positives and false negatives in addition to the slightly lower $F_1$ scores and MCC values; although, these accuracy metrics are comparable to the cross-validation set. For the stratified split-validation set, accuracies are greater than 75% and $F_1$ scores are greater than 0.78 which demonstrate that sufficient accuracies can be obtained with sample sizes of 100. While $F_1$ scores indicate that the predictive model performs well in correctly classifying positive cases (i.e., human classifications are correct Lewis use), the lower MCC values for mechanism prompts that can only be explained by the Lewis model and prompts asking why a compound is a base indicate that the predictive model has a difficult time classifying negative cases (i.e., human classifications are incorrect Lewis use/non-use).

The penultimate validation test was to explore model performance metrics for all data not used in the training data set. The remaining split-validation set (Table 3.5) consists of all the remaining data not used in the training (and cross-validation) set. This validation set allows for appraisal of the predictive model with a large sample size, indicative of overall predictive performance. Accuracies are greater than 80% with $F_1$ scores above 0.80; MCC values are generally above 0.60, with exceptions of the Lewis mechanism prompts (as previously discussed). A lower rate of false negatives is observed with comparably higher rates of false positives. These results suggest that the number of computer-classified correct use may be slightly inflated for this large corpus. For example, in a class of 200 students, if the model predicts that the number of correct use classifications is 175, the actual value may be slightly lower due to higher false positive than false negative rates. Overall, the predictive model performs well for each of the prompt types.

<table>
<thead>
<tr>
<th>Prompt type</th>
<th>N</th>
<th>$\kappa$</th>
<th>Accuracy (%)</th>
<th>$F_1$</th>
<th>MCC</th>
<th>TP (%)</th>
<th>FN (%)</th>
<th>TN (%)</th>
<th>FP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aqueous proton transfer</td>
<td>3,146</td>
<td>0.62</td>
<td>84.01</td>
<td>0.89</td>
<td>0.65</td>
<td>85.71</td>
<td>14.29</td>
<td>79.46</td>
<td>20.54</td>
</tr>
<tr>
<td>Lewis mechanism</td>
<td>100</td>
<td>0.45</td>
<td>90.00</td>
<td>0.94</td>
<td>0.46</td>
<td>96.59</td>
<td>3.41</td>
<td>41.67</td>
<td>58.33</td>
</tr>
<tr>
<td>Why acid?</td>
<td>990</td>
<td>0.73</td>
<td>86.67</td>
<td>0.88</td>
<td>0.73</td>
<td>91.59</td>
<td>8.41</td>
<td>80.59</td>
<td>19.41</td>
</tr>
<tr>
<td>Why base?</td>
<td>1,005</td>
<td>0.59</td>
<td>79.60</td>
<td>0.82</td>
<td>0.60</td>
<td>90.68</td>
<td>9.32</td>
<td>67.43</td>
<td>32.57</td>
</tr>
<tr>
<td>Why amphoteric?</td>
<td>1,184</td>
<td>0.75</td>
<td>87.67</td>
<td>0.87</td>
<td>0.76</td>
<td>90.47</td>
<td>9.53</td>
<td>85.19</td>
<td>14.81</td>
</tr>
<tr>
<td>Overall</td>
<td>6,425</td>
<td>0.67</td>
<td>84.50</td>
<td>0.88</td>
<td>0.67</td>
<td>88.07</td>
<td>11.93</td>
<td>78.59</td>
<td>21.41</td>
</tr>
</tbody>
</table>
The external validation set (Table 3.6) allows us to evaluate performance of the predictive model on a set of new data that includes a variety of new prompts and a new prompt type: specifically, a non-aqueous proton transfer mechanism. We recognized when planning to collect the new, external validation data that the proton-transfer reactions from which the predictive model was developed were all aqueous; thus, we included a non-aqueous proton transfer in the external validation set to further evaluate the generalizability of our predictive model. A summary of the human classifications for the external validation set is given in Table 3.7. This new prompt type has an accuracy of 91.3%, $F_1$ score of 0.95, and MCC of 0.73; thus, we can conclude that the predictive model performs well when used to evaluate these new data. Additionally, not only do we see that kappa, accuracy, $F_1$ scores, and MCC values generally increase across all prompt types, but we find that false negative and false positive rates also decrease. The high true positive rate indicates that the predictive model has high recall, being able to correctly classify a response as correct use of the Lewis model out of all the possible correct classifications given by a human classifier. These external validation results show analogous or better metrics when compared to the cross-validation and split-validations, and to other studies that use machine learning techniques (e.g., Dood et al., 2018, 2020a; Noyes et al., 2020; cf., Zhai et al., 2020). Additionally, the level of prediction accomplished by our model exceeds the 70% accuracy recommendation for use in formative assessments (cf., Haudek et al., 2012; Nehm et al., 2012; Prevost et al., 2016) and is generally within range for accepted measures for summative assessments (cf., Williamson et al., 2012). Therefore, we conclude that an accurate, generalizable, predictive model using machine learning techniques for correct use of the Lewis acid–base model was developed. However, despite this level of accuracy, we reiterate that use of this generalized predictive model should only be used with formative assessments.
Table 3.6. Predictive model results on the external validation set

<table>
<thead>
<tr>
<th>Prompt type</th>
<th>N</th>
<th>(\kappa)</th>
<th>Accuracy (%)</th>
<th>(F_1)</th>
<th>MCC</th>
<th>TP (%)</th>
<th>FN (%)</th>
<th>TN (%)</th>
<th>FP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-aqueous proton transfer</td>
<td>715</td>
<td>0.74</td>
<td>91.33</td>
<td>0.95</td>
<td>0.73</td>
<td>96.01</td>
<td>3.99</td>
<td>75.46</td>
<td>24.54</td>
</tr>
<tr>
<td>Lewis mechanism</td>
<td>716</td>
<td>0.62</td>
<td>94.41</td>
<td>0.97</td>
<td>0.61</td>
<td>97.86</td>
<td>2.14</td>
<td>58.73</td>
<td>41.27</td>
</tr>
<tr>
<td>Why acid?</td>
<td>294</td>
<td>0.81</td>
<td>89.61</td>
<td>0.93</td>
<td>0.79</td>
<td>97.81</td>
<td>2.19</td>
<td>80.18</td>
<td>19.82</td>
</tr>
<tr>
<td>Why base?</td>
<td>292</td>
<td>0.86</td>
<td>93.15</td>
<td>0.94</td>
<td>0.85</td>
<td>94.94</td>
<td>5.06</td>
<td>90.35</td>
<td>9.65</td>
</tr>
<tr>
<td>Why amphoteric?</td>
<td>145</td>
<td>0.88</td>
<td>93.79</td>
<td>0.93</td>
<td>0.88</td>
<td>96.92</td>
<td>3.08</td>
<td>91.25</td>
<td>8.75</td>
</tr>
<tr>
<td>Overall</td>
<td>2,162</td>
<td>0.80</td>
<td>92.74</td>
<td>0.95</td>
<td>0.78</td>
<td>96.87</td>
<td>3.13</td>
<td>80.04</td>
<td>19.96</td>
</tr>
</tbody>
</table>

Table 3.7. Distribution of human-classified correct and incorrect use/non-use of the Lewis acid–base model for the external validation set

<table>
<thead>
<tr>
<th>Prompt type</th>
<th>N</th>
<th>Correct use (%)</th>
<th>Incorrect use/non-use (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-aqueous proton transfer</td>
<td>715</td>
<td>552 (77)</td>
<td>163 (23)</td>
</tr>
<tr>
<td>Lewis mechanism</td>
<td>716</td>
<td>653 (91)</td>
<td>63 (9)</td>
</tr>
<tr>
<td>Why acid?</td>
<td>294</td>
<td>183 (62)</td>
<td>111 (38)</td>
</tr>
<tr>
<td>Why base?</td>
<td>292</td>
<td>178 (61)</td>
<td>114 (39)</td>
</tr>
<tr>
<td>Why amphoteric?</td>
<td>145</td>
<td>65 (45)</td>
<td>80 (55)</td>
</tr>
<tr>
<td>Overall</td>
<td>2,162</td>
<td>1,631 (75)</td>
<td>531 (25)</td>
</tr>
</tbody>
</table>

3.6 Limitations

The immediate findings of this research are constrained by several key limitations: (i) homogeneity of the sample, (ii) accuracy of the predictive scoring model, (iii) machine learning methodology, (iv) students “gaming” the system, (v) when the machine learning model fails, and (vi) use of mixed-models in a response. However, we do note that future work may minimize or nullify several of these limitations based on a broader use and evaluation of the predictive model.

3.6.1 Sample homogeneity

The open-ended written responses collected in this study were from students who took organic chemistry at a single institution with one of three instructors over the span of seven different semesters. We report no evidence for generalizing these constructed response item types and findings to other institutions and to other curricula. However, Ha et al. (2011) found that their machine learning model was able to, in most cases, accurately evaluate the degree of sophistication between biology majors and non-majors at two different institutions. This suggests that machine learning models could be utilized with student populations at different institutions; nonetheless, these models should be tested at multiple institution types (e.g., two-year colleges,
primarily-undergraduate institutions, research-intensive institutions) and with multiple curricula, such as the Chemistry, Life, the Universe and Everything curriculum (Cooper et al., 2019) or Mechanisms before Reactions curriculum (Flynn and Ogilvie, 2015).

3.6.2 Predictive model accuracy

The reported machine learning model has percent accuracies for the given data between 84.5% and 92.7%. While the predictive model is not 100% perfect, the model performs at similar percent accuracies and Cohen’s kappa levels compared to reported models for constructed response assessment items (Haudek et al., 2012; Prevost et al., 2012, 2016; Dood et al., 2018, 2020a; Noyes et al., 2020; Zhai et al., 2020) and is in line with the agreement between authors BJJ and JRR before discussion. As with the supermajority of those developing and disseminating similar models for text analysis of assessment items, we reiterate that the predictive models, ours included, should only serve as a method for evaluating formative assessment items.

3.6.3 Machine learning methodology

While there are many methodological approaches to machine learning model development, our choices could be construed as a limitation. For example, in feature extraction, we chose to use the term frequency weighting in the document-term matrix; that is, each time a term (or feature) \( t \) appears in a student’s response, it will weight that feature more in the matrix. While we found that this weighting method gave the best model metrics overall, students could “game” the system where their response would contain a of surplus key words relating to Lewis acid–base chemistry (e.g., “Lewis,” “electrons,” “nucleophilic,” “electrophilic). This would trigger the predictive model to assign a correct use classification for the response that otherwise would be assigned as incorrect use/non-use when holistically evaluated by a human classifier.
3.6.4 Model shortcomings

3.6.4.1 False positives

Two key limitations are false positive and false negatives, i.e., when the model gets it wrong. In this section we review instances when the model predicted correct Lewis acid–base use, while human-classification suggested not.

There are several reasons why a response could be a false positive. First, student’s use of the term “electrons” while identifying or discussing irrelevant features of a compound or reaction. For example, determining number of valence electrons or formal charge:

*Cyclohexanaminium can act as an acid because it can donate a proton and be left with a lone pair and 2 bonds. This would neutralize its charge since, 2 electrons + 3 bonds = 5, 5 valence electrons – 5 = 0.*

This specific example demonstrates that instances of terms such as, “electrons,” “bonds,” “lone,” and “pair,” terminology associated with the Lewis model, can lead the computer to predict correct Lewis acid–base model use.

Second, students mention “a lone pair” or “lone pairs” without an action verb, usually in the surface-level description of a compound. For example,

*Nitrogen has a lone pair and it is the weakest amine.*

In this specific instance, a structural feature of a compound is noted that is associated with correct Lewis acid–base model; however, the answer as a whole is insufficient to be classified as correct.

Additionally, responses may contain a broad example of the movement of electrons or the arrow-pushing formalism without any specific details. For example,

*The reactants of the mechanism are undergoing a reaction in which the bonds are broke and then reformed to create the two products on the right. The curved arrows represent where the exchange/attraction of the electrons will be moving to create the bonds. This reaction occurs because the reactants are less stable then [sic] what they would be as the products.*
We note that shorter responses without sufficient chemical terminology or longer responses without specific details may trigger our predictive model to give a false positive because the model is simply analyzing term frequencies.

3.6.4.2 False negatives

A false negative means that the model incorrectly classified a response as incorrect use/non-use. Our review of false negatives reveals that shorter responses may cause the model to give a false negative. Term frequencies in shorter responses are lower and therefore, in comparison with other responses, may result in an incorrect classification. For example, the following response was classified as incorrect use/non-use by the model but as correct use by a human classifier:

\[(CH_3)_2CHO^- \text{ can act as a base because it can donate electrons.}\]

At the other end of response length, term frequency can also play a role in longer responses when a student uses both Brønsted–Lowry and Lewis models together in a response. For example,

An amphoteric substance means it can act as both an acid and a base. Water is the most common example of this. Tert-butanol can also act as both an acid or a base. This is because the hydrogen atom attached to the oxygen can be donated making it a Brønsted–Lowry acid. However, the lone pairs on the oxygen can accept a proton (H\(^+\) ion) making another bond making it a Brønsted–Lowry base.

In this instance, greater attention given to the Brønsted–Lowry base model, including explicit naming of the model gave rise to a false negative for this response. We also found instances of this in response to one of the mechanism prompts:

Part A: Negatively charged ethanethiolate transfers electrons to the hydrogen atom of benzoic acid to form a single covalent bond. The electrons from the sigma bond between hydrogen and oxygen in benzoic acid gather around the oxygen atom, causing it to go from a
neutral to a negative formal charge. In this way, enthanethiol [sic] is formed which has a neutral formal charge and negatively charged benzoate is produced. Part B: This reaction occurs because according to the Brønsted-Lowry definition, ethanethiolate is a base or a proton acceptor while benzoic acid is an acid or a proton donor. When ethanethiolate [sic] accepts the H atom or proton it becomes the conjugate acid, ethanethiol, and when benzoic acid gives up its proton it becomes the conjugate base, benzoate.

3.6.4.3 Mixed acid–base model use

In addition to false negatives, mixed acid–base model use serves as a key focus of disagreement between human classifiers. For example, consider this response to a prompt about a proton-transfer mechanism:

The first molecule in the reactant side [methoxide] is serving as a base because it is accepting a proton, while the second molecule in the reactant side [propaninium] is acting as an acid because it is donating a proton. Since the second molecule is donating its H to the first molecule, the single bond transfers as a lone pair to nitrogen. On the product side, the first molecule is the conjugate acid while the second molecule is the conjugate base. On a molecular level, the negative charge means that the atom wants to form a bond while the positive charge wants to donate hydrogen.

This response heavily invokes the Brønsted–Lowry model with concepts about accepting and donating protons in addition to conjugate acids and bases. One human classifier in the interrater discussion classified this response as incorrect use/non-use due to the lack of Lewis model use with too much discussion using the Brønsted–Lowry model. However, the other human classifier argued that while this model does focus on the Brønsted–Lowry model, there are aspects within the response (“the single bond transfers as a lone pair”) that demonstrates correct use of the Lewis model. This response was ultimately classified as correct use and the predictive model also correctly classified the response as correct use.
Prior research has indicated that students hold unclear relationships between acid–base models in their mental models (Schmidt and Volke, 2003; Drechsler and Schmidt, 2005; Bhattacharyya, 2006). Additionally, students struggled to incorporate broader models, such as the Lewis model, into more specific models, such as the Brønsted–Lowry model (Cartrette and Mayo, 2011). While use of the Lewis model has been shown to increase student performance (Dood et al., 2018), it is unclear whether students that use mixed models clearly understand when each model is appropriate, based solely on their responses, or if they default to the most specific model (e.g., Brønsted–Lowry) that can explain a phenomenon over the broader model (e.g., Lewis). The scoring model developed in our research only predicts whether a student has correctly used the Lewis acid–base model in their written response and is limited in understanding if students can differentiate between using mixed models correctly.

3.6.4.4 Gaming the system

Students could potentially “game” the system. Gaming the predictive model would involve generating a response that contains a wide array of key features that relate to the Lewis acid–base model or generating a response with the same feature repeatedly (e.g., stating “electrons” five times). Additionally, students could copy or base their responses off a “good example response.” In any of these cases, the predictive model may assign a correct use classification. We argue, based on using the tool for formative assessment (little to no impact, or only positive impact on course grades), that gaming has limited consequences. For example, as with other reported work (Dood et al., 2019, 2020b), even students demonstrating the highest level of understanding receive further learning opportunities. When constructed-response items, such as those we described herein, are consistently used in courses, and benefit is demonstrated to the student (even when incorrect), we believe that gaming the system will have limited overall impact.
3.7 Implications

3.7.1 Implications for instructors

Instructors should select assessments that send clear and bold messages to students about what is important in the classroom (Holme et al., 2010). A growing number of chemical education researchers (Becker et al., 2016; Cooper et al., 2016; Finkenstaedt-Quinn et al., 2017; Bodé et al., 2019; Caspari and Graulich, 2019; Crandell et al., 2019), in addition to other researchers across the educational research community (Birenbaum and Tatsuoka, 1987; Scouller, 1998), have called for the use of open-ended assessments to facilitate deeper learning and provide diagnostic data for educators. For example, being able to describe how and why a phenomenon (e.g., a reaction mechanism) occurs is critical for efficacious scientific reasoning (Abrams et al., 2001; Cooper, 2015). In chemistry, researchers have advocated for students to answer “why?” in assessments (Goodwin, 2003; Cooper, 2015; Cooper et al., 2016; Stowe and Cooper, 2017; Caspari et al., 2018a; Underwood et al., 2018; Bodé et al., 2019; Caspari and Graulich, 2019; Crandell et al., 2019; Dood et al., 2020a). One type of assessment that instructors should be using in their courses are low-stakes, formative assessments to gauge understanding.

Written formative assessments can help students develop skills in explaining how and why reactions occur. Numerous studies have advocated for students to explain how and why to enrich their productive ideas about how reactions work (e.g., Becker et al., 2016; Cooper et al., 2016; Dood et al., 2018, 2020a; Crandell et al., 2019, 2020). Research has shown that targeted formative feedback allows for students to learn about their competency level, suggestions for improvement, and may positively affect students’ exam scores (Hattie and Timperley, 2007; Hedtrich and Graulich, 2018; Young et al., 2020). Written assessments can reveal students’ understanding or lack thereof; therefore, such assessments should serve as an approach for instructors to use to get students to think more deeply about scientific explanations.

The predictive model developed in this study is a practical, quick, and efficient way to formatively evaluate student understanding of the Lewis acid–base model. We have freely made
available the files along with a set of instructions necessary for instructors to conduct their own analyses (cf., Yik and Raker, 2021); files are to be used with R, a free statistical software environment (R Core Team, 2019). Written formative assessments can be easily scored using computer-scoring models, like the one developed in this study, to support just-in-time teaching in large enrollment courses (Novak et al., 1999; Prevost et al., 2013; Urban-Lurain et al., 2013, Prevost et al., 2016). For example, instructors may administer constructed response questions as homework assignments (with low point value for completion credit) and then utilize the predictive model to classify student responses. Quantitative results are nearly instantaneous, providing quick feedback to instructors and students; the R program outputs result in a spreadsheet with paired student predictive scores that can uploaded into learning management systems with little modification. If instructors intend to hand score responses in addition to using the predictive model, we suggest that hand scoring be conducted first to avoid anchoring bias (Sherif et al., 1958). In the classroom, student responses can be used to create clicker questions and/or as a starting point for classroom discussion. Another option is that instructors can use quantitative results to reshape lessons or homework activities to promote correct understanding and use of the Lewis acid–base model.

Information provided by the predictive model can allow instructors to provide additional resources to students to support learning. One method is to couple predictive models with corresponding topic-specific online tutorials; such tutorials have been shown to increase student understanding and achievement in organic chemistry courses (e.g., O’Sullivan and Hargaden, 2014; Richards-Babb et al., 2015; Dood et al., 2019, 2020b). Tutorial-based learning interventions can be utilized to facilitate better construction of explanations when paired with adaptive learning opportunities based on quick results from computer-assisted scoring. Furthermore, online learning tools can supplement learning. One such open educational resource tool is OrgChem101 (https://orgchem101.com), which contains modules on acid–base reactions, nomenclature, and
organic mechanisms; the latter two modules have been shown to increase students’ learning gains (Bodé et al., 2016; Carle et al., 2020).

3.7.2 Implications for researchers

Tools used to construct our predictive model are open-access and are available for researchers to develop other predictive models. Our machine learning model was developed using R, a free statistical software environment (R Core Team, 2019). By removing the limitation of financial burdens and barriers that other software may impose, researchers are able to more freely use and develop their own predictive models. We have also made available all of our files (cf., Yik and Raker, 2021) containing the R code, custom stopwords dictionary, and patterns and replacements; these tools can be used as a starting point to develop other predictive models. There has been limited exploration in the development of predictive models in chemistry (Haudek et al., 2012; Prevost et al., 2012; Dood et al., 2018, 2019, 2020a; Noyes et al., 2020), which leaves a multitude of different predictive models that can be built to evaluate students’ written responses to constructed-response items. For example, predictive models for other aspects of the postsecondary organic chemistry curriculum are ripe for future exploration; two examples of particular interest are understanding of nucleophiles and electrophiles (Anzovino and Bretz, 2015, 2016) and leaving groups (Popova and Bretz, 2018).

There is ample opportunity to evaluate student understanding of acids and bases in the postsecondary chemistry curriculum using predictive models. For example, in our study, we primarily evaluated first semester organic chemistry students’ understanding of acid–base reaction mechanisms and why compounds can act as an acid, base, or be amphoteric using the Lewis model. As acid–base models are generally first introduced in general chemistry (Paik, 2015), further research can evaluate the effectiveness of our predictive model in this setting. Additionally, our predictive model can be evaluated with students in other organic chemistry courses; Dood et al. (2019) found that students benefited from a tutorial to review Lewis acid–
base concepts at the beginning of the second-semester organic chemistry course. While the Brønsted–Lowry and Lewis models are considered in postsecondary general chemistry and organic chemistry courses, other models to describe acids and bases are introduced in upper-level courses (e.g., Raker et al., 2015). In inorganic chemistry, the Lux–Flood model is based on oxide ion donors and acceptors (Lux, 1939; Flood and Förland, 1947), the Usanovich model defines acids and bases as charged species donors and acceptors (Finston and Rychtman, 1982), and the concept of hard and soft acids and bases (HSABs) describes polarizable acids and bases as soft and nonpolarizable acids and bases as hard (Pearson, 1963). While each of these models have their own advantages and criticisms (Miessler et al., 2014), it may be beneficial for students to preferentially use and understand one of these models in particular scenarios; there are avenues to build predictive models for other acid–base contexts.

Predictive models have the potential to be a practical way to change the nature of formatively assessing student understanding. Researchers have studied reasoning in chemistry, for example: teleological (e.g., Wright, 1972; Talanquer, 2007; Caspari et al., 2018b), mechanistic (e.g., Bhattacharyya and Bodner, 2005; Ferguson and Bodner, 2008; Bhattacharyya, 2013; Galloway et al., 2017; Caspari et al., 2018a), causal (e.g., Ferguson and Bodner, 2008; Crandell et al., 2019, 2020), and causal mechanistic (e.g., Becker et al., 2016; Cooper et al., 2016; Bodé et al., 2019; Noyes and Cooper, 2019; Crandell et al., 2020). Studies have also been conducted on student reasoning through case comparisons between reaction mechanisms (e.g., Graulich and Schween, 2018; Bodé et al., 2019; Caspari and Graulich, 2019; Watts et al., 2021). Student understanding via argumentation has likewise been investigated (e.g., Moon et al., 2016, 2017, 2019; Pabuccu, 2019; Towns et al., 2019). Regardless of the mode of reasoning, these different routes offer researchers comparative and contrasting means to study student understanding. While predictive models have begun to consider classification or levels of reasoning (Haudek et al., 2012; Prevost et al., 2012; Dood et al., 2018, 2019, 2020a; Noyes et al., 2020), the question remains: can generalized predictive models be built for other chemistry concepts? The
development of future predictive models has to potential to expand work communicated in this study by using levels of reasoning and additionally analyses that instructors can use to further support student learning.

3.8 Conclusion

This study has shown that a generalized predictive model using machine learning techniques can be developed to accurately predict correct use of the Lewis acid–base model with overall accuracies ranging from 84.5% to 92.7%. We have demonstrated that this predictive model is applicable to a total of six different prompt types: what is happening and why for (i) an aqueous proton-transfer reaction, (ii) a non-aqueous proton-transfer reaction, (iii) a mechanism that can be explained by only the Lewis model, (iv) why is this compound an acid? (v) why is this compound a base? and (vi) why is this compound amphoteric? Our results suggest promising avenues for the development of machine learning-based scoring tools to efficiently and accurately evaluate student understanding of chemical concepts beyond acid–base chemistry. Additionally, our predictive model was built using open-access statistical software (i.e., R) and is freely accessible to be used by instructors as a formative assessment tool to predict students’ correct use of the Lewis acid–base model.

3.9 Conflicts of interest

There are no conflicts to declare.

3.10 Acknowledgments

We would like to thank all the students who participated in our study.
3.11 References


Chapter 4

Generalized rubric for level of explanation sophistication for nucleophiles in organic chemistry reaction mechanisms

4.1 Note to Reader

This chapter is a published manuscript in Chemistry Education Research and Practice. This chapter is reproduced from:


with permission from the Royal Society of Chemistry. Further copyright information can be found in Appendix E.2.

This work has co-authors. Amber J. Dood was a former graduate student at the University of South Florida who collected some data used in this study and helped in conceptualizing the rubric reported in this study. Stephanie J. H. Frost is a current graduate student at the University of South Florida who helped in conceptualizing the rubric reported in this study and provided assistance with interrater reliability. Daniel Cruz-Ramírez de Arellano and Kimberly B. Fields are organic chemistry instructors at the University of South Florida who allowed for data collection in their classes. Jeffrey R. Raker is the principal investigator for this project.
4.2 Abstract

Reaction mechanisms are central to organic chemistry and organic chemistry education. Assessing understanding of reaction mechanisms can be evaluated holistically, wherein the entire mechanism is considered; however, we assert that such an evaluation does not account for how learners variably understand mechanistic components (e.g., nucleophile, electrophile) or steps (e.g., nucleophilic attack, proton transfer). For example, a learner may have proficiency of proton transfer steps without sufficient proficiency of a step where a nucleophile and electrophile interact. Herein, we report the development of a generalized rubric to assess the level of explanation sophistication for nucleophiles in written explanations of organic chemistry reaction mechanisms from postsecondary courses. This rubric operationalizes and applies chemistry education research findings by articulating four hierarchical levels of explanation sophistication: absent, descriptive, foundational, and complex. We provide evidence for the utility of the rubric in an assortment of contexts: (a) stages of an organic chemistry course (i.e., first or second semester), (b) across nucleophile and reaction types, and (c) across prompt variations. We, as well, present a case study detailing how this rubric could be applied in a course to collect assessment data to inform learning and instruction. Our results demonstrate the practical implementation of this rubric to assess understanding of nucleophiles and offer avenues for establishing rubrics for additional mechanistic components, and understanding and evaluating curricula.

4.3 Introduction

Reaction mechanisms are ubiquitous with organic chemistry. Understanding of reaction mechanisms, both how to draw and the meaning conveyed, is critical for success as an organic chemist (Bhattacharyya, 2013; Nedungadi and Brown, 2021). However, studies have consistently shown that learners have difficulty using the electron-pushing formalism, the language of organic chemistry (Bhattacharyya and Bodner, 2005; Anderson and Bodner, 2008; Ferguson and Bodner, 2008; Kraft et al., 2010; Grove et al., 2012b; Bhattacharyya, 2014). Learners’ ability in interpreting
reaction mechanisms influences their learning (Daniel, 2018). Therefore, we are interested in learners’ mechanistic reasoning, which broadly includes learners’ descriptions of how a reaction occurs through the movement of electrons and changes in bonding (Machamer et al., 2000; Russ et al., 2008; Yan and Talanquer, 2015; Bodé et al., 2019; Dood and Watts, 2022).

Sociologists (e.g., Collins, 2011) suggest that knowing language surrounding a practice (e.g., mechanistic reasoning; see Talanquer, 2018; Dood and Watts, 2022) may better advance associated understanding than simply engaging in the practice (e.g., drawing a reaction mechanism). In other words, knowing how to communicate within the context of a practice is necessary for participation in a practice. We assert that educators must aim to cultivate learners’ ability to put their mechanistic reasoning into words (i.e., lexical ability; Connor et al., 2021) when creating reaction mechanism learning experiences; knowing and applying the mechanistic lexicon and language has the potential to promote better understanding of reaction mechanisms.

As Cooper and others have argued, learners must have opportunities to demonstrate and wrestle with any outcome or objective we seek through our instruction (Cooper, 2015; Becker et al., 2016; Stowe and Cooper, 2017; Bodé et al., 2019; Galloway et al., 2019; Watts et al., 2020, 2022). In other words, if we want learners to ascribe meaning to reaction mechanism drawings, then we must ask learners to communicate that meaning, e.g., through written explanations. Scholarship in this area tends to favor single or few step reactions wherein an overall assessment of understanding, reasoning, etc. can be declared (Cooper et al., 2016; Caspari et al., 2018; Bodé et al., 2019; Crandell et al., 2019, 2020). However, even the simplest of reaction mechanisms to draw (arguably the bimolecular substitution reaction, i.e., one-step, two mechanistic arrows, two starting materials, two products), has multiple components that need to be jointly considered when explaining the reaction mechanism drawing; this is noted by Flynn and Ogilvie (2015) in their implementation of a carbonyl-reaction-first approach to learning organic chemistry. As well, research on learner understanding of nucleophiles and electrophiles (two entities “chunked” as
one by organic chemists) has shown that deep understanding of one is not indicative of deep understanding of the other (Anzovino and Bretz, 2015, 2016).

Constructed response items provide an important assessment means for capturing the meaning ascribed to reaction mechanisms. Constructed response items with associated evaluative rubrics (e.g., Becker et al., 2016; Cooper et al., 2016; Caspari et al., 2018; Bodé et al., 2019; Crandell et al., 2019, 2020; Deng and Flynn, 2021; Noyes et al., 2022), and sometimes computer-based scoring models (e.g., Dood et al., 2019, 2020a; Noyes et al., 2020; Yik et al., 2021), are now frequently reported in the STEM education literature. As we have asserted before, a significant limitation of these assessment items is the focus on highly specific assessment items and examples or contexts. For example, even though broad concepts such as nucleophiles are being assessed, evaluation of responses are built around specifics (e.g., understanding of nucleophiles in the electrophilic aromatic substitution reaction of benzene with chlorine and iron(III) chloride). While such scholarship is valuable, it is restrictive in that an educator wishing to use the constructed response items in their courses is limited to the specific assessment item and evaluative rubric, i.e., the scholarship does not account for variations in the prompt type or specific example. We argue that such items and associated rubrics need to be defined by concept and not specifics: for example, learning goals for understanding of a nucleophile–electrophile reaction step and an associated rubric have more utility for educators and ultimately promoting learning.

In this work, we report the development of a rubric that evaluates the level of explanation sophistication for understanding of nucleophiles in organic reaction mechanisms through written responses to open-ended formative assessment items. We start from the chemistry education and broader education research literature to define the levels of explanation sophistication including identifying areas of confusion and opportunities for learning noted in the research literature. This work also builds on prior work evaluating learner understanding of acid–base reactions (Yik et al., 2021). In that study, we noted that the language organic chemists use to
describe and explain reaction mechanisms is irrespective of the substrates and reagents in the reaction; therefore, the language used to describe mechanistic components can be conceptualized as non-reaction-type or non-reactant(s) specific. The work reported herein advances our understanding of assessing reaction mechanism learning by: (1) operationalizing levels of explanation sophistication for nucleophiles and (2) reporting a generalized rubric for evaluating a mechanistic component (i.e., nucleophiles) in the context of an entire reaction.

4.3.1 Understanding of organic chemistry reaction mechanisms

Bhattacharyya and Bodner (2005) argued that the ability to use and understand the electron-pushing formalism in reaction mechanisms is vital for comprehending organic chemistry. Educators routinely note that drawing reaction mechanisms is a foundational topic in organic chemistry education (Goodwin, 2003; Ferguson and Bodner, 2008; Goodwin, 2008; Bhattacharyya, 2013; Bhattacharyya and Harris, 2018). Results from a recent survey of organic chemistry educators in the United States reaffirmed that understanding of reaction mechanisms is important for learner success (Nedungadi and Brown, 2021). While many educators believe that when students draw a reaction mechanism, they ascribe some meaning to the structures and arrows; many students lack a deep understanding of reaction mechanisms and simply adorn their pictures with arrows because the students saw no benefit from these representations (Grove et al., 2012b). As such, students may use techniques, such as rote memorization and surface-level studying approaches, to construct reaction mechanisms (Ferguson and Bodner, 2008; Cooper and Stowe, 2018). In other words, reaction mechanism “pictures” have little to no meaning for learners.

Studies report that both undergraduate and graduate students have difficulties using the electron-pushing formalism (Bhattacharyya and Bodner, 2005; Anderson and Bodner, 2008; Ferguson and Bodner, 2008; Kraft et al., 2010; Grove et al., 2012b; Bhattacharyya, 2014). We should note that graduate student-level work in this area is often done with learners pursuing
graduate-level studies in organic chemistry; thus, if there was a population of chemists that we would expect to have advanced understanding, it would be those learners. Some have argued that this may stem from students having difficulty in simultaneously ascribing meaning to curved arrows and using chemical language to describe what is happening in reaction mechanisms (Galloway et al., 2017; Bhattacharyya and Harris, 2018). In general, students favor product-oriented learning over the deep engagement of core ideas needed in process-oriented learning that are reaction mechanisms (Anderson and Bodner, 2008; Kraft et al., 2010; Grove et al., 2012a; Graulich, 2015).

Learners struggle with reaction mechanisms as a whole, but also often lack skills necessary to reason about individual steps in a mechanism (Graulich, 2015). Cruz-Ramírez de Arellano and Towns (2014) reported that students were unsuccessful in recognizing components of alkyl halide reactions, and thus failed to understand the holistic nature of these reactions. Watts et al. (2020) analyzed students’ written explanations of an acid-catalyzed amide hydrolysis reaction to identify features of mechanistic reasoning; it was found that students were consistent in using appropriate language to describe mechanistic steps. Organic chemistry instructors support the notion that learners need to understand components of reactions, saying that the identification of functional groups and reagents along with their classification as electron donors or acceptors are necessary for students to become proficient in mechanistic reasoning using the electron-pushing formalism (Bhattacharyya, 2013). This indicates that learners need to build foundational skills in the components of reaction mechanisms to understand the whole of reaction mechanisms more meaningfully.

An overwhelming conclusion from the chemistry education research literature is that students rely on memorized pieces of knowledge and information to solve mechanistic problems with little understanding of the fundamental concepts (e.g., acid–base chemistry, nucleophiles, electrophiles) that underpin reaction mechanisms (Ferguson and Bodner, 2008; Graulich, 2015). However, students are more apt to concentrate on structural features over function and reasoning
(Domin et al., 2008; Ferguson and Bodner, 2008; Kraft et al., 2010; McClary and Talanquer, 2011; Cruz-Ramírez de Arellano and Towns, 2014; Anzovino and Bretz, 2015; Galloway et al., 2017; Graulich and Bhattacharyya, 2017; Galloway et al., 2019; Dood et al., 2020a; Lapierre and Flynn, 2020; Petterson et al., 2020; Xue and Stains, 2020). Students report that constructing reaction mechanisms is meaningless to them and that instruction focused on how reaction mechanisms proceed with little emphasis on the why (Ferguson and Bodner, 2008). However, expert-level mechanistic reasoning requires integrating multiple electrostatic concepts such as formal and partial charges, polarity, electronegativity, electron density, and also identifying and classifying nucleophiles and electrophiles (Bhattacharyya, 2013).

4.3.2 Understanding of nucleophiles

In this paper, we focus on the understanding of nucleophiles. The research literature supports the notion that nucleophiles and electrophiles are not equally understood (Strickland et al., 2010; Anzovino and Bretz, 2015, 2016; Putica and Trivic, 2016) with learners being more successful in identifying and explaining nucleophiles. While organic chemists may view these two entities as a combination, we note that learners need a more foundational understanding of each as they develop understanding of the relationship between the two. We chose nucleophiles to begin our broader work in this area all-the-while noting that this is merely a first step.

Learners hold alternative conceptions regarding nucleophiles and struggle to articulate conceptualizations about nucleophiles. It is documented that students believe nucleophiles are positively charged, electron deficient, and tend to accept electrons; these are all the opposite of what nucleophiles are and how nucleophiles participate in bonding. In other words, students confuse the properties of nucleophiles with those of electrophiles (Akkuzu and Uyulgan, 2016). In a study by Putica and Trivic (2016), a majority of students made incorrect attempts at defining nucleophiles, again conflating nucleophilic properties with those of electrophiles. Additionally, chemistry undergraduates at the end of their four-year degree studies have been reported to
struggle in articulating their conceptions of implicit chemical properties, such as nucleophilicity and the electronic interactions that drive chemical reactivity (DeFever et al., 2015).

Students use rote memorization and surface-level practices to identify nucleophiles and nucleophilic behavior (Anzovino and Bretz, 2015). Students also engage in rote memorization of surface-level features related to nucleophilic behavior rather than engaging in deeper-level relationships between structure and reactivity (Anzovino and Bretz, 2015). In another study, students were more successful at identifying nucleophiles than electrophiles (Strickland et al., 2010) and are prone to prioritize structural features to identify nucleophiles in reaction mechanisms (Anzovino and Bretz, 2015, 2016; Weinrich and Sevian, 2017). However, students generally struggle at identifying nucleophiles without a reaction mechanism; this suggests that students are not using mechanistic reasoning when engaging with reaction mechanisms (Strickland et al., 2010).

Students have difficulty relating nucleophiles to other chemical concepts. The development of an assessment of concepts important for developing proficiency in organic reaction mechanisms showed that items related to nucleophiles are difficult; this is compounded with items related to resonance, inductive effects, and intermediate stability being even more challenging (Nedungadi et al., 2020). Multiple studies have reported that students have difficulty in being able to distinguish between nucleophiles and bases, and the connection between the two entities (Cartrette and Mayo, 2011; Cruz-Ramírez de Arellano and Towns, 2014; Anzovino and Bretz, 2015, 2016).

4.3.3 Moving toward meaningful assessments

Much of the research literature focused on student understanding of reaction mechanisms has been done by evaluating students’ construction of reaction mechanism drawings and predicting products using the formalism (Bhattacharyya and Bodner, 2005; Kraft et al., 2010; Strickland et al., 2010; Grove et al., 2012b; Flynn and Featherstone, 2017; Galloway et al., 2017;
DeCocq and Bhattacharyya (2019) reported that exercises, such as drawing mechanistic reaction arrows and predicting reaction products, are not successful at eliciting and evaluating reasoning. Numerous studies have concluded that increased cognitive demand from proposing mechanisms may cause students to not use mechanisms when predicting products, and thus further resorting to rote memorization and surface-level approaches to learning (Ferguson and Bodner, 2008; Grove et al., 2012a; Flynn and Featherstone, 2017; Galloway et al., 2017; Cooper and Stowe, 2018). Before expecting learners to ascribe meaning to reaction mechanisms they have drawn, we assert that learners need opportunities to construct meaning from “correct” reaction mechanisms given to them by cultivating their mechanistic reasoning through written formative assessments by scaffolding learning.

Work by multiple research teams (e.g., Cooper, 2015; Cooper et al., 2016; Dood et al., 2018, 2020a; Bodé et al., 2019; Yik et al., 2021; Raker et al., 2023), have asked students what is happening in reaction mechanism and why it happens. This “what and why?” scaffold has been used to prompt mechanistic understanding in acid–base chemistry (Cooper et al., 2016; Dood et al., 2018, 2019; Crandell et al., 2019; Yik et al., 2021), unimolecular nucleophilic substitution reactions (Dood et al., 2020a), and bimolecular nucleophilic substitution reactions (Crandell et al., 2020). Requiring students to explicitly address each step of a reaction mechanism using underlying fundamental concepts will aid students in appreciating the utility of the electron-pushing formalism (Bhattacharyya and Bodner, 2005). The electron-pushing formalism and reaction mechanisms are models and representations that are essential to reasoning in organic chemistry and provide information about the conceptions of chemical transformations, and therefore are used in the construction of explanations (Goodwin, 2003).

Constructing explanations of reaction mechanisms engages students in key scientific practices. The Framework for K-12 Science Education and the Next Generation Student Standards highlight using models, constructing explanations, and communicating information as essential practices for learning and doing science (National Research Council, 2012; NGSS Lead
States, 2013). Engagement in these scientific practices, such as constructing explanations, will aid students in deepening their understanding of science. For example, the writing process aids in deeper thinking and engagement with course content (Rivard, 1994; Bangert-Drowns et al., 2004; Reynolds et al., 2012).

Writing assessments using purposeful constructed-response prompts can promote understanding of chemical concepts (Birenbaum and Tatsuoka, 1987; Cooper, 2015; Stowe and Cooper, 2017; Underwood et al., 2018). Open-ended written assessments should prompt students what is happening in the reaction mechanism and why it happens to promote deeper learning and allow instructors to gain deeper insight into students’ understanding (Bell and Cowie, 2001; Cooper, 2015; Cooper et al., 2016; Underwood et al., 2018). While there are many benefits to open-ended written assessments, providing consistent and prompt feedback using such assessments is difficult.

Rubrics are a possible solution (Wolf and Stevens, 2007; Brookhart, 2018). According to Andrade (2000), rubrics are beneficial because they: (i) are easy to use and explain, (ii) convey clear expectations, (iii) provide students with more feedback than traditional assessments, and (iv) support learning through the development of skills and understanding. Rubrics help articulate instructors’ expectations by describing the qualities of an assessment across a continuum and for providing feedback on student work (Andrade, 2000; Brookhart and Chen, 2015). In developing a rubric, Dawson (2017) recommends a broad set of design elements that should be considered. This objective to assess students can be practically used in, but not limited to, competency-based grading (see Voorhees, 2001) and specifications grading approaches (see Nilson, 2015). In these grading systems, the instructor defines a passing or proficiency threshold on an assessment, for example, reaching a certain level in a rubric for understanding of nucleophiles in a reaction mechanism, and the token system in specifications grading provides students with options to improve and resubmit work that does not meet satisfactory criteria (see Nilson, 2015; Howitz et al., 2021). Specifications grading has been used in various disciplines, including lecture-based
chemistry courses (e.g., Diegelman-Parente, 2011; Boesdorfer et al., 2018; Martin, 2019), and therefore, rubrics provide a means to assess such learning in a variety of classroom contexts.

4.3.4 Rubric for evaluating understanding of nucleophiles

Students’ explanations of reaction mechanisms have been evaluated using a variety of frameworks: levels of complexity of relations (Caspari et al., 2018), Chemical Thinking Learning Progression (Bodé et al., 2019), causal versus mechanistic (Cooper et al., 2016; Crandell et al., 2019, 2020), and levels of explanation sophistication (Dood et al., 2020a). For a more comprehensive review of reasoning frameworks, see a review by Dood and Watts (2022) on mechanistic reasoning in organic chemistry.

Levels of complexity of relations. Caspari et al. (2018) characterized students’ comparisons of the activation energy of two mechanistic steps rooted in the complexity of relations (i.e., low, middle, and high). Relations with low complexity use explicit structural differences as a cause to describe relation, or non-electronic effects, to explain change. Relations with middle complexity use explicit structural differences to infer implicit structural causes that are used to describe relation, or non-electronic effects, to explain change. Relations with high complexity use explicit structural differences to infer implicit structural causes that are used to describe an electronic effect to explain change.

Chemical Thinking Learning Progression. Bodé et al. (2019) explored students’ explanations of the likelihood of reaction progression between two reaction mechanisms. The Chemical Thinking Learning Progression framework (Sevian and Talandquer, 2014) was used to characterize these explanations. The framework characterizes responses by mode of reasoning: descriptive, relational, linear causal, and multicomponent causal. Descriptive responses state explicit features without further explanation. Relational responses use explicit and implicit features to draw connections, but do not explain a cause–effect relationship. Linear causal responses build upon relational responses and describe a cause-and-effect relationship using a single variable,
and multicomponent causal responses describe cause-and-effect relationships using multiple variables.

Causal mechanistic reasoning. Cooper with others (Cooper et al., 2016; Crandell et al., 2019, 2020) have characterized student reasoning about acid–base and bimolecular nucleophilic substitution (S\textsubscript{N}2) reactions using a causal mechanistic reasoning framework. Descriptive general responses provide a simplistic description of bond breaking/forming processes whereas descriptive causal response also discusses electrostatic interaction and demonstrates understanding of electrostatic attraction. Descriptive mechanistic responses recognize electrons and explains the movement of electrons, whereas causal mechanistic responses build upon descriptive mechanistic by understanding the electrostatic attraction between chemical species.

Levels of explanation sophistication. Dood et al. (2020a, 2020b) reported an emergent framework for evaluating explanations of a unimolecular nucleophilic substitution (S\textsubscript{N}1) reaction. Responses were characterized into three levels: Level 1 (descriptive), Level 2 (surface-level why), and Level 3 (deeper-level why). Level 1 responses describe reactions using explicit features with no explanation. Level 2 responses use surface level features to explain why the reaction is happening or uses implicit features with no elaboration. Level 3 responses describe why the reaction happens using implicit features that have been inferred from explicit features or uses electronic effects to explain the reaction.

The levels of explanation sophistication rubric proposed herein synthesizes and operationalizes these four frameworks as informed by the research literature on understanding of reaction mechanisms and nucleophiles. Our rubric for evaluating the understanding of nucleophiles is based on a conceptualization of a generalized rubric for levels of explanation sophistication for the understanding of electrophiles reported by Raker et al. (in press). This rubric is general as it can be applied to a variety of written formative assessment items (see specificity in Dawson, 2017). In operationalizing our nucleophile rubric, four hierarchical levels of sophistication were identified: Absent, Descriptive, Foundational, and Complex (Table 4.1).
rubric builds upon our previous levels of explanation sophistication framework (Dood et al., 2020a, 2020b) with the addition of the *Absent* level and mirrors other studies in the literature (Sevian and Talanquer, 2014; Cooper et al., 2016; Weinrich and Talanquer, 2016; Caspari et al., 2018; Bodé et al., 2019; Graulich et al., 2019; Crandell et al., 2020; Dood et al., 2020a; Watts et al., 2020) that suggest students’ reasoning abilities progress through a hierarchical relationship. Additionally, we have changed the level names from our previous iteration of the framework so that the level names may be better understood and applied by learners and instructors in regard to description and hierarchy of the levels. This analytic scoring strategy applies a single level for the evaluative criterion of nucleophiles (see *evaluative criteria* and *scoring strategy* in Dawson, 2017). We envision that this rubric can be a paper-based copy, a digital-based document file, or be translated into a rubric in an online learning management system (see *presentation* in Dawson, 2017).

The *Absent* level has been left out of prior frameworks for evaluating understanding of reaction mechanisms, by us and others (Cooper et al., 2016; Bodé et al., 2019; Crandell et al., 2019, 2020; Dood et al., 2020a, 2020b). It is included here as we note that learners can and do respond with “I’m not sure” or “I haven’t studied this yet” when asked to explain (Dood et al., 2020a). Thus, an assessment tool for practical use in instruction needs a means to characterize and address such responses. The *Absent* level also addresses instances where explanations of reaction mechanisms may glance over a mechanistic step or provide no understanding about a concept: For example, “the product is formed through an S_N2 reaction;” this response does not address the nucleophile in the reaction, nor understanding of nucleophiles in a bimolecular substitution reaction.

The *Descriptive* level characterizes simple narratives of bond forming processes or states what is happening without any explanation of why: “the ethoxide attacks the carbon with the
<table>
<thead>
<tr>
<th>Level</th>
<th>Description of level</th>
<th>Key features of level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absent</td>
<td>• No response</td>
<td>• The nucleophile is not identified</td>
</tr>
<tr>
<td></td>
<td>• Non-normative</td>
<td>• The nucleophile-electrophile reaction step is missing from the explanation</td>
</tr>
<tr>
<td>Descriptive</td>
<td>• Describes the nucleophile engaging in bond forming processes</td>
<td>• The nucleophilic molecule or atom is identified</td>
</tr>
<tr>
<td></td>
<td>• Simplistic description of bond forming processes</td>
<td>• Electrons are not used to describe nucleophilic behavior</td>
</tr>
<tr>
<td></td>
<td>• Nucleophilic behavior is described at a surface, atomic level</td>
<td>• Bond forming processes are described (e.g., “nucleophile attacks/forms a bond with electrophile”)</td>
</tr>
<tr>
<td>Foundational</td>
<td>• Nucleophilic behavior is described at a surface, electronic level</td>
<td>• Electrons are central to nucleophilic behavior and arrows represent the movement of electrons</td>
</tr>
<tr>
<td></td>
<td>• Explicit features are mentioned</td>
<td>• Example descriptors of electrons: sigma electrons/bond, pi electrons/bond, lone pair (and not alkene or double bond)</td>
</tr>
<tr>
<td></td>
<td>• Implicit features may be mentioned, but not fully explained</td>
<td>• Bond forming processes are described using electrons (e.g., “the lone pair on the nucleophile attacks/forms a bond with electrophile”)</td>
</tr>
<tr>
<td>Complex</td>
<td>• Describes why the nucleophile is involved in bond forming processes</td>
<td>• Implicit electronic features are used to describe nucleophilic behavior</td>
</tr>
<tr>
<td></td>
<td>• Nucleophilic behavior is described at a deeper, electronic level</td>
<td>• Examples of electronic properties: electron density, electronegativity, partial charges, molecular orbital descriptions (e.g., HOMO/LUMO)</td>
</tr>
<tr>
<td></td>
<td>• Explicit features are used to infer implicit features that are sufficiently explained</td>
<td>• Bond forming processes are described using electrons and electronic properties (e.g., “the high electron density on the nucleophile is attracted to the area of low electron density on the electrophile”)</td>
</tr>
</tbody>
</table>
bromide” and lacks recognition of electrons. Here, students are associating the general language used to describe reaction mechanisms using chemical names and processes with a primary focus on explicit structural features (see descriptive in Bodé et al., 2019; see descriptive general in Crandell et al., 2020; see Level 1 in Dood et al., 2020a, 2020b). A key component at this level is that responses properly identify the nucleophile (Anzovino and Bretz, 2015).

The Foundational level is given to responses that recognize that electrons are responsible for interactions between chemical species leading to bond forming/breaking processes (cf. descriptive mechanistic in Crandell et al., 2020). A response stating, “the lone pair on the ethoxide attacks the carbon with the bromide” is an example at the Foundational level where the key difference is in the addition of the phrase “lone pair.” Implicit features may be mentioned but not fully explained at this level (cf. Level 2 in Dood et al., 2020a, 2020b).

Lastly, the Complex level characterizes responses at a deeper, electronic understanding. These responses describe both explicit and implicit features and includes descriptions of electronic properties or interactions that lead to reactivity, such as electron density, electronegativity, partial charges, and molecular orbital theory descriptions (see causal mechanistic in Crandell et al., 2020; see Level 3 in Dood et al., 2020a, 2020b).

4.3.5 Research goal

Our work is guided by one primary goal: to develop a generalized rubric to evaluate the level of explanation sophistication for nucleophile understanding in organic chemistry reaction mechanisms. We seek to demonstrate that the rubric has utility in a variety of courses (e.g., first or second term), for a variety of nucleophile-containing reactions (e.g., unimolecular nucleophilic substitution reactions or acetal formation reactions), and with a variety of cued assessment prompts.
4.4 Methods

This work was conducted under application Pro#00028802, “Comprehensive evaluation of the University of South Florida’s undergraduate and graduate chemistry curricula,” as reviewed by the University of South Florida’s Institutional Review Board on December 13, 2016; the activities were determined to not constitute research involving human subjects per Institutional Review Board criteria and thus approval was not needed.

4.4.1 Constructed response items

The constructed response item format used in this study was initially reported for use in a single-step proton transfer reaction (Cooper et al., 2016), and modified and expanded for use in other acid–base (Dood et al., 2018, 2019; Crandell et al., 2019; Yik et al., 2021) and multi-step nucleophilic substitution reactions (Crandell et al., 2020; Dood et al., 2020a, 2020b). This format includes asking respondents to describe what is happening in a reaction and to explain why the reaction occurs. In total, 85 reaction mechanisms were used in this work. A summary of these reaction mechanisms is given in Table 2; example reaction mechanisms are provided in Figure 4.1 (see exemplars in Dawson, 2017). A full list of assessment items is located in Appendix B.

Two prompt variations were used to collect explanations (Table 4.3). These are labeled Original and More Cued to convey the intent of changes employed in the second prompt. In the Results and Discussion, we will discuss the intent of each prompt and any observable differences in response patterns.

4.4.2 Data collection

Data were collected from seven semesters (Fall 2017, Spring 2018, Spring 2019, Fall 2019, Spring 2020, Fall 2020, and Fall 2021) of the first semester (Organic Chemistry 1) and two semesters (Spring 2020 and Fall 2021) of the second semester (Organic Chemistry 2) of a year-long organic chemistry course at the University of South Florida, a large, research-intensive,
Table 4.2. Reaction mechanisms used

<table>
<thead>
<tr>
<th>Functional group</th>
<th>Reaction</th>
<th>Variations</th>
<th>Number of explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alkyl halide</td>
<td>Bimolecular nucleophilic substitution (S₉₂)</td>
<td>4</td>
<td>820</td>
</tr>
<tr>
<td></td>
<td>Unimolecular nucleophilic substitution (S₉₁)</td>
<td>8</td>
<td>5,050</td>
</tr>
<tr>
<td>Alkene</td>
<td>Halogenation/hydrohalogenation</td>
<td>5</td>
<td>1,262</td>
</tr>
<tr>
<td></td>
<td>Halohydrin formation</td>
<td>3</td>
<td>717</td>
</tr>
<tr>
<td></td>
<td>Hydration/dehydration</td>
<td>4</td>
<td>1,276</td>
</tr>
<tr>
<td>Alkyne</td>
<td>Alkylation</td>
<td>1</td>
<td>411</td>
</tr>
<tr>
<td>Alcohol</td>
<td>Conversion to good leaving group, then S₉₂</td>
<td>4</td>
<td>932</td>
</tr>
<tr>
<td></td>
<td>Epoxidation/ring-opening</td>
<td>6</td>
<td>1,284</td>
</tr>
<tr>
<td>Aromatic ring</td>
<td>Acylation</td>
<td>1</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td>Addition-elimination</td>
<td>1</td>
<td>136</td>
</tr>
<tr>
<td></td>
<td>Alkylation</td>
<td>1</td>
<td>196</td>
</tr>
<tr>
<td></td>
<td>Azo coupling</td>
<td>1</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>Electrophilic aromatic substitution</td>
<td>6</td>
<td>657</td>
</tr>
<tr>
<td></td>
<td>Nucleophilic aromatic substitution</td>
<td>2</td>
<td>225</td>
</tr>
<tr>
<td>Carbonyl</td>
<td>Acyl substitution</td>
<td>3</td>
<td>403</td>
</tr>
<tr>
<td></td>
<td>Aldol addition/condensation</td>
<td>2</td>
<td>273</td>
</tr>
<tr>
<td></td>
<td>Alpha-halogenation</td>
<td>1</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>Condensation of an ester/diester</td>
<td>2</td>
<td>402</td>
</tr>
<tr>
<td></td>
<td>Conjugate addition</td>
<td>2</td>
<td>255</td>
</tr>
<tr>
<td></td>
<td>Enamine/imine synthesis</td>
<td>2</td>
<td>367</td>
</tr>
<tr>
<td></td>
<td>Esterification</td>
<td>1</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td>Grignard</td>
<td>6</td>
<td>961</td>
</tr>
<tr>
<td></td>
<td>Hemiacetal/acetal formation</td>
<td>4</td>
<td>885</td>
</tr>
<tr>
<td></td>
<td>Hydration/dehydration</td>
<td>2</td>
<td>137</td>
</tr>
<tr>
<td></td>
<td>Hydrolysis</td>
<td>1</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Reduction</td>
<td>9</td>
<td>1,825</td>
</tr>
<tr>
<td></td>
<td>Saponification</td>
<td>1</td>
<td>202</td>
</tr>
<tr>
<td>Conjugated diene</td>
<td>Electrophilic addition</td>
<td>2</td>
<td>553</td>
</tr>
</tbody>
</table>

public university in the southeastern United States. These courses were taught by four instructors: authors DCR, KBF, and JRR, and F. Costanza (see Acknowledgments). Data were collected in author DCR’s Organic Chemistry 1 course in Spring 2018, Spring 2019, Fall 2019, Spring 2020, Fall 2020, and Fall 2021. Data were collected in author KBF’s Organic Chemistry 1 course in Fall 2017, Fall 2019, Fall 2020, and Fall 2021; and Organic Chemistry 2 course in Spring 2021. Data were collected in F. Costanza’s Organic Chemistry 2 course in Spring 2020. Data were collected in JRR’s Organic Chemistry 1 course in Fall 2017, Fall 2019, and Spring 2020; and Organic Chemistry 2 course in Fall 2021. For both courses (Organic Chemistry 1 and Organic Chemistry 2), between Fall 2017 and Spring 2019, the textbook used was Solomons, Fryhle, and Snyder’s
Figure 4.1. Example reaction mechanisms

Table 4.3. Constructed response item prompt variations

<table>
<thead>
<tr>
<th>Original</th>
<th>More Cued</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Part A</strong>: Describe in full what you think is happening on the molecular level for this reaction. Be sure to discuss the role of the reactant and intermediate. <strong>Part B</strong>: Using a molecular level explanation, explain why this reaction occurs. Be sure to discuss why the reactants form the products shown.</td>
<td><strong>Part A</strong>: Describe in full detail the sequence of events that occur at the molecular level for this reaction. Be sure to discuss the role of each reactant and intermediate. <strong>Part B</strong>: Using a molecular level explanation, explain why each of the reactants and intermediates interact.</td>
</tr>
</tbody>
</table>

**Organic Chemistry, 12th edn** (2016); between Fall 2019 and Spring 2021, the textbook used was Klein’s **Organic Chemistry, 3rd edn** (2017), and Klein’s **Organic Chemistry, 4th edn** (2021) was used in Fall 2021. For all semesters, constructed response items used the **Original** prompt; data collected in Fall 2021 used the **Original** and **More Cued** prompts.
Constructed response items were administered via Qualtrics surveys. In total, 85 unique reaction mechanism prompts were administered (see Table 4.2) resulting in 19,936 responses. Participants received extra credit towards their term or final examination grade for completing the assessment. Study participants may have completed up to four surveys in a given semester and participants may have completed assessment items in both their Organic Chemistry 1 and Organic Chemistry 2 courses. Some data for unimolecular nucleophilic substitution reactions have been analyzed analogously and reported in other studies (Dood et al., 2020a, 2020b).

4.5 Results and discussion

The value of a rubric is in its use as an assessment tool to provide feedback to students and evaluative information for instructors. In addition, students are able to use a rubric to monitor their learning and instructors are able to use a rubric for grading or evaluation (Panadero and Jonsson, 2013). Here we describe the use of our nucleophile rubric in the context of postsecondary Organic Chemistry 1 and Organic Chemistry 2 courses.

4.5.1 Refining and operationalizing the rubric levels

Our levels of explanation sophistication rubric is reported in Table 4.1, above. In applying this rubric, responses from first and second semester courses that explain reaction mechanisms with multiple nucleophile types (e.g., lone pair, sigma, pi), and span a variety of reaction families (e.g., aromatic reactions, nucleophilic additions, reduction reactions) were first analyzed by author BJY. In this process, author BJY applied the rubric (Table 4.1) and grouped similar responses based on the language used to explain nucleophiles; characteristics of these responses that distinguished responses from others were noted. Through peer debriefing discussions (see Lincoln and Guba, 1985), author BJY presented descriptions, key features, and examples of responses in each of the four levels of the rubric to authors SJHF and JRR in which these aspects
were refined, and inclusion criteria were solidified; this led the characterization of key features for each level in the nucleophile rubric (see “key features of level” in Table 4.1).

To ensure reliability of the data obtained by using the rubric, interrater reliability checks were performed (see the quality processes criteria in Dawson, 2017). Two raters were used in the interrater process; authors BJY and SJHF are chemistry graduate students in chemistry education research and have served as teaching assistants for at least one of the organic chemistry courses. As the rubric should be accessible and useful for a variety of learners and educators, authors BJY and SJHF are well suited for establishing judgment complexity criteria (Dawson, 2017). Author BJY applied the rubric (Table 4.1) using the refined key features (Table 4.1; see Appendix B.1 for complete rubric including exemplars) to classify all 19,936 responses. For reactions that have two nucleophile interaction components, the higher of the two individual levels was used to classify a response. Author SJHF then independently classified a set of 200 randomly selected responses (~1% overall) representative of the distribution of nucleophile type and reaction family type of the entire data set. Authors BJY and SJHF initially agreed on 53% (n = 107) of the responses with a weighted (equal weights) kappa of 0.37; after discussion about disagreements, author BJY changed classifications for 10% (n = 20) of the responses, and author SJHF changed classifications for 35% (n = 71) of the responses. There was a final agreement of 96% (n = 193) with a weighted (equal weights) kappa of 0.94. As a result of these discussions, author BJY reevaluated responses with a particular focus on responses classified under the Complex level for unimolecular substitution reactions and reactions involving a carbocation intermediate, and the rubric was further refined to better convey inclusion criteria for different levels which resulted in the rubric and key features as presented in Table 4.1 (see Appendix B.1).

To demonstrate reliability of the data using the refined rubric (see Table 4.1), a second round of interrater reliability evaluation was performed. Two raters were used in the interrater process: author BJY and an organic chemistry instructor. The refined rubric (Table 4.1) was applied to a set of 200 randomly selected responses (~1% overall), different from the data set
used in the original interrater reliability evaluation. The two raters agreed on 87% ($n = 174$) of the responses with a weighted (equal weights) kappa of 0.81; disagreements differed no more than one level.

### 4.5.2 Rubric use by Organic Chemistry 1 and Organic Chemistry 2 courses

The distribution of revised and finalized classifications for the level of explanation sophistication for nucleophiles for the overall data set from Organic Chemistry 1 and Organic Chemistry 2 contexts are provided in Table 4. The distribution for these data sets show that most responses are centered within the middle of the rubric (i.e., the *Descriptive* and *Foundational* levels) with fewer students at the extreme ends of the rubric (i.e., the *Absent* and *Complex* levels). When describing this data by course (i.e., Organic Chemistry 1 and Organic Chemistry 2), this same distribution pattern holds.

<table>
<thead>
<tr>
<th>Level</th>
<th>Organic Chemistry 1</th>
<th>Organic Chemistry 2</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>13,559</td>
<td>6,377</td>
<td>19,936</td>
</tr>
<tr>
<td>Absent</td>
<td>1,788 (13.2)</td>
<td>1,019 (16.0)</td>
<td>2,807 (14.1)</td>
</tr>
<tr>
<td>Descriptive</td>
<td>7,407 (54.6)</td>
<td>3,473 (54.4)</td>
<td>10,880 (54.6)</td>
</tr>
<tr>
<td>Foundational</td>
<td>3,842 (28.3)</td>
<td>1,605 (25.2)</td>
<td>5,447 (27.3)</td>
</tr>
<tr>
<td>Complex</td>
<td>522 (3.9)</td>
<td>280 (4.4)</td>
<td>802 (4.0)</td>
</tr>
</tbody>
</table>

Scale attenuation effects occur when a response format includes too few options to reflect, in this case, response sophistication about nucleophiles. Two types of scale attenuation effects are floor and ceiling effects. Floor effects describe when too many responses are at the lower limit of the scale and ceiling effects describe when too many responses are at the upper limit of the scale (see Šimkovic and Träuble, 2019); scale attenuation effects are considered problematic in measurement theory because the measured variable becomes insensitive to changes in the latent variable and affects the performance of statistical models. A distribution of 15% at the floor or ceiling of a scale has been historically accepted as demonstrating moderate scale attenuation.
effects (McHorney and Tarlov, 1995); however, this standard has been widely used in the medical community where floor or ceiling effects can have substantive effects on patient health. While our results show nearing a floor effect, our scale is sufficient as the Absent level describes the lack of the concept of nucleophiles in a response and the next level in the scale (i.e., Foundational) describes the next logical step where the response acknowledges the existence of a nucleophile. Additionally, other frameworks describe the highest level as containing multicomponent, causal, or electrostatic descriptions, which are similar to our Complex level, and are based in three to six level scales (Caspari et al., 2018; Bodé et al., 2019; Crandell et al., 2020). Moreover, these frameworks suggest that these ideas are the highest observable descriptions at the target course level. Therefore, we argue that a ceiling effect would suggest that student responses are at the highest level of sophistication and meet or exceed instructor expectation.

Table 4.5 shows exemplar response excerpts for the levels in the rubric for the concept of nucleophiles (see exemplars in Dawson, 2017). Reactions A through E given in Figure 4.1 correspond respectively to those reactions noted in Table 5. Here, we provide exemplars of all levels for a single reaction (Example 1; responses from Organic Chemistry 1) and exemplars for different reactions at each of the levels (Example 2; responses from Organic Chemistry 2).

4.5.3 Rubric use by nucleophile type

To provide further evidence for the generalizability of our nucleophile rubric, Table 4.6 shows the distribution by nucleophile type. Again, most responses reside within the Descriptive and Foundational levels with very few responses at the Complex level. In addition to prompts that feature a single nucleophile type, we also collected responses for prompts that featured combinations of nucleophile types (Table 4.6); similar distributions are observed which speaks to the further generalizability of the rubric.
4.5.4 Rubric use by reaction family

Responses were also analyzed by reaction family (Table 4.6). We classified reactions into five major family types: reactions involving a carbocation intermediate, aromatic reactions, nucleophilic addition, reductions reactions, and bimolecular nucleophilic substitution reactions \((S_N2)\). Although a type of nucleophilic addition reaction, we segregated reduction reactions as its own family type due to shared commonality in reducing a carbonyl; as all Grignard reactions included reaction with a carbonyl, Grignard reactions are also classified in this reduction reaction category. Similar to our other analyses, we observe the majority of responses are in the middle levels of the rubric, and that there are more responses at the lower extreme than the upper extreme showing that there is more room to demonstrate improvement. This pattern is also observed for responses that wrote about a reaction that contained two family types (e.g., nucleophilic addition and \(S_N2\)).

In prior work with acid–base reactions (Dood et al., 2018; Yik et al., 2021), we observed that students may use analogous language for nucleophiles to describe Lewis bases (Anzovino and Bretz, 2015, 2016). For example, a response could include, “a lone pair on the alkoxide will attack the proton on water.” While this statement describes a proton transfer reaction, the statement relies on language indicative of nucleophiles. In comparing prompts that do contain a proton transfer step with those that do not (Table 4.8), we observe similar distributions between the two that resemble similar patterns in the other comparisons. This evidence suggests that other aspects of a reaction mechanism do not readily influence how learners describe nucleophiles.

4.5.5 Rubric use by prompt variations

The last validity evidence for our rubric is the level of explanation sophistication for nucleophiles collected with multiple prompt types. In the midst of our data collection reported herein, Crandell et al. (2020) modified their original prompt to investigate how wording might affect the responses by activating different resources. This is corroborated by a study that found that
Table 4.5. Example excerpts from responses for the levels of explanation sophistication for nucleophiles\textsuperscript{a}

<table>
<thead>
<tr>
<th>Level</th>
<th>Example 1</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absent</td>
<td>[Reaction A] “The first step of the S_N2 reaction is a nucleophilic attack followed by the loss of a leaving group. The interaction happens at the same time.”</td>
<td>[Reaction B] “The phenyl magnesium bromide is the Grignard reagent used in this Grignard reaction. The Grignard reagent turns the benzoic acid into a carboxylic acid.”</td>
</tr>
<tr>
<td>Descriptive</td>
<td>[Reaction A] “The iodine is acting as a leaving group and detaching from the 1-iodopropane. At the same time, the cyanide is attacking the molecule and substituting the leaving group. The reactants and intermediates interact because this is a substitution reaction. Iodine is a good leaving group so therefore this triggers the substitution reaction. Prompting the cyanide to replace it.”</td>
<td>[Reaction C] “The negatively charged OCH$_3$ will attack the carbonyl group, pushing one of the bonds up onto the O making it negatively charged. A lone pair will come back down to make a double bond again and the Cl will leave negatively charged.”</td>
</tr>
<tr>
<td>Foundational</td>
<td>[Reaction A] “The I is a good leaving group and it is leaving while the electron pair on the carbon connected to the triple bond is attacking the carbon that the leaving group is leaving from. They all interact in the same step simultaneously.”</td>
<td>[Reaction D] “The alcohol group removes a hydrogen which creates a double bond and makes the carbonyl oxygen single bonded with a negative charge. The electrons in the double bond attack the other carbonyl carbon forming a bond between the two molecules. The electrons on the negatively charged oxygen attack a hydrogen forming it into an alcohol.”</td>
</tr>
<tr>
<td>Complex</td>
<td>[Reaction A] “The cyanide is acting as a nucleophile attacking the partially positive alpha carbon kicking off the iodine leaving group, thus forming the product. The electron rich cyanide is attracted to the alpha carbon because of its partial positive charge.”</td>
<td>[Reaction E] “The oxygen molecules on the sulfur molecule are more electronegative therefore pulling out density and making the sulfur atom have a strong partial positive charge. The electrons [in the] benzene ring are very attracted to this charge and attach the sulfur breaking on [one] of the double bonds to oxygen making it a single. This reaction occurs because benzene is a very stable and strong nucleophile and is attracted to strong electrophiles.”</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Reactions A through E correspond respectively to Reactions A through E in Figure 4.1

Table 4.6. Distribution of levels of explanation sophistication by nucleophile type

<table>
<thead>
<tr>
<th>Level</th>
<th>Lone pair &amp; sigma</th>
<th>Sigma</th>
<th>Pi</th>
<th>Lone pair &amp; sigma &amp; pi</th>
<th>Lone pair &amp; sigma</th>
<th>Sigma &amp; pi</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>12,990</td>
<td>2,786</td>
<td>2,117</td>
<td>532</td>
<td>1,426</td>
<td>85</td>
</tr>
<tr>
<td>Absent</td>
<td>1,623 (12.5)</td>
<td>586 (21.0)</td>
<td>409 (19.3)</td>
<td>50 (9.4)</td>
<td>134 (9.4)</td>
<td>5 (5.9)</td>
</tr>
<tr>
<td>Descriptive</td>
<td>7,163 (55.2)</td>
<td>1,798 (64.5)</td>
<td>913 (43.1)</td>
<td>327 (61.5)</td>
<td>629 (44.1)</td>
<td>50 (58.8)</td>
</tr>
<tr>
<td>Foundational</td>
<td>3,590 (27.6)</td>
<td>347 (12.5)</td>
<td>718 (33.9)</td>
<td>148 (27.8)</td>
<td>623 (43.7)</td>
<td>21 (24.7)</td>
</tr>
<tr>
<td>Complex</td>
<td>614 (4.7)</td>
<td>55 (2.0)</td>
<td>77 (3.7)</td>
<td>7 (1.3)</td>
<td>40 (2.8)</td>
<td>9 (10.6)</td>
</tr>
</tbody>
</table>
students only considered nucleophiles and electrophiles when prompted to do so (Cartrette and Mayo, 2011). In the context of work reported herein and in response to Crandell and colleagues’ work, we changed “explain why this reaction occurs. Be sure to discuss why the reactants form the products shown” in the *Original* prompt to “explain why each of the reactants and intermediates interact” in the *More Cued* prompt to make explicit that we would like for responses to explain why for each mechanistic step.

Responses in our work were collected primarily with the *Original* prompt (see Table 4.3), with some collected using the *More Cued* prompt. The distribution of responses among the levels (Table 4.9) is again similar between the two prompt types and parallels our prior analyses. It is incumbent upon us to note that the data, as reported in Table 4.9, were not collected per an experimental design that would allow for comparison of response distributions between the two. We recognize that such work would be important; however, our data set does not allow for such an analysis.

---

Table 4.7. Distribution of levels of explanation sophistication by reaction family

<table>
<thead>
<tr>
<th>Level</th>
<th>Carbocation</th>
<th>Aromatic</th>
<th>Nucleophilic addition</th>
<th>Reduction</th>
<th>S&lt;sub&gt;N&lt;/sub&gt;2</th>
<th>Nucleophilic addition &amp; S&lt;sub&gt;N&lt;/sub&gt;2</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>7,432</td>
<td>1,581</td>
<td>3,264</td>
<td>2,786</td>
<td>4,612</td>
<td>261</td>
</tr>
<tr>
<td>Absent</td>
<td>1,008</td>
<td>252</td>
<td>493</td>
<td>586</td>
<td>437</td>
<td>(11.9)</td>
</tr>
<tr>
<td>Descriptive</td>
<td>3,757</td>
<td>780</td>
<td>1,891</td>
<td>1,798</td>
<td>2,504</td>
<td>(57.4)</td>
</tr>
<tr>
<td>Foundational</td>
<td>2,238</td>
<td>463</td>
<td>780</td>
<td>347</td>
<td>1,544</td>
<td>(28.7)</td>
</tr>
<tr>
<td>Complex</td>
<td>429</td>
<td>86</td>
<td>100</td>
<td>55</td>
<td>127</td>
<td>(2.0)</td>
</tr>
</tbody>
</table>

Table 4.8. Distribution of levels of explanation sophistication by reactions involving proton transfers

<table>
<thead>
<tr>
<th>Level</th>
<th>No proton transfer</th>
<th>Proton transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>3,160</td>
<td>16,776</td>
</tr>
<tr>
<td>Absent</td>
<td>357 (11.3)</td>
<td>2,450 (14.6)</td>
</tr>
<tr>
<td>Descriptive</td>
<td>1,653 (52.3)</td>
<td>9,227 (55.0)</td>
</tr>
<tr>
<td>Foundational</td>
<td>996 (31.5)</td>
<td>4,451 (26.5)</td>
</tr>
<tr>
<td>Complex</td>
<td>154 (4.9)</td>
<td>648 (3.9)</td>
</tr>
</tbody>
</table>
In summary, we established consistency between rubric users with an interrater reliability investigation and showed utility of the rubric in a variety of contexts (i.e., Organic Chemistry 1 and Organic Chemistry 2 courses), with a variety of reaction mechanisms including varied nucleophile types and reaction families, and with two variations of the constructed response prompt.

Table 4.9. Distribution of levels of explanation sophistication by prompt type

<table>
<thead>
<tr>
<th>Level</th>
<th>Original</th>
<th>More Cued</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>15,654</td>
<td>4,282</td>
</tr>
<tr>
<td>Absent</td>
<td>2,379 (15.2)</td>
<td>428 (10.0)</td>
</tr>
<tr>
<td>Descriptive</td>
<td>8,309 (53.1)</td>
<td>2,571 (60.0)</td>
</tr>
<tr>
<td>Foundational</td>
<td>4,354 (27.8)</td>
<td>1,093 (25.5)</td>
</tr>
<tr>
<td>Complex</td>
<td>612 (3.9)</td>
<td>190 (4.5)</td>
</tr>
</tbody>
</table>

4.6 Case study – Using the rubric in teaching practice

In this section, we offer a case study (Woodside, 2010; Patton, 2015; Yin, 2018) on how this rubric could be used within the context of a course in place of a conventional “implications for instructors” section. Our goal is to demonstrate how assessment results from the rubric can formatively and iteratively inform learning and instruction by developing reflective educators (Henderson et al., 2011).

For this case study, we chose an Organic Chemistry 1 course that was coordinated among two instructors in Fall 2021. Assessment data were collected from all sections of the course, including from sections taught by both instructors, at three time points (i.e., Time Points 1 through 3) throughout the semester, occurring after instruction on examination material but before an examination that assessed learning related to nucleophiles and reaction mechanisms (see Figure 4.2 for timeline). The course textbook for Fall 2021 was Klein’s Organic Chemistry, 4th edn (2021). The textbook takes a traditional approach to topic order based on functional groups (e.g., alkyl halides, alkenes, alkynes, alcohols, and ethers in Organic Chemistry 1) after initial skill development in chemical reactivity and mechanisms; in the textbook chapter presenting mechanisms, students learn to draw a mechanism and four characteristic arrow-pushing motifs:
nucleophilic attack, loss of a leaving group, proton transfer, and carbocation rearrangement. The reaction motifs structure is carried out throughout the course (into Organic Chemistry 2, as well) and aims to help learners develop more expert-like thinking by understanding the isolated pieces of a mechanism (Ferguson and Bodner, 2008; Grove et al., 2012a) and integrate this knowledge into the whole (Galloway et al., 2018) as new functional groups and mechanisms are studied.

Time Point 1 is a measure of where students are when they first encounter nucleophiles and reaction mechanisms; chronologically it is the third of five total course examinations (see Figure 4.2). Table 4.10 shows the distribution of levels using the nucleophile rubric for three different reactions of alkenes (i.e., a halogenation, hydration, and halohydrin formation) that were assessed at this time point. A participating student would have only seen and written about one of these reactions. For the course context and students, the Foundational level is considered as the level of satisfactory understanding; in other words, this is the level of explanation sophistication desired to be exhibited by students. As found in our analysis, the bulk of the responses are generally at the Descriptive level with fewer responses at the target Foundational level. We should note that students received extra credit for completing the assessment; and thus, motivation may have played a role in how and why they responded. Overall, the majority of students identified the nucleophile (i.e., at the Descriptive level or above) in each reaction, which is key.

Figure 4.2. Timeline of the case study: course content, examinations, and assessment time points
Table 4.10. Distribution of levels of explanation sophistication for Case Study Time Point 1: reactions of alkenes

<table>
<thead>
<tr>
<th>Level</th>
<th>Halogenation</th>
<th>Hydration</th>
<th>Halohydrin formation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>276</td>
<td>273</td>
<td>277</td>
<td>826</td>
</tr>
<tr>
<td>Absent</td>
<td>31 (11.2)</td>
<td>37 (13.6)</td>
<td>27 (9.7)</td>
<td>95 (11.5)</td>
</tr>
<tr>
<td>Descriptive</td>
<td>120 (43.5)</td>
<td>167 (61.2)</td>
<td>153 (55.2)</td>
<td>440 (53.3)</td>
</tr>
<tr>
<td>Foundational</td>
<td>116 (42.0)</td>
<td>57 (20.9)</td>
<td>93 (33.6)</td>
<td>266 (32.2)</td>
</tr>
<tr>
<td>Complex</td>
<td>9 (3.3)</td>
<td>12 (4.3)</td>
<td>4 (1.5)</td>
<td>25 (3.0)</td>
</tr>
</tbody>
</table>

Time Point 2 is characteristic wherein students have had the opportunity to reflect on the topics of the previous examination and potentially hone their skills relating to nucleophiles and reaction mechanisms as more reaction chemistry is taught; this time point occurs prior to the fourth of five course examinations (see Figure 4.2). For this assessment, students wrote about one of the four nucleophilic substitution reactions in Table 4.11 and one of the two reactions of an alkene or alkyne in Table 4.12; the letter designation A or B after the substitution reaction type designates two different variations of the reaction. Thus, each student wrote about a total of two reaction mechanisms; these were randomly presented to each student, and thus, students working at the same time on completing the extra credit assessment likely saw different reaction mechanisms or the same reaction mechanisms but in a different order.

Table 4.11. Distribution of levels of explanation sophistication for Case Study Time Point 2: substitution reactions

<table>
<thead>
<tr>
<th>Level</th>
<th>$S_N^{1A}$</th>
<th>$S_N^{1B}$</th>
<th>$S_N^{2A}$</th>
<th>$S_N^{2B}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>206</td>
<td>203</td>
<td>206</td>
<td>207</td>
<td>822</td>
</tr>
<tr>
<td>Absent</td>
<td>9 (4.4)</td>
<td>18 (8.9)</td>
<td>13 (6.3)</td>
<td>18 (8.7)</td>
<td>58 (7.0)</td>
</tr>
<tr>
<td>Descriptive</td>
<td>124 (60.2)</td>
<td>122 (60.1)</td>
<td>135 (65.5)</td>
<td>136 (65.7)</td>
<td>517 (62.9)</td>
</tr>
<tr>
<td>Foundational</td>
<td>68 (33.0)</td>
<td>54 (26.6)</td>
<td>51 (24.8)</td>
<td>52 (25.1)</td>
<td>225 (27.4)</td>
</tr>
<tr>
<td>Complex</td>
<td>5 (2.4)</td>
<td>9 (4.4)</td>
<td>7 (3.4)</td>
<td>1 (0.5)</td>
<td>22 (2.7)</td>
</tr>
</tbody>
</table>

Letter A and B designations refer to two different variations of each reaction.

Table 4.12. Distribution of levels of explanation sophistication for Case Study Time Point 2: reactions with alkenes and alkynes

<table>
<thead>
<tr>
<th>Level</th>
<th>Alkene hydrohalogenation</th>
<th>Alkyne alkylation</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>410</td>
<td>411</td>
</tr>
<tr>
<td>Absent</td>
<td>31 (7.6)</td>
<td>54 (13.1)</td>
</tr>
<tr>
<td>Descriptive</td>
<td>174 (42.4)</td>
<td>251 (61.1)</td>
</tr>
<tr>
<td>Foundational</td>
<td>162 (39.5)</td>
<td>102 (24.8)</td>
</tr>
<tr>
<td>Complex</td>
<td>43 (10.5)</td>
<td>4 (1.0)</td>
</tr>
</tbody>
</table>
For all reactions at this time point, the majority of responses are again at the *Descriptive* level demonstrating that students can identify the nucleophile with nucleophilic behavior described at a surface-level. Generally, there are fewer responses at the *Absent* level and more at the *Foundational* level when compared with Time Point 1 (Table 4.10) while noting that there is some variation depending on the exact reaction mechanism; for example, there is a larger number of responses in the *Complex* level for the alkene hydrohalogenation reaction which has aspects of halogenation reactions from Time Point 1. While we resist a causal explanation, per the assessment measures at Time Point 2, students’ level of explanation for nucleophiles appears to be becoming more sophisticated since Time Point 1; this is somewhat plausible given broader metrics of causality including temporality, time on task, etc. However, we still note that, on average, students are not at the designated *Foundational* level goal.

Time Point 3 is the final measure of students’ understanding of nucleophiles in reaction mechanisms at the end of the Organic Chemistry 1 course, occurring directly before the last of the five course examinations (see Figure 4.2). Here, students were asked to write about one of the three substitution reactions with alcohols (see Table 4.13), and one of the four reactions of an aldehyde or ketone to form an alcohol (see Table 4.14). Again, these distributions show that students are not at the target *Foundational* level. In fact, students may be at the same level of understanding for previously assessed mechanisms; for example, the reaction of an alcohol with hydrobromic acid (HBr) and the alkyne alkylation reaction are analogous to an *S_N2* reaction with respect to the nucleophilic attack by a lone-pair-containing, anionic nucleophile and has a similar distribution to those reactions. Additionally, these reactions assess understanding of reduction reactions (see Table 4.14) for the first time. The reduction reactions with lithium aluminum hydride (LiAlH₄) or sodium borohydride (NaBH₄) have a larger number of students at the *Absent* level than previously observed with other reaction types which may be the result of less time to understand this newly taught reaction type. These hydride reduction reactions contrast Grignard reactions,
Table 4.13. Distribution of levels of explanation sophistication for Case Study Time Point 3: substitution reactions with alcohols

<table>
<thead>
<tr>
<th>Level</th>
<th>Using HBr</th>
<th>Using PBr₃</th>
<th>Using SOCl₂</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>262</td>
<td>266</td>
<td>261</td>
<td>789</td>
</tr>
<tr>
<td>Absent</td>
<td>20 (7.7)</td>
<td>13 (4.9)</td>
<td>31 (11.9)</td>
<td>64 (8.1)</td>
</tr>
<tr>
<td>Descriptive</td>
<td>173 (66.0)</td>
<td>121 (45.5)</td>
<td>150 (57.5)</td>
<td>444 (56.3)</td>
</tr>
<tr>
<td>Foundational</td>
<td>64 (24.4)</td>
<td>111 (41.7)</td>
<td>75 (28.7)</td>
<td>250 (31.7)</td>
</tr>
<tr>
<td>Complex</td>
<td>5 (1.9)</td>
<td>21 (7.9)</td>
<td>5 (1.9)</td>
<td>31 (3.9)</td>
</tr>
</tbody>
</table>

Table 4.14. Distribution of levels of explanation sophistication for Case Study Time Point 3: reactions of aldehydes and ketones to form alcohols

<table>
<thead>
<tr>
<th>Level</th>
<th>LiAlH₄ with aldehyde</th>
<th>NaBH₄ with ketone</th>
<th>Grignard with aldehyde</th>
<th>Grignard with ketone</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>197</td>
<td>197</td>
<td>200</td>
<td>195</td>
<td>789</td>
</tr>
<tr>
<td>Absent</td>
<td>37 (18.8)</td>
<td>50 (25.4)</td>
<td>16 (8.0)</td>
<td>12 (6.2)</td>
<td>115 (14.6)</td>
</tr>
<tr>
<td>Descriptive</td>
<td>132 (67.0)</td>
<td>117 (59.4)</td>
<td>162 (81.0)</td>
<td>154 (79.0)</td>
<td>565 (71.6)</td>
</tr>
<tr>
<td>Foundational</td>
<td>25 (12.7)</td>
<td>26 (13.2)</td>
<td>19 (9.5)</td>
<td>26 (13.3)</td>
<td>96 (12.2)</td>
</tr>
<tr>
<td>Complex</td>
<td>3 (1.5)</td>
<td>4 (2.0)</td>
<td>3 (1.5)</td>
<td>3 (1.5)</td>
<td>13 (1.6)</td>
</tr>
</tbody>
</table>

which the latter are taught analogously, but has higher levels of explanation sophistication with a larger proportion of responses in the **Descriptive** level.

In summary, these assessment results show that the majority of students are not at the **Foundational** target level when explaining nucleophiles in the context of reaction mechanisms. One proposition is that students fail to describe or interpret chemical reactions due to insufficient practice with engaging with these rich descriptions in textbooks and course materials (McCollum and Morsch, 2022). As we will later note, there is an opportunity for more assessments throughout the learning experience to emphasize and model the desired learning level. Overall, we do observe a trend that explanations became more sophisticated over the semester-long course, which may be the result of continuing practice in writing about reaction mechanisms.

### 4.6.1 Reflective teaching practices

From a summative, post-semester reflection, these assessment results are an opportunity to consider bigger changes we might make to the course to better achieve our goal of a majority of students explaining nucleophiles at the **Foundational** level or higher.
The assessments, as noted, were given as extra credit assignments prior to each of the last three course examinations of the term. While the instructors fully expected students to develop a deep understanding of the reaction mechanisms, our assessment evidence suggests there is room for improvement. An advantage of using a rubric, such as the one reported herein, is that students are able to use the rubric to monitor their learning and instructors are able to use the rubric to assess learning (Panadero and Johnson, 2013; see users and uses in Dawson, 2017); thus, we need to share the rubric with students as part of the formative assessment and feedback process (see secrecy in Dawson, 2017). Additionally, exemplary responses are key to the learning process. We need to share quality responses for each level of the rubric that fit within the specific assessment item and context of our course (see exemplars in Dawson, 2017); more importantly, we need to communicate our Foundational level goal to students in the classroom and through examples. The rubric in and of itself can act as a centerpiece for feedback discussion. Instructor comments about a learner’s writing can aid in students’ explanations (see accompanying feedback information in Dawson, 2017).

From a formative perspective, the rubric and associated assessment data provide a means to address learning in the ongoing context of the course. Having students complete such writing tasks throughout instruction, modeling expected language in instruction (Andrade, 2005), and using such items on for-credit assessment would further emphasize the importance of this type of learning. As a particular example based on the assessment data we presented above, some of the responses at Time Point 1 are at the Absent level; this is a point of reflection for us, the educator, to ensure we note and label parts of reaction mechanisms (e.g., nucleophile, electrophile, leaving group) and mechanistic steps (e.g., nucleophilic attack, loss of leaving group, proton transfer) as they are presented, and used, and repeated, and appear in homework, etc. making sure to explicitly note the nucleophile in every context. We, as practicing chemists, do not need to be this explicit; however, to a learner, an overemphasis can be key. In considering moving from Descriptive to Foundational, educators have an opportunity to again be routine, in ad
nauseum, in making the implicit explicit: we should note the most nucleophilic part of the nucleophile (e.g., the atom and lone pair, sigma bond, pi bond) in each reaction mechanism and mechanistic step, as appropriate. Instances of being “overly explicit” are not meant to be made for the entirety of the year-long course; at some point, learning should be expected. However, if we assume that some level of explicitness is made and thus some level of scaffolding is provided to students, our work and the research literature we summarized herein would support the notion that we are not yet doing enough cueing or scaffolding to reach our desired levels of learning. Our goal then is that later in the semester, when we notice more students are at the Foundational level (i.e., Time Point 2), we can better introduce the “next” level of learning and emphasize electrostatic interactions using differences in electronegativity, electron density, and partial charges to model language we would expect at the Complex level.

4.6.2 Implications for teaching practice

Throughout the collection of the data reported in the broader manuscript and the data specifically reported in this case study, students’ only reference for developing an importance of this type of written explanation assessment was in completing the extra credit assignments that had these prompts; if writing about the “how and why?” is desirable, instructors must provide students with opportunities to practice and appropriately evaluate these skills. Thus, we must be intentional about using explanations in our assessments (formative and summative) and correspondingly, using this rubric and other means to clearly communicate our learning expectations (Cooper, 2015; Stowe and Cooper, 2017).

These assessment data provide a means to consider broader teaching, learning, and assessment transformations, and more importantly, conduct quasi-experimental investigations in our course to make decisions about what those transformations could be. For example, the same prompt and reaction mechanism could be used at the beginning of the semester (or when nucleophiles and reaction mechanisms are introduced) and then again at the end of the semester;
the rubric could then assist in longitudinal evaluation. To make that example slightly more explicit, a bimolecular nucleophilic substitution reaction could be used to evaluate nucleophile understanding shortly after instruction or near an examination of their understanding; then, towards the end of the term, the same reaction could be used, and the results of the second assessment can be used to make claims about progression of explanation sophistication of nucleophiles. Additionally, modification of teaching strategies could be rooted in analyzing distributions of responses to discern variation in reactions or reaction types. For example, the distribution between the reactions of aldehyde and ketones with hydride reagents compared with Grignard reagents as noted previously; these are similar mechanisms, and yet, distributions varied and thus as instructors, we have an opportunity to use the assessment to talk about those similarities. Additionally, the prompt could be varied throughout the course to be more or less cued, as needed, or to fit the specifications of our expectations and learning objectives of the course. For example, as we have presented in this case study, our expectations were for Organic Chemistry 1 students; expectations for students enrolled in a non-majors organic chemistry course or for a general, organic, and biochemistry course may be different.

4.7 Limitations

There are three limitations of note for this work: (i) instructor implementation of the rubric, (ii) homogeneity of the sample, and (iii) limited number of prompt variations.

4.7.1 Instructor implementation

Differences in instructor expectations may yield differing implementations of the rubric when evaluating level of explanation sophistication. We note several differences in this section that emerged during the interrater process and in peer debriefing discussions; examples of such disagreements are:
Authors had contrasting beliefs regarding the strength of a nucleophile. For example, for a reduction reaction, a response stated, “the ketone is reduced by a moderately strong hydrogen nucleophile … there are two enantiomer products because originally the molecule was planar and could be attacked either from on top or from bottom.” This response was classified in the Descriptive level by one author that believed the strength of a nucleophile was out of the scope regarding nucleophilic behavior in reaction mechanisms and another author that considered descriptions about the strength of nucleophiles classified this response in the Foundational level.

Beliefs about the strength of nucleophiles evolve into conceptions about what makes a “good” nucleophile. A response could have stated that the “oxygen acts as a nucleophile and performs a substitution reaction … and attacks the beta-carbon” and thus be classified at the Descriptive level; however, the added language for an epoxidation reaction that says, “hydroxide has a high affinity for hydrogen [on the alcohol] and removes it. The oxygen becomes a good nucleophile and attacks the beta-carbon” could be argued to be of higher understanding of the nucleophile and classified at the Foundational level.

The Foundational level is characterized as explanations that demonstrate understanding that electrons are central to nucleophilic behavior and arrows represent the movement of electrons. For example, consider this partial response to an aldol condensation mechanism:

“The base comes in and grabs an alpha proton to carry out deprotonation of the alpha carbon on the right because it is more stable. The electrons of the carbonyl group move up to the oxygen forming a negative charge and from deprotonation a double bond is formed between the carbonyl carbon and the alpha carbon or rather the alpha and beta carbon. This intermediate now acts as the nucleophile and attacks another ketone molecule at the carbonyl carbon because it has electrophilic character, pushing electrons up onto the oxygen which bears a negative charge.”

This response has language concerning electrons before and after the nucleophilic step, and thus demonstrates sufficient understanding that nucleophilic behavior and mechanistic arrows represent the movement of electrons at the Foundational level. However, the rubric could
be applied such that responses must explicitly state the use of electrons in the nucleophilic step to be at the *Foundational* level; because this response does not make that explicit, the response could be classified at the *Descriptive* level.

Electrophilic aromatic substitution reactions, for example, resulted in differing applications of the rubric. In a response to an electrophilic aromatic chlorination, a student wrote, “formation of the electrophile occurs because of the nucleophilic attack from the lone pair [from] chlorine to AlCl₃.” This could be classified at the *Descriptive* level or at the *Foundational* level; we noted in prior work (Dood et al., 2018; Yik et al., 2021) that some students used nucleophile–electrophile interactions to describe the behavior of Lewis bases and acids, respectively.

These examples show how the rubric could be interpreted and applied differently based on one’s ideas about nucleophiles. This, however, when made transparent, can catalyze discussions between educators, exemplify the nature of chemistry and science, and clarify course learning objectives. Instructors can have differing opinions rooted in their beliefs and expectations for students; in the end, though, instructors have a responsibility to make those expectations clear to students and consistently evaluate those expectations accordingly.

### 4.7.2 Sample homogeneity

Responses used in this study were written by students taught by one of four instructors at a single institution that uses a single curriculum. We should note that those instructors routinely coordinate their course sections; thus, an analogous curriculum is enacted in the courses from which data were collected, with students having analogous learning experiences. We therefore caution against blind adoption of the rubric without consideration of the learners, curriculum, and context. Adaptation of the rubric may be necessary in the context of other curricula such as a spiral organic curriculum (Grove et al., 2008), Mechanisms before Reactions curriculum (Flynn and Ogilvie, 2015), or the Organic Chemistry, Life, the Universe and Everything (OCLUE) curriculum (Cooper et al., 2019). We note that these curricula are associated with different
reasoning frameworks; for example, the causal mechanistic reasoning framework (Cooper et al., 2016; Crandell et al., 2019, 2020) is associated with OCLUE.

We also note the emergence of discussions related to equity in STEM when assessments require written explanation or other written work; this is particularly important when the language required to be used in the assessment is not the learner’s first language. For example, Deng et al. (in press) reported that students who learned English-as-an-additional language demonstrated lower levels of explanation sophistication compared to their English-first language counterparts.

At our study site (i.e., the University of South Florida), we are aware that a growing number of learners are refining their English skills (i.e., the language of instruction at our study site) while simultaneously learning the language of organic chemistry, chemistry, science, etc. As we further develop our rubric and understand the utility of using it in our enacted curriculum, we intend to be attentive to the barriers to learning, access, and retention consequences that our assessments may have.

4.7.3 Limited prompt variations

While our work reports the wide utility of the rubric to evaluate the level of explanation sophistication for nucleophiles with nearly 20,000 responses, these responses were gathered using two almost identical prompts (see Table 4.3). These limited prompt variations restrict instructors to rely on two similarly worded, formulaic prompts that focus on describing what is happening and why it is happening on a molecular level. Our decision to modify the Original prompt to the More Cued prompt was based in work to understand students’ use of causal mechanistic reasoning (Cooper et al., 2016; Crandell et al., 2019, 2020). Our prompts have parallel structure: “Consider the mechanism for the [reaction name/type] to form [product]” or “consider the mechanism for the [reaction] between [substrate] and [reagent] to form [product].”

We advocate for instructors to use this general form when first applying our reported rubric; such
use can provide a foundation to consider more situational variations to match the learning goals and context in which the rubric is used.

Future endeavors need to consider how cued, scaffolded, and other varied prompts are associated with the types and kinds of responses obtained and how explanations might be integrated into a course. In our work, we demonstrate the generalizability of this rubric with two prompt variations (see Table 4.3). We can envision prompts that further cue students into activating resources about electrostatic interactions or scaffold responses such that students first identify reaction components before describing the how and why of a reaction mechanism. Writing assignments that ask students to explain reaction mechanisms may take the form of formative assessments or through homework. These assignments could also be embedded into larger summative projects; one example could include a literature project at the end of the year-long organic chemistry course sequence where students find a reaction covered in the course and explain the mechanism for the noted transformation; the literature project could have multiple components wherein explanation of a reaction mechanism is only a component to the summative assessment experience.

4.8 Implications

We report in this paper a rubric for evaluating student level of explanation sophistication for a single reaction mechanism concept: nucleophiles. Using the same sophistication levels (i.e., absent, descriptive, foundational, and complex), the overall rubric structure could be applied to other reaction mechanism concepts (e.g., electrophiles, leaving groups, proton transfers). We envision a broad set of rubrics to evaluate explanations of reaction mechanisms with every possible component of a reaction mechanism or holistically (i.e., a single level for the entirety of a response).

Data used in this study were collected over several years and give a broad picture of how students understand and explain nucleophiles throughout the course. At the same time, our data
do not show how this understanding is associated with student achievement (e.g., course grades) and development of understanding over time. Studies have shown that a rubric used to evaluate student work on constructed response items was associated with higher student achievement possibly due to operationalizing definitions of achievement that students could understand (Shafer et al., 2001). We do not yet report external measures to demonstrate increase in student understanding (e.g., examination scores) which could serve as an additional metric to evaluate courses and curricula. Furthermore, as mentioned in our Case Study discussion, assessment of understanding in a longitudinal setting would provide a stronger reference point for course evaluation and revision. A longitudinal approach could take form throughout a single-semester course, a year-long course, or across the undergraduate curriculum. In combination with other evaluative criteria, understanding across time may serve as a useful tool for evaluating curricula and teaching practices.

Scoring student written responses is tedious, especially for large enrollment courses such as Organic Chemistry. Computer-assisted predictive scoring models can be used to evaluate written assessments. The speed to quickly evaluate many responses with predictive scoring models makes written responses a viable option for in-class use (e.g., Haudek et al., 2011, 2012; Prevost et al., 2013; Dood et al., 2018, 2020a; Noyes et al., 2020; Yik et al., 2021). While predictive scoring models are becoming more commonplace, these models are specific to single assessment items; previous work (Yik et al., 2021), though, has demonstrated that generalized predictive models can be developed. A limitation, however, of that work is the binary classification for correct or incorrect/non-use of a very specific conceptual model (Yik et al., 2021). In moving toward more meaningful assessment measures, researchers must aim to (a) develop generalized predictive models to eliminate limitations for item-specific measures, and (b) have means to evaluate responses on a continuum or scale, such as using the rubric reported herein.
4.9 Conclusion

We report a generalized rubric for level of explanation sophistication for nucleophiles in organic chemistry reaction mechanisms. We characterized this rubric with a set of design elements for a quality rubric. Our results suggest that this rubric is generalizable throughout a year-long organic chemistry course, across multiple prompt types, with different nucleophile types, and in a variety of reaction families and types. We also present a case study where the rubric was used to classify responses which informed reflective practices and actionable items to enhance teaching and learning experiences. This nucleophile-focused rubric is a starting point for other reaction mechanism components (e.g., intermediates or electrophiles) that can be applied in assessing curricula and teaching practices. Future work might involve using the levels of explanation sophistication described in this rubric as categories for a machine learning-based predictive scoring model.

As one student reflected in their explanation of a reaction mechanism, “I realize that if I had a better understanding of why the reactants are interacting it would make the mechanism easier to comprehend instead of just memorizing the mechanism.” Our overarching principal goal is to create environments that catalyze such learning. This begins with clear outcomes as expressed in our reported rubric, use of that rubric to assess learning, and revising and refining teaching practices to promote such learning.

4.10 Conflicts of interest

There are no conflicts to declare.

4.11 Author contributions

BJY and JRR conceived the project. BJY, AJD, and JRR collected data assisted by DCR and KBF. BJY, AJD, SJHF, and JRR conceptualized the rubric. BJY cleaned, prepared, and
analyzed the data. BJY and JRR wrote the manuscript. All authors read, edited, and approved the final manuscript.

4.12 Acknowledgments

We would like to thank all the students who participated in our study. We would like to thank Frankie Costanza (University of South Florida) for providing access to students for data collection. Author BJY would like to thank the Office of Graduate Studies at the University of South Florida for support through a Dissertation Completion Fellowship.

4.13 References


Becker, N., Noyes, K. and Cooper M. (2016). Characterizing students’ mechanistic reasoning about london dispersion forces. *Journal of Chemical Education* 93, 1713-1724. doi: 10.1021/acs.jchemed.6b00298


Cooper, M. M. (2015). Why ask why?. *Journal of Chemical Education* 92, 1273-1279. doi: 10.1021/acs.jchemed.5b00203


Crandell, O. M., Lockhart, M. A. and Cooper, M. M. (2020). Arrows on the page are not a good gauge: Evidence for the importance of causal mechanistic explanations about nucleophilic substitution in organic chemistry. *Journal of Chemical Education* 97, 313-327. doi: 10.1021/acs.jchemed.9b00815


DeFever, R. S., Bruce, H. and Bhattacharyya, G. (2015). Mental rolodexing: Senior chemistry majors’ understanding of chemical and physical properties. *Journal of Chemical Education* 92, 415-426. doi: 10.1021/ed500360g


143


Šimkovic, M. and Träuble, B. (2019). Robustness of statistical methods when measure is affected by ceiling and/or floor effect. *PLOS ONE* 14:e0220889. doi: 10.1371/journal.pone.0220889


Chapter 5

Evaluating the impact of malleable factors on percent time lecturing in gateway chemistry, mathematics, and physics courses

5.1 Note to Reader

This chapter is a published manuscript in the International Journal of STEM Education. This chapter is reproduced from:
with permission from the Springer Nature. Further copyright information can be found in Appendix E.3.

This work has co-authors. Charles Henderson, Melissa H. Dancy, Estrella Johnson, Jeffrey R. Raker, and Marilyne Stains conceived the larger survey project and Naneh Apkarian aided in designing the survey. Naneh Apkarian cleaned and compiled the data. Jeffrey R. Raker, Naneh Apkarian, and Marilyne Stains contributed to analyzing the data. Jeffrey R. Raker is the principal investigator for this study.

5.2 Abstract

Background: Active learning used in science, technology, engineering, and mathematics (STEM) courses has been shown to improve student outcomes. Nevertheless, traditional lecture-orientated approaches endure in these courses. The implementation of teaching practices is a
result of many interrelated factors including disciplinary norms, classroom context, and beliefs about learning. Although factors influencing uptake of active learning are known, no study to date has had the statistical power to empirically test the relative association of these factors with active learning when considered collectively. Prior studies have been limited to a single or small number of evaluated factors; additionally, such studies did not capture the nested nature of institutional contexts. We present the results of a multi-institution, large-scale (\(N = 2,382\) instructors; \(N = 1,405\) departments; \(N = 749\) institutions) survey-based study in the United States to evaluate 17 malleable factors (i.e., influenceable and changeable) that are associated with the amount of time an instructor spends lecturing, a proxy for implementation of active learning strategies, in introductory postsecondary chemistry, mathematics, and physics courses.

Results: Regression analyses, using multilevel modeling to account for the nested nature of the data, indicate several evaluated contextual factors, personal factors, and teacher thinking factors were significantly associated with percent of class time lecturing when controlling for other factors used in this study. Quantitative results corroborate prior research in indicating that large class sizes are associated with increased percent time lecturing. Other contextual factors (e.g., classroom setup for small group work) and personal contexts (e.g., participation in scholarship of teaching and learning activities) are associated with a decrease in percent time lecturing.

Conclusions: Given the malleable nature of the factors, we offer tangible implications for instructors and administrators to influence the adoption of more active learning strategies in introductory STEM courses.

5.3 Introduction

It is established that using active learning instructional approaches (i.e., less time spent lecturing) are associated with higher conceptual understanding and persistence in postsecondary (i.e., undergraduate) STEM courses (i.e., undergraduate) STEM courses (Springer et al., 1999; Lorenzo et al., 2006; Ruiz-Primo et al., 2011; Freeman et al., 2014; Theobald et al., 2020). This
result holds across a variety of class sizes, disciplines, and levels (Freeman et al., 2014). Importantly, many studies on the use of active learning have also demonstrated a reduction in achievement gaps for minoritized populations (i.e., low-income students or underrepresented minorities) in STEM (Lorenzo et al., 2006; Kogan and Laursen, 2014; Theobald et al., 2020). In particular, active learning has been shown to reduce achievement gaps in exam scores by a third between minoritized groups and non-minoritized groups, narrow gaps in passing rates by nearly half (Theobald et al., 2020), decrease failure rates (as defined by the percentage of students receiving a D or F grade, or withdrawing from the course), and increase achievement across all STEM disciplines when compared to traditional lecture courses (Freeman et al., 2014). Despite evidence of the benefit of active learning strategies (Freeman et al., 2014; Haak et al., 2011; Ballen et al., 2017; Styers et al., 2018; Harris et al., 2020; Theobald et al., 2020), observation-based studies have confirmed that lecture-oriented pedagogical approaches remain as significant components in most STEM courses (Stains et al., 2018), which may be due in part to the lack of departmental norms for using research-based instructional methods (Henderson and Dancy, 2007; Shadle et al., 2017), faculty reward structures (Brownell and Tanner, 2012; Michael, 2007; Shadle et al., 2017), and student resistance (Henderson and Dancy, 2007; Michael, 2007; Shadle et al., 2017).

Prior work in understanding the adoption of active learning strategies has strived towards identifying factors or has considered a single or small number of associated factors. For example, studies have used faculty discussions (Shadle et al., 2017), interviews (e.g., Henderson and Dancy, 2007; Oleson and Hora, 2014), and survey methodologies (e.g., Lund and Stains, 2015; Gibbons et al., 2018; Johnson et al., 2019) to identify factors related to the influence of pedagogical decisions; large class sizes, fixed-seat classroom layouts, lack of pedagogical knowledge in research-based instructional practices, and insufficient faculty assessment methods and processes have all been reported as barriers to uptake of active learning strategies. Lund and Stains (2015) explored departmental influences (e.g., perceived norms towards teaching)
and classroom influences (e.g., class size and layout) on adopted pedagogies also finding that departments place high expectations on research output and such classroom elements further restrict instructor teaching. Studies by Gibbons et al. (2018) and Popova et al. (2020) explored the link between postsecondary instructors’ thinking and enacted instructional practices, showing a connection between the two. Later, Popova et al. (2021) found evidence for the interconnectedness of instructors’ beliefs about teaching and learning with personal (e.g., nature and extent of instructors’ preparation and learning efforts) and contextual factors (e.g., course, department, and broader cultural contexts).

However, due to small sample sizes and lack of statistical power, these studies could not empirically test the relative association of these factors when considered all together with active learning. In addition, these studies did not account for the nested nature of their institutional contexts (i.e., instructors within departments within institutions). Therefore, a large-scale, multidisciplinary study of malleable factors (i.e., things that can be changed and altered) related to adoption of such active learning pedagogies in postsecondary STEM courses is needed to complement the research literature and provide further opportunity for actionable changes at instructor, department, and institution levels.

The study reported herein focuses specifically on instructors of introductory chemistry, mathematics, and physics and the malleable factors that influence their uptake of active learning practices as measured by a proxy, i.e., percent time not lecturing. Specifically, we use multilevel modeling to account for the nested nature of our data by discipline and institution, and evaluate 17 factors situated within the three categories (i.e., contextual factors, personal factors, and teacher thinking) as to their relationship with reported percent time lecturing. The teacher-centered systemic reform model (Woodbury and Gess-Newsome, 2002; Gess-Newsome et al., 2003) suggests that factors within these categories are related to enacted teaching practices. In the next section, we describe this conceptual framework and detail the literature that report these 17 factors related to the adoption of active learning.
5.3.1 Conceptual framework

Research on pedagogy adoption and colloquial anecdotes about why instructors choose to enact active learning pedagogies informed the selection of modeled factors in our study. The teacher-centered systemic reform model was developed from an exhaustive review of the literature as a mechanism to understand the evolution of classroom practices as a result of reform initiatives (Woodbury and Gess-Newsome, 2002) and was later modified to better reflect the nature of teaching in a university context (Gess-Newsome et al., 2003). This framework is useful as it considers the situational teaching contexts along with an instructor’s educational influences and their beliefs about teaching and learning all within a complex educational system.

Fundamentally, the TCSR framework is focused on teacher change as the source of grander changes within the larger institutional system. An instructor is the ultimate authority of the enacted practices that occur in a classroom and is aligned with the TCSR model’s theoretical underpinnings that instructors’ beliefs influence their practices as embedded with a larger system, such as in classrooms, departments, institutions, and disciplines (Woodbury and Gess-Newsome, 2002). This study aims to quantify the extent that malleable factors have on the uptake of active learning and the TCSR framework is best suited, since its focus is on the instructor and their personal and teaching contexts which are the most malleable along with nested contexts (i.e., instructors within departments that are within institutions) in which teaching reform occurs.

According to Gess-Newsome et al. (2003), the TCSR model for a university context has three broad categories: contextual factors, personal factors, and teacher thinking factors. Situated within contextual factors are the broader cultural context (e.g., teacher development and teaching materials), school context (e.g., institution type, physical space, and technology), department and subject area content (e.g., department and cultural norms and teacher’s class load), and classroom context (e.g., class size and physical organization of the room). Situated within personal factors are instructors’ demographic profile, types and years of teaching experience, and nature and extent of teacher preparation and continued learning efforts. Situated within teacher
thinking factors are instructors’ sense of dissatisfaction with current practices; and knowledge and beliefs about teachers and teachers’ roles, students and learning, schooling and schools, and content being taught.

Extensive literature used to develop the TCSR model aims at capturing as many of the intricacies that can be situated under the broad factors that comprise the general context of reform. Nevertheless, the framework does not fully capture all of the complexities of higher education institutions when the model was used to frame the study reported herein. As a consequence of the literature review on malleable factors that affect pedagogical change reported herein, our conceptualization of the TCSR model necessitated further modifications. For example, factors were found that necessitated inclusion in our study as distinctly different from department contexts (e.g., discipline and cultural norms), such as tenure status (e.g., non-tenure-track lecturers, tenure-track faculty, and tenured faculty), teaching load, and instructors’ teaching evaluation. It is crucial to delineate department appointment expectations as a subcategory under contextual factors apart from department contextual factors to aid in accounting for the different department and institutional policies that differ with the distinct instructional positions (e.g., lecturers versus tenure-track professors). Higher education systems undergo gradual change and theoretical frameworks on change theories need to be explored and reevaluated in light of new research. Findings since the original conceptualization necessitated modification of the TCSR model with the ones we present in this study.

Our study intends to evaluate the effects of malleable factors related to adoption of active learning pedagogies when controlling for institutional and disciplinary differences. Thus, non-malleable factors, such as race/ethnicity, are not included. In addition, while this study attempts to account for as many of the malleable factors reported in the literature, it is not possible to (1) ask respondents about all possible aspects of the TCSR model in a survey and (2) statistically test all factors that may be present in graphical representations of the TCSR model (cf. Woodbury and Gess-Newsome, 2002; Gess-Newsome et al., 2003); the associated sample size requisite
for sufficient statistical power grows with a larger number of tested factors. To balance complexity and parsimony, we include malleable factors found in the literature that have been previously cited as barriers to implementation or as reasons for uptake of active learning strategies. While our study contains factors from all three broad categories (i.e., contextual, personal, and teacher thinking) and all of their subcategories, some factors identified in the model (cf. Gess-Newsome et al., 2003) are either not malleable (e.g., physical location and college president) or difficult to quantify in a statistical model (e.g., instructor’s daily/weekly schedule and student personal expectations).

In this study, we include malleable factors that have been found and discussed in many and different STEM fields as specific disciplines may lack literature in that area. While STEM fields may show some disciplinary differences, these factors can be assumed to affect all STEM fields to some extent (Lund and Stains, 2015). In the next sections, we describe the evidence-based factors grounded in the literature that have been found to affect the uptake of active learning under each of the three broad categories in the TCSR model that were tested in this study (see Figure 5.1).

Figure 5.1. Conceptualization of the Teacher-Centered Systemic Reform (TCSR) model for change in higher education with control variables (discipline and highest degree awarded) and malleable factors included in this study
5.3.1.1 Contextual factors

Contextual barriers to instructional change have been widely studied (e.g., Fairweather and Rhoads, 1995; Hora, 2012; Lund et al., 2015; Shadle et al., 2017; Stains et al., 2018). In this study, three types of contextual factors were explored: department characteristics, department appointment expectations, and classroom contexts.

5.3.1.1.1 Department characteristics

Department characteristics that have been reported to be associated with the adoption of active learning include: (1) discipline (i.e., chemistry, mathematics, or physics) and (2) the highest degree the department awards (i.e., associates, bachelor’s, or graduate). While these factors are not malleable, they are included as control variables for departmental characteristics when testing the association between malleable factors. It has been reported that instructional practices differ between STEM disciplines (Fairweather and Rhoads, 1995; Hora and Anderson, 2012; Lund et al., 2015; Stains et al., 2018). Several studies have indicated that the balance between teaching and research at one’s institution and department impact how teaching is approached. The highest degree awarded in the department has been shown to be a viable proxy for the extent of focus a department places teaching versus research (cf. Cox et al., 2011; Srinivasan et al., 2018). For example, instructors that teach in departments with graduate degree offerings are presumed to have a greater focus on research than their counterparts in departments with associate degree programs.

5.3.1.1.2 Department appointment expectations

Department appointment expectations that have been reported to be associated with the adoption of active learning include: (1) teaching load, (2) tenure status, (3) the role of student evaluations, and (4) the role assessment of teaching in review, promotion, or tenure. While the department may have set teaching loads and standards for the role of student evaluations and
assessment of teaching performance in review, promotion, or tenure, there is the possibility, especially in larger departments with an array of teaching personnel, that evaluation may be unique or differentiated by appointment.

Teaching load has been reported to be associated with teaching practices, with higher teaching loads being attributed to an instructor being pressed for time (Hora, 2012). Lack of available time devoted to teaching activities has the potential to result in a lack of innovative pedagogies. Henderson and Dancy (2007) noted that one of the largest barriers reported by physics instructors was a heavy teaching load. In this study, we separate teaching load as a distinct factor from tenure status and institution type. From a broader view, teaching loads might be an indicator of tenure status; for example, an instructor with no opportunity for tenure (e.g., a lecturer or visiting instructor) may have higher teaching loads. In addition, a teaching load could correlate with institution type; an instructor at an institution with a larger teaching focus (e.g., a primarily undergraduate institution; PUI) may have a higher teaching load. However, upon further inspection, different instructional positions can hold different tenure statuses and be at different institution types. For example, a tenured professor at a PUI may have a large teaching load or a lecturer at a large research-intensive institution may have a high teaching load.

In this study, respondents are grouped into three appointment categories: (1) no opportunity for tenure, (2) tenure-track, and (3) and tenured. Tenure status has been shown to have an association with the amount of adoption of research-based instructional strategies (Landrum et al., 2017; Shadle et al., 2017). Those teaching undergraduate STEM courses, especially those with the privilege of obtaining tenure at research-intensive institutions, have the ability to identify as both an educator/teacher and a researcher. Implementation of active learning strategies may more time consuming (Beatty et al., 2005; Drinkwater et al., 2014), thus, tenure-track and tenured faculty members, in theory, have to weigh time spent on teaching and research, among other responsibilities. Fairweather (2008) reported that untenured faculty members are least likely to be productive in both teaching and research compared to being productive in either
teaching or research; therefore, tenure status is suggested to be influential in instructional decisions. Landrum et al. (2017) corroborates this idea and found significant differences in evidence-based instructional practice adoption between tenure/tenure-track faculty and their non-tenure-track counterparts, with the former reporting significantly higher use of these practices. Being tenured can allow for more freedom and flexibility to use innovative teaching methods (Hora, 2012).

Instructor’s perceived value of how their department or institution values teaching, both in their assessment of their teaching and from student evaluations, plays an important role in the instructor’s role as a teacher; studies have reported that if it was the norm for instructors in a department to integrate research-based methods into their teaching then it was easier for others to do so and that there is no uniform method of evaluating and rewarding one’s teaching (e.g., Prosser and Trigwell, 1997; Gess-Newsome et al., 2003; Henderson and Dancy, 2007; Walczyk et al., 2007; Seymour et al., 2011; Brownell and Tanner, 2012; Hora and Anderson, 2012; Sturtevant and Wheeler, 2019). Lund and Stains (2015) reported that a substantial number of STEM instructors describe departmental and tenure pressures as influential to their teaching practices, while few faculty say student evaluations of teaching influence their teaching approaches (Erdmann et al., 2020). In a study of mathematics instructors, Johnson et al. (2018) reported that while the most popular reason reported for not attempting instructional change was a lack of time for course redesign, roughly 20% of respondents reported that they believed their departments would not support them and that instructional change would not be valued in their annual review, promotion, or tenure process.

5.3.1.1.3 Classroom contextual factors

Classroom contextual factors include characteristics of the classroom learning environment that influence pedagogical decisions. Classroom contextual factors that have been
reported to be associated with the adoption of active learning include: (1) class size, (2) classroom layout, and (iii) decision making authority over instructional choices.

Class size and classroom layout have been reported to be influential and strong barriers to the implementation of active learning strategies (e.g., Prosser and Trigwell, 1997; Henderson and Dancy, 2007; Michael, 2007; Lund et al., 2015; Lund and Stains, 2015; Shadle et al., 2017; Sturtevant and Wheeler, 2019). STEM instructors cite factors such as large class sizes (i.e., over 100 students) as a reason why they have not chosen to adopt interactive teaching methods (Hora and Anderson, 2012; Shadle et al., 2017) and is corroborated by a significant correlation between class size and percent time spent lecturing (Smith et al., 2014). However, fixed classroom layouts are not prohibitive to enacting active learning pedagogies; in an observation study by Lund et al. (2015), various levels of student–student interactions (i.e., active learning) were observed in large classes taught in fixed-seat lecture halls. This may indicate that there is a relationship between class size and classroom layout.

In a case study of doctoral degree granting institutions, it was found that course coordination was one of the seven features that contributed to successful calculus programs (Rasmussen et al., 2014; Bressoud and Rasmussen, 2015;). Rasmussen et al. (2019) reported that high-quality active learning can be attributed, in part, by the support systems that course coordination affords. Communication channels, both formal and interpersonal, are key to the diffusion of innovations (Rogers, 2003), for example the dissemination of research-based instructional practices. In the context of adopting innovative teaching practices, such communication could be facilitated through a course coordinator with positional and personal instructional influence (Apkarian and Rasmussen, 2017; Lane et al., 2019; Bazett and Clough, 2021; Golnabi et al., 2021).

By leveraging common tools and resources, encouraging collaboration and shared objectives, and promoting professional development, course coordinators can act as change agents (Williams et al., 2022). Coordination can catalyze community-building and collaboration
between instructors; information was also shared at department meetings and retreats which led to meaningful conversations centered around teaching (Bazett and Clough, 2021; Williams et al., 2022). Visitors, instructors, teaching assistants, adjuncts, and lecturers (VITAL; Levy, 2019) that teach sections of coordinated mathematics courses have access to course coordinators to discuss pedagogical approaches and active learning activities (Golnabi et al., 2021). In addition, coordinators of active learning courses in mathematics have also been shown to utilize local data (e.g., student performance data, grades in subsequent courses, and student-generated data) to inform curriculum and pedagogy (Martinez and Pilgrim, 2021). In physics, co-teaching between a new instructor and an experienced instructor that uses active learning showed immediate uptake of these teaching practices by the new instructor and a positive shift in their beliefs and intentions of using these strategies in the future (Henderson et al., 2009).

5.3.1.2 Personal factors

Personal factors include the nature and extent of teaching preparation and experience, and teaching-related training. Personal factors that have been reported to be associated with adoption of active learning that were explored in this study include: (1) experience with research-based instructional strategies (RBIS) as a student, (2) completion of teaching-focused coursework, (3) participation in new faculty experiences or workshops, and (4) participation in scholarship of teaching (SOTL) or discipline-based education research (DBER).

Previous experience with RBIS as a student impacts instructional decisions (Oleson and Hora, 2014; Lund and Stains, 2015). In a study of upper-division mathematics instructors, it was reported that, by far, the two most influential factors on instructional practices were both their experiences as a teacher and as a student (Fukawa-Connelly et al., 2016). In interviews of STEM instructors, many cite that knowledge regarding teaching explained the selection of teaching techniques including active learning strategies (Oleson and Hora, 2014). In a study of biologists,
chemists, and physicists, instructors who had experienced RBIS as a student were more likely to implement RBIS in their own teaching (Lund and Stains, 2015).

Additionally, instructors’ previous coursework, which can be a part of their teacher preparation, influences teaching decisions (Windschitl and Sahl, 2002; Lotter et al., 2007; Southerland et al., 2011a, 2011b; Hora, 2012). It has been reported that doctoral training influences teaching approaches for some individuals (Southerland et al., 2011a, 2011b; Hora, 2012), and more specifically, university coursework also alters teaching pedagogy (Windschitl and Sahl, 2002).

Dissemination of research-based instructional strategies through new faculty experiences and workshops have been reported to increase instructor awareness and inclusion of these instructional strategies in many STEM fields (e.g., Henderson, 2008; Ebert-May et al., 2015; Stains et al., 2015). In particular, the chemistry education community has introduced instructors to RBIS and advocated for the adoption of RBIS through the Multi-Initiative Dissemination Project workshops via four reform projects: ChemConnections, Molecular Science, New Traditions—now known as Process Oriented Guided Inquiry Learning (POGIL), and Peer-Led Team Learning (cf. Landis et al., 1998; Peace et al., 2002; Burke et al., 2004;). More recently, the Cottrell Scholars Collaborative New Faculty Workshop was established to prepare chemistry instructors at becoming teacher–scholars by engaging with evidence-based teaching methods (Baker et al., 2014; Stains et al., 2015). Other initiatives include the Core Collaborators Workshops for biochemistry (Murray et al., 2011), POGIL workshops for physical chemistry laboratory (Stegall et al., 2016), and Active Learning in Organic Chemistry workshops (Houseknecht et al., 2020).

In the mathematics community, Project NExT (New Experiences in Teaching) and MathFest minicourses, along with other workshops through the Mathematical Association of America are meant to disseminate new teaching pedagogies. However, Fukawa-Connelly et al. (2016) report that only very small percentages of workshop participants found them to be very influential in their teaching and little importance were assigned to these aforementioned
workshops. In addition, instructors may participate in workshops through the Academy of Inquiry Based Learning to shape their teaching of inquiry-based learning (Fukawa-Connelly et al., 2016).

Within the physics education community, a long-standing workshop spanning more than two decades has been run for new physics and astronomy faculty to increase awareness of RBIS (Henderson, 2008; Henderson et al., 2012). In addition, these types of workshops that expand awareness and utilization of RBIS have been established in biology through the Faculty Institutes for Reforming Science Teaching programs and National Academies Summer Institutes on Undergraduate Education (Wood and Gentile, 2003; Handelsman et al., 2004; Ebert-May et al., 2015; Derting et al., 2016).

It is the assumption that conducting or participating in SOTL or DBER allow instructors use the knowledge gained from that scholarly work to better their teaching practices. While no study, to our knowledge, exists that specifically explores the enacted instructional practices of DBER instructors, SOTL and DBER are closely related pursuits and participation in either will yield similar changes (Henderson et al., 2012). In a meta-analysis of undergraduate STEM instructional practices (Henderson et al., 2011), instructor participation in SOTL has been found to be associated with improvements in course and program-level curricula.

5.3.1.3 Teacher thinking factors

Teacher thinking factors include an instructor’s beliefs about teaching and level of dissatisfaction with their current practices and student learning. Teacher thinking factors that have been reported to be associated with the adoption of active learning that were explored in this study include: (1) satisfaction with student learning and (2) holding a growth mindset.

Erdmann et al. (2020) reported a small relationship between the level of instructor dissatisfaction and pedagogical revisions in a range of STEM disciplines in a sample of primarily biology, chemistry, mathematics, and physics instructors. Teacher thinking has been shown to bolster and hinder the use of new ideas and technologies in constructing classroom and course
learning environment (Moore, 2002; Rogers, 2003). Adoption of new teaching strategies begins with an instructors’ dissatisfaction with the current instruction or a belief that students learn better with strategies not being currently utilized (e.g., Windschitl and Sahl, 2002; Gess-Newsome et al., 2003; Lotter et al., 2007; Bauer et al., 2013; Andrews and Lemons, 2015; Gibbons et al., 2018). Pedagogical dissatisfaction is when an instructor realizes a misalignment of their instructional goals with their instructional practice (Southerland et al., 2011a, 2011b). This disconnect between goals and practice can result in a revision of teaching practice and the adoption of new pedagogical strategies (Feldman, 2000).

Instructors’ mindset beliefs have been reported to likely influence how their courses are structured (Rattan et al., 2012). A fixed or growth mindset is a belief in the inflexibility or malleability, respectively, of a human characteristic (e.g., intelligence; Dweck, 1999). Holding a particular mindset, for example beliefs that traits (e.g., student intelligence) are rigid and cannot be changed (fixed mindset) or can be developed with time and experience (growth mindset), is related to teaching practice choices (Canning et al., 2019). In a longitudinal study of STEM faculty, including chemistry, mathematics, and physics, who endorsed growth mindsets used more motivating pedagogical practices (Canning et al., 2019), which can include active learning strategies (Springer et al., 1999; Prince, 2004; Armbruster et al., 2009).

Holding a growth mindset has been reported to be associated with the uptake of evidenced-based practices, such as active learning, in STEM faculty (Bathgate et al., 2019). Instructors’ mindset beliefs were found to be related to the adoption of active-learning practices in biology; fixed mindset instructors taught using a teacher-centered focus (e.g., lecturing) and growth mindset instructors taught using a student-centered approach (e.g., active learning; Aragón et al., 2018). In mathematics, fixed mindsets were associated with using teaching strategies that would reduce student engagement and achievement (Rattan et al., 2012); in addition, growth mindset beliefs were held by mathematics instructors that were more willing to consider non-lecture pedagogies (Johnson et al., 2018). In observations of instructors that teach
introductory STEM courses, Ferrare (2019) reported that instructors’ beliefs about student learning were linked to certain instructional styles; for example, instructors espousing fixed mindset beliefs were more likely to teach using “chalk talks.”

Interventions based on growth mindset have been reported to be effective (e.g., Dweck and Leggett, 1988; Dweck, 1999; Yeager et al., 2016a, 2016b). In addition, mindset interventions are generalizable and replicable (Bettinger et al., 2018; Yeager et al., 2019). However, there have been some contest as to their replicability (e.g., Bahník and Vranka, 2017; Li and Bates, 2017). An additional challenge comes from practitioners’ misinterpretations of growth mindset and ways to promote it (Dweck 2019). While current research efforts are underway to evaluate the applications of growth mindset and interventions (McMahon et al., 2019), the meta-analyses have shown effectiveness in students (Sisk et al., 2018).

5.3.1.4 Framework conceptualization

The TCSR model is functional for understanding the adoption of active learning strategies due to the interconnectedness of contextual factors, personal factors, and teacher thinking factors along with the interactions between them, when considering why particular teaching practices are enacted in the classroom situated within the larger institutional context. Work in STEM education (e.g., Henderson and Dancy, 2007; Henderson et al., 2011; Oleson and Hora, 2014; Lund and Stains, 2015; Shadle et al., 2017; Johnson et al., 2018, 2019) corroborates these three factors associated with adoption (or barriers to adoption) of more active learning instruction: (1) contextual factors, (2) personal factors, and (3) teacher thinking factors. Figure 5.1 (see above) summarizes the malleable factors included in this study situated within our conceptualization of the TCSR model.
5.3.2 Research question

The conceptual framework described informed the development of the following research question we seek to answer in this study:

To what extent are contextual factors, personal factors, and teacher thinking factors associated with percent time lecturing in gateway chemistry, mathematics, and physics courses when controlling for all other factors and accounting for the nested nature of the data (i.e., instructors within departments within institutions)?

5.4 Methods

We employed survey methodology and used quantitative approaches to answer the research question. Quantitative analysis of the data included multilevel modeling to account for the nested structure of the data (i.e., instructors within departments within institutions). Below, survey development and the nature of the participants included in this study is first described. Then, the multilevel modeling methods are described along with the specific malleable factors (i.e., variables) evaluated in this study.

5.4.1 Survey development

The survey instrument from which specific items are used in this study was developed and informed by previous large-scale studies in postsecondary chemistry (Lund and Stains, 2015; Gibbons et al., 2018; Stains et al., 2018), mathematics (Johnson et al., 2018; Apkarian et al., 2019), and physics (Henderson and Dancy, 2009; Walter et al., 2016, 2021). The full survey asked instructors about five main topics: (1) course context, (2) instructional practices, (3) awareness and usage of active learning instructional techniques, (4) perceptions, beliefs, and attitudes related to students, learning, and departmental context, and (5) personal demographics and experience. Where applicable, previous instruments and scales with reliability and validity evidence were used (e.g., mindset: Dweck et al., 1995). Single-item constructs in the survey show
content and face validity with expert review. Our interpretation of the survey results in this study is the inherent consequential validity that is presented in the Discussion. Survey items used in this study are detailed in Appendix C.1.

5.4.2 Participants

A database was constructed of instructors teaching postsecondary introductory chemistry, mathematics, and physics in the United States \( (n_{\text{total}} = 18,337) \). Instructors in this database were identified through stratified random sampling based on institution type; the goal was to create a representative sample of institution types: 2-year institutions (i.e., associate degree-granting, 4-year institutions (i.e., bachelor’s degree-granting), and universities (i.e., graduate degree-granting). All instructors at each of the selected institutions teaching the targeted introductory level courses were added to the database and invited to participate in the study.

The database includes 9,404 instructors at 4-year institutions (including bachelor’s and graduate degree-granting institutions) in the United States that had conferred at least one bachelor’s degree in all three disciplines (chemistry, mathematics, and physics) between 2011 and 2016 as recorded by the National Center for Education Statistics’ Integrated Postsecondary Education Data System. In addition, the database also includes 8,933 instructors at 2-year institutions in the United States that offer all three of the courses. Contact information for these instructors \( (n_{\text{total}} = 18,337) \) was compiled by the American Institute of Physics Statistical Research Center using publicly accessible online information and through communication with department chairs at the target institutions. Potential survey respondents needed to have taught general chemistry, single-variable calculus, or quantitative-based introductory physics as the primary instructor in 2 years prior to data collection (i.e., in the 2017–18 and/or 2018–19 academic year); in addition, the survey respondents had to have not taught the course exclusively online.
5.4.3 Data collection

Data were collected via the custom-built survey, as described above, was overseen by the American Institute of Physics Statistical Research Center between March 2019 and May 2019; the survey was open for 58 days. Data were collected in accordance with the Western Michigan University Institutional Review Board application no. 17-06-10 approved June 20, 2017, with informed consent obtained digitally. Instructors were sent up to four email notifications: (1) an announcement about the survey, (2) an invitation to participate, and (3 & 4) up to two follow-up invitation emails. A small number of instructors were contacted via follow-up phone calls (N = 603) from the American Institute of Physics Statistical Research Center asking them to complete the survey; instructors were prioritized from institutions, where there were two of the three disciplines already represented to have a data set with representation from all three disciplines.

Respondents included 3769 instructors (20.5% unit response rate) comprised of 2,670 instructors at 4-year institutions and 1099 instructors at 2-year institutions; respondents included 1,244 chemistry, 1,349 mathematics, and 1,176 physics instructors. In total, 1,387 respondents were listwise deleted from the study described herein, because they did not answer all the survey items used in the construction of the multilevel models. The study sample thus included 2,382 instructors from 1,405 departments at 749 institutions for which complete data were collected for the survey items used in this study. The study sample included 795 chemistry, 778 mathematics, and 809 physics instructors with 1,764 instructors at 4-year and 618 instructors at 2-year institutions.

5.4.4 Multilevel models

A three-level model was used to evaluate the impact of malleable factors on amount lecturing in introductory courses in chemistry, mathematics, and physics to account for the nested structure of the data. In this model, instructors (level 1) can be thought of being nested within departments (level 2) which are nested within institutions (level 3). Instructors may, therefore, be
affected by grouping effects at the department and institution levels; this violates the independence of observations assumption required by traditional ordinary least squares regression techniques but can be accounted for in multilevel regression models (Raudenbush and Bryk, 2002; Snijders and Bosker, 2012; Theobald, 2018). If this nested nature of the data is not accounted for, then data may be analyzed at one level with conclusions drawn at another level; this phenomenon is known as an ecological fallacy (Robinson, 1950). Several studies advocate for this specific three-level model (Porter and Umbach, 2001; Smart and Umbach, 2007) and other studies implement variations of this nesting model to align with their research questions (e.g., Marsh et al., 2002; Smeby and Try, 2005; Sonnert et al., 2007). Descriptive statistics for instructor- and department-level factors are given in Appendix C.2 and C.3. Correlations among instructor- and department-level factors are reported in Appendix C.4 and C.5. Variance inflation factors (VIF), reported in Appendix C.6, are well under the suggested cutoff value of 10 and thus do not indicate multicollinearity between predictor variables (Myers, 1990).

Models were constructed using the lme4 package (Bates et al., 2015) in RStudio version 1.2.5033 (R Core Team, 2019) using the full maximum likelihood estimation method due to the large sample size (N = 2382 respondents). The lmerTest package was used to obtain p values for fixed effects (Kuznetsova et al., 2017). T tests used the Satterthwaite approximations for degrees of freedom. An iterative model building process was used by adding and subtracting predictor factors using statistical tests for model comparison to obtain the reported model. Models were compared using theory, fit statistics (i.e., deviance), statistical tests (i.e., χ2 tests), and explained variances. Raudenbush and Bryk notation (2002) is used to describe the multilevel models. The complete final model is reported in Equation 5.1.

Effect sizes indicate the magnitude of the relationship between a predictor variable and the outcome variable. Typically, effect sizes are used as a standardized measure for the comparison of effects within a study or between studies. However, there is no consensus method for calculating effect size in multilevel models (Selya et al., 2012; Lorah, 2018). Cohen’s \( \hat{f} \) is
advantageous as it is compatible with the nested nature of the data and will be calculated as a measure of local fixed effect sizes and global effect size (Cohen, 1988). Local effect size is the proportion of explained variance by a given effect relative to the proportion of the unexplained outcome (i.e., percent lecturing) variance, whereas global effect size is the proportion of explained variance by all effects relative to the proportion of the unexplained outcome variance (Selya et al., 2012; Lorah, 2018). Random effect sizes are related to the intraclass correlation coefficient (ICC), which is the proportion of variance in the outcome accounted for by a level in a multilevel model and is, therefore, a measure of strength of association between level membership (department or institution) and the outcome (Lorah, 2018). The ICC represents and effect size index as the magnitude of the association can be interpreted analogously to a correlation coefficient (Snijders and Bosker, 2012).

5.4.5 Explanation of variables

Table 5.1 includes all factors used in the multilevel model and their coding. In the final multilevel model (Equation 5.1), $LECTURE_{ijk}$ is the overall percent lecturing for an instructor $i$ in department $j$ within institution $k$. Subscripts $i$ and $j$ correspond to variables that apply to individual instructors or their departments, respectively. In the final multilevel model, $\gamma_{000}$ represents the overall mean intercept, $\beta_{001}$ through $\beta_{004}$ represent the department-level (level 2) predictor coefficients, and $\pi_{100}$ through $\pi_{1600}$ represent the instructor-level (level 1) predictor coefficients.
In the custom-built survey, the item corresponding to class size ("What was the approximate enrollment in a typical lecture section?") featured text entry for the response. As a result, this resulted in a range of entry formats from specific values to ranges. In addition, we do not expect instructors to make meaningful instructional decisions based on the exact number of students enrolled in a lecture section (e.g., 32 vs. 33 students), we opted to use bins for the class size variable. The Common Data Set Initiative (n.d.), which is used by *U.S. News & World Report* rankings, uses well-known bins for class sizes: 2–19, 20–29, 30–39, 40–59, 60–99, and 100 or more students. The bins allow for a more interpretable multilevel coefficient versus the value for a single individual student in a class (i.e., class size as a continuous variable). For these reasons, binning of the class size variable was chosen to better inform pedagogical decisions.

### 5.5 Results

We report, herein, the results of a national survey on instructional practices (i.e., percent time lecturing) in introductory, gateway chemistry, mathematics, and physics courses (i.e., general chemistry, single-variable calculus, and introductory quantitative physics). These gateway courses have long been identified as a cause for students not completing STEM degrees (Seymour and Hewitt, 1997; Koch, 2017; Seymour and Hunter, 2019). Data were collected and modeled based on the nested nature of departments and institutions from which the 2,382 respondents were sampled.

A multilevel regression model was constructed to evaluate the association of malleable factors with percent lecturing in introductory chemistry, mathematics, and physics courses. Data were modeled by department and institution to account for the lack of independence of observations at these levels: instructors (*i*, level 1) were nested within departments (*j*, level 2) nested within institutions (*k*, level 3). The unconditional model has an ICC of 0.13 for level 2 (department) and an ICC of 0.11 for level 3 (institution); thus ~13% of variability in the dependent variable (i.e., percent time lecturing) is accounted by nesting observations at the department level.
Table 5.1. Factors used in the final multilevel model and their coding

<table>
<thead>
<tr>
<th>Factor</th>
<th>Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable</strong></td>
<td></td>
</tr>
<tr>
<td>LECTURE: Overall percent lecturing</td>
<td>The overall percent of time during regular class meetings that students spend listening to the instructor lecture or solve problems</td>
</tr>
<tr>
<td><strong>Department characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>CHEM: Chemistry (reference group: math)</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>PHYS: Physics (reference group: math)</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>BACH: Bachelor’s degree is highest offered by department (reference group: associate degree)</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>GRAD: Graduate degree is highest offered by department (reference group: associate degree)</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td><strong>Department appointment expectations</strong></td>
<td></td>
</tr>
<tr>
<td>LOAD: Teaching load</td>
<td>0 = 1 course, 1 = 2 courses, 2 = 3 courses, 3 = 4 courses, 4 = 5+ courses</td>
</tr>
<tr>
<td>TENURE: Tenured instructor (reference group: no opportunity to earn tenure)</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>TENURETRACK: Tenure-track instructor (reference group: no opportunity to earn tenure)</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>SET: Role of student evaluation of teaching in review, promotion, or tenure compared to other measures</td>
<td>0 = not used, 1 = less weight, 2 = equal weight, 3 = more weight, 4 = only used</td>
</tr>
<tr>
<td>APT: Role of assessment of teaching performance in review, promotion, or tenure compared to other measures</td>
<td>0 = not used, 1 = not influential, 2 = somewhat influential, 3 = influential, 4 = very influential</td>
</tr>
<tr>
<td><strong>Classroom contextual</strong></td>
<td></td>
</tr>
<tr>
<td>SIZE: Class size</td>
<td>(-2 = 2–19, \ -1 = 20–29, 0 = 30–39, +1 = 40–59, +2 = 60–99, +3 = 100+ students)\</td>
</tr>
<tr>
<td>ROOM: Classroom setup</td>
<td>0 = fixed seats, 1 = allows for group work</td>
</tr>
<tr>
<td>SIZE × ROOM: Class size × classroom setup interaction effect</td>
<td>Interaction effect of class size and classroom setup with coding used above</td>
</tr>
<tr>
<td>DECISION: Decision making authority</td>
<td>0 = respondent has sole decision-making authority, 1 = in collaboration with others to make decisions</td>
</tr>
<tr>
<td><strong>Personal factors</strong></td>
<td></td>
</tr>
<tr>
<td>RBIS: Had been in a course as a student that had used RBIS</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>SOTL: Conducts scholarship of teaching and learning or discipline-based education research</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>TFC: Has taken teaching-focused coursework at the undergraduate, graduate, or postdoctoral level</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>WKSP: Has participated in teaching-related workshops</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>NFE: Has participated in teaching-related new faculty experiences</td>
<td>0 = no, 1 = yes</td>
</tr>
</tbody>
</table>
Table 5.1 (Continued)

<table>
<thead>
<tr>
<th>Teacher thinking</th>
<th>GROWTH: Degree of holding a growth mindset by the instructor</th>
<th>Satisfaction: Instructor’s level of satisfaction with student learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(-2.5 = ) fixed mindset to (+2.5 = ) growth mindset(^b)</td>
<td>(-2 = ) very dissatisfied to (+2 = ) very satisfied(^c)</td>
</tr>
</tbody>
</table>

\(^a\)Grand-median centered at 30–39 students. \(^b\) Average of three items on a six-point Likert scale from 1 to 6 that describe fixed mindset (Dweck et al., 1995); items were reverse coded to represent increasing growth mindset and centered at the middle of the scale. \(^c\) The single item was answered on a five-point Likert scale from 1 (very dissatisfied) to 5 (very satisfied); values were centered at middle of the scale.

and \(\sim 11\%\) of variability in the outcome variable by nesting at the institution level. When more than 10% of variance occurs between levels, a multilevel model is appropriate (Tai et al., 2005).

Seventeen factors—ten contextual, five personal, and two teacher thinking—were examined (see full model in Table 5.2); 10% of respondent-level, 68% of department-level, and 52% of institutional-level variances (analogous to \(R^2\) in multiple regression) were accounted for in the full model. Additionally, all the factors considered in the full model collectively explained 22% of the variance in the data; when all factors are accounted for in the full model, there is a medium to large global effect size with Cohen’s \(f^2 = 0.28\) (Cohen, 1992). The full model intercept (i.e., 82.66) represents the percent time lecturing reported by an instructor in a mathematics department that awards an associate degree as the highest degree (i.e., reference instructor) at zero (or the median value) for all other evaluated factors (see “Methods”). All multilevel regression coefficients for an individual factor are reported with all other factors held constant, with the exception of the single reported interaction effect between class size and classroom setup.

This model is visualized in Figure 5.2; the regression is plotted with the intercept at 82.66 with multilevel regression coefficients below the intercept showing a decrease in percent lecturing and estimates above the intercept indicating an increase in percent lecturing. For scaled factors, estimates for each scale point is shown. Multilevel regression coefficients indicate the strength of the relationship between a predictor variable and the outcome variable (i.e., percent lecturing) when all other predictor variables are accounted for and held constant.
Table 5.2. Factors in explaining percent lecturing in introductory chemistry, mathematics, and physics courses

<table>
<thead>
<tr>
<th>Estimate</th>
<th>SE</th>
<th>p</th>
<th>Cohen’s $f^2$ (size)$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>82.66</td>
<td>2.78</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Chemistry</td>
<td>0.34</td>
<td>1.25</td>
<td>0.789</td>
</tr>
<tr>
<td>Physics</td>
<td>-4.30</td>
<td>1.20</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Bachelor's program</td>
<td>0.95</td>
<td>1.43</td>
<td>0.505</td>
</tr>
<tr>
<td>Graduate program</td>
<td>1.56</td>
<td>1.63</td>
<td>0.340</td>
</tr>
<tr>
<td>Teaching load</td>
<td>-0.08</td>
<td>0.45</td>
<td>0.855</td>
</tr>
<tr>
<td>Tenured faculty</td>
<td>0.47</td>
<td>1.12</td>
<td>0.671</td>
</tr>
<tr>
<td>Tenure-track faculty</td>
<td>-1.18</td>
<td>1.44</td>
<td>0.414</td>
</tr>
<tr>
<td>Student evaluation of teaching</td>
<td>-0.41</td>
<td>0.47</td>
<td>0.391</td>
</tr>
<tr>
<td>Assessment of teaching performance</td>
<td>-0.04</td>
<td>0.55</td>
<td>0.939</td>
</tr>
<tr>
<td>Class size</td>
<td>1.14</td>
<td>0.50</td>
<td>0.021</td>
</tr>
<tr>
<td>Classroom setup</td>
<td>-10.71</td>
<td>1.08</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Class size × classroom setup</td>
<td>-1.17</td>
<td>0.67</td>
<td>0.080</td>
</tr>
<tr>
<td>Decision making</td>
<td>-4.60</td>
<td>1.34</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>RBIS use as a student</td>
<td>-3.83</td>
<td>1.31</td>
<td>0.003</td>
</tr>
<tr>
<td>Scholarship of teaching and learning</td>
<td>-8.56</td>
<td>0.96</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Teaching-focused coursework</td>
<td>-3.22</td>
<td>0.97</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Teaching-related workshops</td>
<td>-8.24</td>
<td>1.59</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Teaching-related new faculty experiences</td>
<td>-4.69</td>
<td>1.14</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Growth mindset</td>
<td>-3.08</td>
<td>0.41</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Satisfaction with student learning</td>
<td>-1.17</td>
<td>0.67</td>
<td>0.080</td>
</tr>
</tbody>
</table>

$^a$Cohen’s (1992) $f^2$ effect size: < 0.02 = negligible (n); ≥ 0.02 = small (s); ≥ 0.15 = medium (m); ≥ 0.35 = large (l)

5.5.1 Department characteristics factors

Department characteristics factors evaluated in this model include discipline and highest degree awarded by department.

Academic discipline (i.e., chemistry, mathematics, physics) is evaluated with mathematics as reference: Holding all other evaluated variables constant, percent time lecturing for instructors from chemistry departments ($\beta_j = 0.34, p > 0.05, f^2 < 0.01$) is not statistically different from instructors in mathematics departments; percent time lecturing for instructors from physics departments is 4.30% less ($p < 0.001, f^2 < 0.01$) than instructors from mathematics departments.

Highest degree awarded by the department is evaluated with an associate degree as the highest degree awarded as reference: Holding all other evaluated variables constant, average percent time lecturing for instructors in departments awarding as the highest degree bachelor’s degrees ($\beta_j = 0.95, p > 0.05, f^2 < 0.01$) or graduate degrees ($\beta_j = 1.56, p > 0.05, f^2 < 0.01$) are not
Figure 5.2. Comparative association of variables with percent time lecturing: (A) Department characteristics, personal factors, and teacher thinking. (B) Classroom contextual factors and department appointment expectations. The line at 82.66 represents the intercept of the model statistically different from instructors in departments awarding associate degrees and its highest degree.

5.5.2 Department appointment expectations

Department appointment expectations evaluated in this model include teaching load, tenure status, and perceived professional review and reward structure factors.

Teaching load and tenure status are evaluated: Teaching load is non-significant, negatively associated with percent time lecturing ($\pi_j = -0.08$, $p > 0.05$, $\ell < 0.01$). In addition, tenure-status is evaluated (“not in a tenure-track position” is the reference): Changes in percent
time lecturing for instructors in a tenure-track position ($\pi_j = 0.47, p > 0.05, f^2 < 0.01$) or in a tenured position ($\pi_j = -1.18, p > 0.05, f^2 < 0.01$) are not statistically from instructors without opportunity for tenure.

Perceived role of assessment of teaching performance in review, promotion, and tenure is evaluated: A non-significant, negligible decrease in percent time lecturing ($\pi_j = -0.41, p > 0.05, f^2 < 0.01$) is associated with increases in the role of student evaluations of teaching in evaluating teaching performance. Additionally, a non-significant, negligible decrease in percent time lecturing ($\pi_j = -0.04, p > 0.05, f^2 < 0.01$) is associated with an increase in the perceived role overall assessment of teaching performance matters in decisions of review, promotion, and tenure.

5.5.3 Classroom contextual factors

Classroom contextual factors evaluated in this model include class size, physical layout, and course administration (i.e., involvement in decision making).

Physical space, including maximum number of students for a given classroom, and configuration and type of furniture, are evaluated: A positive, small increase in percent time lecturing is associated with larger course sizes ($\pi_j = 1.14, p = 0.021, f^2 < 0.01$). A large decrease in percent time lecturing is associated with classroom spaces conducive to group work ($\pi_j = -10.71, p < 0.001, f^2 = 0.05$, small effect). There is a non-significant, small interaction effect between course size and classroom setup ($\pi_j = -1.17, p > 0.05, f^2 < 0.01$); this interaction effect essentially cancels out fluctuations in course size for classrooms spaces conducive to group work.

The primary decision maker for course instructional methods is evaluated: A significant decrease in percent time lecturing ($\pi_j = -4.60, p < 0.001, f^2 < 0.01$) is associated with courses in which decisions are made primarily by the instructor in conjunction with at least one additional instructor of the department.
5.5.4 Personal factors

Instructor’s personal factors evaluated in this model include prior experience in courses taught with research-based instructional strategies, participation in SOTL or DBER, and also involvement in pedagogical professional development.

Experience as a student in courses taught with research-based instructional strategies is evaluated: Such experience is associated with a significant decrease in percent time lecturing ($\pi_j = -3.38, p < 0.01, \bar{f} < 0.01$).

Participation in SOTL or conducting DBER is evaluated: A significant decrease in percent time lecturing ($\pi_j = -8.56, p < 0.001, \bar{f} = 0.03, \text{small effect}$) is associated with such engagement.

Participation in teaching-related professional development experiences is evaluated: A significant decrease in percent time lecturing ($\pi_j = -3.22, p < 0.001, \bar{f} < 0.01$) is associated with participation in teaching-focused coursework at the undergraduate, graduate, or postdoctoral levels. A significant decrease in percent time lecturing ($\pi_j = -8.24, p < 0.001, \bar{f} < 0.01$) is associated with participation in teaching-focused workshops including half-day to multiple day workshops and attending teaching-focused conferences. A significant decrease in percent time lecturing ($\pi_j = -4.69, p < 0.001, \bar{f} < 0.01$) is associated with participation in teaching-related new faculty experiences internal or external to the respondent’s institution.

5.5.4 Teacher thinking factors

Teacher thinking factors evaluated in this model include holding a growth mindset and satisfaction with student learning.

Growth mindset (i.e., the belief that ability can be developed) is evaluated: A significant decrease in percent time lecturing ($\pi_j = -3.08, p < 0.001, \bar{f} = 0.02, \text{small effect}$) is associated with an increase in growth mindset. Satisfaction with student learning is evaluated: A non-significant decrease in percent time lecturing ($\pi_j = -1.17, p > 0.05, \bar{f} < 0.01$) is associated with increased satisfaction with student learning.
5.6 Discussion

Six themes emerged from our multilevel regression model results that are associated with decreased percent time lecturing, at the instructor-level, when all other factors are held constant: (1) classroom spaces conducive to group work, (2) shared decision making on instructional methods, (3) participation as a student in courses utilizing research-based instructional strategies, (4) participation in teaching-related professional development experiences, (5) participation in scholarship of teaching and learning and discipline-based education research, and (6) espousing a growth mindset.

Only malleable factors are considered and interpreted in this Discussion as only these factors can lead to tangible implications for institutions and potential actions to support the adoption of more engaging instructional practices. In addition, these malleable factors have tangible implications for professional organizations and communities of practice to support transformation and reform efforts. Therefore, we do not consider department-level factors (e.g., discipline and highest degree awarded) as these cannot be changed; it is not practical for instructors to abandon their disciplinary training and switch to a new STEM field or take up a new position at a different institution for the sake of enacting more active learning practices.

5.6.1 Classroom spaces

Instructors consistently indicate that large class sizes, coupled with fixed-seating classroom layouts (i.e., lecture halls with bolted seats to the floor), are not conducive to interactions between students and make it difficult to implement research-based instructional strategies (Gess-Newsome et al., 2003; Henderson and Dancy, 2007; Hora, 2012; Lund and Stains, 2015). Studies routinely show that smaller-sized classes held in spaces that allow for active learning (e.g., moveable tables or desks) are associated with implementation of more student-centered teaching methods (Cotner et al., 2013; Lund and Stains, 2015; Shadle et al.,
Our findings corroborate these studies indicating that class size and classroom spaces matter, with the latter having a large, significant association with decreased percent time lecturing.

These findings beg the question: “If we build it, will they come?” Or more pointedly, “If classrooms spaces are built or renovated to be more conducive to group work, will instructors implement more active-learning pedagogies in courses taught in such spaces?” The underlying ambiguity is: what is the cause and what is the effect? Our results can be interpreted that the space catalyzes implementation of non-lecture-based pedagogies. Some evidence indicates that building a classroom for active learning increases the like likelihood for such intended purposes, because it motivates and encourages instructors to try active learning pedagogies (Foote et al., 2016). Conversely, our results could be interpreted in such a way that instructors wanting to use pedagogies that are more easily implemented in classroom spaces conducive to group work seek out and request to teach in such spaces. Having a decided active learning classroom increases the sustainability of active learning implementations as it takes effort to undue significant structural change because of the buy in and support of many individuals, including instructors and administrators (Knaub et al., 2016). Both causal explanations are only possible if such classroom spaces exist and are available. Regardless, we argue that such classrooms spaces should be advocated for and built regardless of the cause–effect relationship.

There are resources readily available for designing classroom spaces that promote implementation of active learning pedagogies: One such example is SCALE-UP (Student-Centered Active Learning Environment with Upside-down Pedagogies) which aims to reform teaching practices by manufacturing physical changes to the classroom layout that in turn minimize lecture (SCALE-UP, 2011; Knaub et al., 2016). SCALE-UP has demonstrated improvements in students’ problem-solving abilities, conceptual understanding, attitudes toward science, retention in introductory courses, and later performance in subsequent courses in chemistry, mathematics, and physics; studies have also shown a reduction in failure rates, especially for women and minoritized students in SCALE-UP classrooms (Beichner et al., 2007;
Beichner, 2008). Additionally, FLEXspace® (Flexible Learning Environments Exchange space) is a tool that supports communities of experts, practitioners, and institutional decision makers to improve active learning space planning, design, and implementation (FLEXspace, 2018).

5.6.2 Shared decision on instructional methods

Introductory courses in chemistry, mathematics, and physics are typically large enrollment courses, even at smaller-sized institutions (Seymour and Hewitt, 1997; Koch, 2017; Seymour and Hunter, 2019). We define “large” to be considered relative to other upper-level courses within a given institution. Irrespective of institution size, though, typically large enrollment courses are divided into smaller classes (i.e., sections). It is common for more than one instructor to be teaching the set of sections for a given course. The degree of coordination of these sections can vary from complete independence (including different textbooks and syllabi) to complete coordination, wherein all aspects of the course are common across sections including examinations and instructional practices (including different textbooks and syllabi) to complete coordination wherein all aspects of the course are common across sections including examinations and instructional practices (Apkarian and Kirin, 2017). Decision making authority for aspects of the course may lie with the individual instructor or with a committee. Our findings suggest that when an instructor shares decision making authority with one or more instructors on the instructional methods used, that such coordination is associated with less time lecturing.

Decisions on teaching methods are best when done as a team. A recent study revealed that instructors at three research-intensive universities who use innovative teaching practices preferentially interact with other users due to their similar teaching values (Lane et al., 2020). As Lane et al. (2020) note, co-teaching and teaching teams can encourage the interaction between instructors of different teaching practices (Henderson et al., 2009). Instructors that have experience in implementing active learning strategies should collaborate or co-teach with non-users to assist in the uptake of these pedagogies (Gess-Newsome et al., 2003; Henderson and
Dancy, 2007; Henderson et al., 2009). Instructors state that it is easier to integrate research-based methods if other instructors are implementing new methods at the same time (Henderson and Dancy, 2007; Foote et al., 2016); implementation requires supportive departmental and institutional mentors (Henderson and Dancy, 2007; Shadle et al., 2017). Coordinating a course across multiple class sections is a possible route to execute this change and offers an opportunity for instructional support outside of formal avenues of professional development (Henderson et al., 2009; Apkarian et al., 2019; Golnabi et al., 2021; Martinez and Pilgrim, 2021; Williams et al., 2022).

We advocate for course coordination to encourage the exchange of ideas and experience among instructors as means to reduce the time constraints and uncertainties of implementing new instructional practices (Shadle et al., 2017). As lack of time has been previously noted as a barrier for the implementation of new teaching pedagogies (e.g., Henderson and Dancy, 2007; Brownell and Tanner, 2012; Lund and Stains, 2015), the appointment of course coordinators with long-term roles would provide necessary and ample time to support implementation of active learning pedagogies (Rasmussen and Ellis, 2015). In addition, co-teaching can be used in cases of a two-section course, where one instructor is experienced in the implementation of active learning strategies (Henderson et al., 2009). Course coordinators can discuss RBIS implementation with instructors in the teaching team (Golnabi et al., 2021) and others at department meetings (Bazett and Clough, 2021; Williams et al., 2022) to facilitate conversation around active learning uptake.

5.6.3 Experience as a student in a course using research-based instructional strategies

It has been reported that teaching methods experienced by STEM instructors when they were students influence their current teaching practices (Oleson and Hora, 2014; Lund and Stains, 2015; Fukawa-Connelly et al., 2016). Prior classroom experiences as an instructor, professional development programs, and interactions with other instructors are also influential in
teaching pedagogies (Oleson and Hora, 2014). Our results corroborate these findings with experiences of RBIS as a student aligning with lower reported percent time lecturing.

A continual cycle of future instructors (i.e., current undergraduate and graduate students) using active learning strategies can pave the way for an effective, longitudinal, and sustainable method for implementing pedagogical reform supported the results of this study and the notion that previous experiences a student influence present teaching pedagogies practices (Oleson and Hora, 2014; Lund and Stains, 2015; Fukawa-Connelly et al., 2016). For this to succeed though, current instructors need to be made aware of and be willing to implement RBIS. Active learning pedagogies used in today’s classrooms will influence future instructors.

5.6.4 Participation in teaching-related professional development

Our findings show that instructors who have taken teaching-focused coursework, participated in teaching-related workshops, or teaching-related new faculty experiences report a lower percentage time lecturing than instructors who have not engaged in these opportunities. Centers for teaching and learning and professional organizations (e.g., the American Chemical Society, the Mathematical Association of America, and the American Physical Society) provide opportunities for instructors, in addition to graduate students and postdoctoral scholars, to participate in professional development workshops. Teaching-focused workshops are shown to be effective in informing teaching decisions (Henderson et al., 2011; Oleson and Hora, 2014). In addition, some centers for teaching and learning now provide structured for-credit teaching-focused coursework independently or in collaboration with colleges of education. Topics of these professional development opportunities range from discipline-specific pedagogical training to teaching and learning theory, effective pedagogical practices, and design of instructional materials (cf. Gardner and Jones, 2011; Wyse et al., 2014; Coppola, 2016; Wheeler et al., 2017;). Although few instructors receive pedagogical training as a part of their graduate programs, STEM instructors who have received training were found to more likely to have referenced sources of
instructional innovation (Walczyk et al., 2007). While most professional development programs focus on new or early-career faculty and instructors, programs should also be developed for and tailored to those at mid-career or late-stages (Austin and Sorcinelli, 2013), and institutionalized and sustained professional development is necessary for lasting pedagogical change (Henderson et al., 2011; Borda et al., 2020).

In addition to pedagogical training, centers for teaching and learning can provide practical assistance to all instructors in their active learning endeavors. These centers can also sponsor communities of practice for instructors to discuss their teaching practices and instructional projects, which have also been shown to increase use of RBIS and transfer knowledge between disciplines to enhance instruction (Henderson et al., 2017; Pelletreau et al., 2018; Dancy et al., 2019; Tomkin et al., 2019; Benabentos et al., 2021); for adopters of active learning pedagogies, participation in a community of practice has been shown to have greater use of student-centered practices (Benabentos et al., 2021). If these centers for teaching and learning do exist, we propose that centers are given sufficient resources to achieve their goals. A common reason why these centers are under-utilized is because center staff, while experts in education, may not have broad disciplinary expertise; to increase credibility, centers for teaching and learning should incorporate more discipline-specific skills training and hire more persons with a broad spectrum of DBER experience (Seymour et al., 2011; Pelletreau et al., 2018). At institutions, where centers are not financially feasible, peer-coaching may be a useful form of professional development (Desimone and Pak, 2017); instructors can observe one another’s teaching and provide feedback and discuss teaching methods (Gormally et al., 2014; Smith et al., 2014).

Dissemination of RBIS through new faculty experiences has been shown to increase instructor awareness and inclusion of these instructional strategies in biology (Wood and Gentile, 2003; Handelsman et al., 2004; Ebert-May et al., 2015; Derting et al., 2016), chemistry (Murray et al., 2011; Baker et al., 2014; Stains et al., 2015; Stegall et al., 2016; Houseknecht et al., 2020), and physics (Henderson, 2008; Henderson et al., 2012). Institutions may require, or highly
incentivize, participation in these professional development experiences for all new instructors, or through tenure and promotion requirements (Seymour et al., 2011; Bathgate et al., 2019a, 2019b). Institutions can also help sustain implementation of RBIS through, at the very minimum highly incentive, programs designed for mid- and late-stage instructors (Austin and Sorcinelli, 2013; Borda et al., 2020).

5.6.5 Scholarship of teaching and learning or discipline-based education research

Conducting or participating in SOTL or DBER had the largest impact on reducing the percent time teaching of all non-contextual factors in our study. In an analytical review of literature on undergraduate STEM practices, Henderson et al. (2011) noted that engagement in SOTL is a means for developing reflective educators. Instructors that engaged in STEM education research were found to use more student-centered instructional practices (Henderson et al., 2017; Pelletreau et al., 2018; Dancy et al., 2019; Tomkin et al., 2019; Benabentos et al., 2021). This is corroborated by our findings that participation in SOTL is associated with adoption of more active learning pedagogies. Thus, instructors should be encouraged to engage and be recognized for SOTL, and by extension, DBER work (Henderson et al., 2012). Institutions should encourage and reward instructors for their efforts in SOTL and treat it as a valuable scholarly outlet (Walczyk et al., 2007). Collaborative projects with education scholars, both discipline-based and within colleges of education, can serve to catalyze purposeful investigation of teaching and learning in STEM courses (Oleson and Hora, 2014; Shadle et al., 2017; Dancy et al., 2019). In addition, collaborative SOTL projects between instructors and graduate students enrolled in future faculty programs provide an additional pathway for involving current and future educators in teaching-oriented scholarship (e.g., Coppola, 2016), which would give both short- and long-term benefits to individuals and the field.
5.6.6 Holding a growth mindset

Our results suggest that growth mindset beliefs are associated with a reduction in the amount of time spent lecturing. Instructors of different identities (i.e., gender and race/ethnicity) and experiences (i.e., teaching experience and tenure status) across multiple STEM disciplines have been found to espouse fixed mindsets (Canning et al., 2019). This is problematic such that fixed mindsets may make stereotype threats more evident and concerning; stereotyped stigmatized students have been shown to experience more anxiety, lower sense of belonging, and become less interested (Emerson and Murphy, 2015; Bian et al., 2018). In addition, instructors holding fixed mindsets may inhibit the pursuit of graduate-level education by women and minoritized students (Leslie et al., 2015). Fixed mindsets are malleable and large-scale studies have shown that more of a growth mindset can be developed (e.g., Yeager and Dweck, 2012; Yeager et al., 2016a, 2016b; Broda et al., 2018), although do not yet know the effect that growth mindset interventions have on the implementation of active learning (Aragón et al., 2018).

Studies have shown that instructors’ growth mindsets have the potential to improve student learning as well as address equity in the classroom (e.g., Gasiewski et al., 2012; Leslie et al., 2015; T. Smith et al., 2018; Canning et al., 2019). Safe environments promote engagement; students felt more comfortable asking questions when instructors held a growth mindset, which is exemplified by their feeling that a too rudimentary question was nonexistent (Gasiewski et al., 2012). Instructor feedback can communicate their mindset beliefs; students that received growth mindset comments themselves moved toward more growth mindset beliefs and scored higher on a summative assessment than their counterparts who received fixed mindset comments (Smith et al., 2018). This is advantageous as growth mindsets have been associated with higher achievement in students (e.g., Blackwell et al., 2007; Yeager and Dweck, 2012; Burnette et al., 2013; Yeager et al., 2019). Thus, professional development activities including dissemination of active learning pedagogies, teaching assistant training, and new faculty experiences should include a component on beliefs about teaching and learning. These training and experiences
should aim to make instructors aware of how their mindset influences the culture in their classes and student motivation and achievement (Canning et al., 2019).

5.6.7 Non-significant factors

Several factors are not statistically significant in explaining variability in percent time lecturing. These factors include teaching load, tenure status, student evaluations of teaching, assessment of teaching performance, and satisfaction with student learning.

The insignificance of teaching load and tenure status potentially indicates that instructors have a high work expectation that is irrespective of the division of time between teaching and research; in other words, instructors are busy. We know that current incentives are lacking and there is a limited focus on teaching in annual and merit reviews for faculty members to enact research-based instructional strategies (Lund and Stains, 2015; Shadle et al., 2017). When the association between percent time lecturing and the importance of student evaluations of teaching or the role of teaching performance in reviews are considered without controlling for other predictors, a significant, but small association has been found (Apkarian et al., 2021); this suggests that there may be a more nuanced relationship between instructional practices and evaluations of teaching performance than what is reported in the study reported herein. Nonetheless, departmental and institutional pressures or incentives can become more influential if a larger emphasis is placed on the assessment of teaching performance in annual, tenure, or promotion evaluations (Lund and Stains, 2015).

Finally, decreased satisfaction with student learning has been shown in other studies to be associated with adoption of active learning strategies, with dissatisfaction being a central tenant to the TCSR model that frames our study and analyses (Gess-Newsome et al., 2003). However, we acknowledge that the relationship between this factor and percent time lecturing may be more complex than what we have modeled in our study given the array of modeled factors.
One way to interpret this is that instructors who lecture less are happier with their student outcomes; this would be consistent with dissatisfaction leading to change.

5.6.8 Limitations

This study has a limited scope; we only sampled instructors who teach introductory courses from chemistry, mathematics, and physics. In addition, previous studies show differences exist between lower- and upper-division STEM courses (Lund et al., 2015; Benabentos et al., 2021). To better characterize STEM courses as a whole, a broader study should be conducted that includes a wider array of STEM disciplines at both introductory and advanced levels from a representative sample of different institution types.

The TCSR model was originally developed to understand reform in K-12 education (Woodbury and Gess-Newsome, 2002) and was then adapted for a college classroom (Gess-Newsome et al., 2003). Consequently, due to our review of the literature, our conceptualization of the TCSR framework required further modification of the framework to delineate department appointment expectations apart from departmental contextual factors to account for the complex nature of different instructional positions in higher education. In addition, it should be noted that the TCSR framework has not been previously used in multilevel analyses; therefore, some of the considerations of testing a statistical model such as the nested nature of the data (i.e., instructors within departments within institutions) have not been previously addressed. While a modification of the model (i.e., an articulated differentiation of appointment expectations) was necessary, this does not invalidate the TCSR model, but furthers theoretical and empirical possibilities for using the model to evaluate teaching practices.

The self-reported nature of our outcome measure (i.e., percent time lecturing) and a respondent’s contextual, personal, and beliefs about teaching and learning factors result in some loss of empirical strength due to potential reliability threats and may potentially mischaracterize the complex nature of the classroom. Discrepancies between self-reported data and researcher-
evaluated observations of observed classroom practices have previously been found (e.g., Koziol and Burns, 1986; Bodzin and Beerer, 2003; Herrington et al., 2016). Although, observational studies are prone to observer subjectivity, particularly in regard to rater agreement (Waxman and Padrón, 2004; Hill et al., 2012). Low numbers of observations, insufficient training of raters, and non-representative snapshots of instructional practices can raise uncertainty of observation data (Hill et al., 2012; Cohen and Goldhaber, 2016). However, evidence suggests that self-reported data about teaching practices align well with observational studies (Durham et al., 2018; Gibbons et al., 2018; Hayward et al., 2018). While large-scale studies are necessary to identify teaching practices (Williams et al., 2015), we note that observational studies, though, would parallel self-report data but is impractical with the large sample of our study (i.e., $N = 2,382$). Balancing error associated with self-report data and the opportunity to conduct large-scale study such as ours, supported by the work of others (Durham et al., 2018; Gibbons et al., 2018; Hayward et al., 2018), the results of the study herein are trustworthy.

5.7 Conclusions

Based on these results from a national survey of gateway chemistry, mathematics, and physics instructors that considers a large number of factors associated with uptake of active learning and accounts for the nested nature of institutional contexts, we provide four broad recommendations for sustaining active learning strategies in introductory STEM courses:

1. Construct classroom spaces that support and promote active learning (i.e., moveable table/desks for shared group work and activities, whiteboards to support collaboration, etc.). Provide and incentivize professional development to assist instructors in maximizing the use of active learning spaces.

2. Coordinate large enrollment courses with multiple course sections. Collaborate with other instructors on instructional methods, allowing for discussion and reflection on instructional practices.
3. Offer and encourage participation in professional development programs and communities of practice for widespread awareness and implementation of research-based instructional strategies. Promote a growth mindset and develop constructive beliefs about teaching and learning in professional development opportunities.

4. Engage in the scholarship of teaching and learning. Recognize and value SOTL work accomplished by instructors.

Our results demonstrate to instructors, departmental, and institutional leaders on how contextual, personal, and teacher thinking factors are associated with decisions about instructional practices at their institutions. The research literature suggests that instructor-centered reforms and instructor-, departmental-, and institutional-led reform initiatives are essential to meaningful and sustained cultural change about how STEM courses are taught (Beichner et al., 2007; Henderson et al., 2011; Shadle et al., 2017; Reinholz and Apkarian, 2018).

5.8 Abbreviations

ATP: assessment of teaching performance; DBER: discipline-based educational research; FLEXspace: Flexible Learning Environments Exchange space; ICC: intraclass correlation coefficient; NeXT: New Experiences in Teaching; NFE: new faculty experience; POGIL: Process Oriented Guided Inquiry Learning; RBIS: research-based instructional strategies; SCALE-UP: Student-Centered Active Learning Environment with Upside-down Pedagogies; SET: student evaluations of teaching; SOTL: scholarship of teaching and learning; STEM: science, technology, engineering, and mathematics; TFC: teaching-focused coursework; TCSR: Teacher-Centered Systemic Reform; VIF: variance inflation factor; VITAL: visitors, instructors, teaching assistants, adjuncts, and lecturers; WKSP: workshop.
5.9 Acknowledgments

The authors wish to acknowledge Eun Sook Kim (University of South Florida) for insightful conversations about the construction and interpretation of the multilevel models. We would like to thank all the instructors that participated in this study.

5.10 Author contributions

C.H., M.H.D., E.J., J.R.R. and M.S. conceived of the study and designed the survey. B.J.Y, J.R.R., N.A., and M.S. analyzed the data. B.J.Y. and J.R.R. wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

5.11 Funding

This material is based upon work supported by the National Science Foundation under Grant Nos. DUE 1726126 (J.R.R.), 2028134 (M.S.), 1726328 (C.H.), 1726042 (M.H.D.), and 1726281 (E.J.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

5.12 Competing interests

The authors declare that they have no competing interests.

5.13 References


Canning, E. A., Muenks, K., Green, D. J., and Murphy, M. C. (2019). STEM faculty who believe ability is fixed have larger racial achievement gaps and inspire less student motivation in their classes. *Science Advances, 5*:eaau4734. doi: 10.1126/sciadv.aau4734


Chapter 6

Association of malleable factors with adoption of research-based instructional strategies in introductory chemistry, mathematics, and physics

6.1 Note to Reader

This work has co-authors. Charles Henderson, Melissa H. Dancy, Estrella Johnson, Jeffrey R. Raker, and Marilyne Stains conceived the larger survey project and designed the survey. Naneh Apkarian aided in designing the survey, cleaned, and compiled the data. Jeffrey R. Raker, Naneh Apkarian, and Marilyne Stains contributed to analyzing the data. Jeffrey R. Raker is the principal investigator for this study.

6.2 Abstract

Active learning pedagogies are shown to enhance the outcomes of students, particularly in disciplines known for high attrition rates. Despite the demonstrated benefits of active learning, didactic lecture continues to predominate in science, technology, engineering, and mathematics (STEM) courses. Change agents and professional development programs have historically placed emphasis on develop–disseminate efforts for the adoption of research-based instructional strategies (RBIS). With numerous reported barriers and motivators for trying out and adopting active learning, it is unclear to what extent these factors are associated with adoption of RBIS and the effectiveness of change strategies. We present the results of a large-scale, survey-based study of introductory chemistry, mathematics, and physics instructors and their courses in the United States. Herein, we evaluate the association of 17 malleable factors with the tryout and adoption of RBIS. Multilevel logistic regression analyses suggest that several contextual,
personal, and teacher thinking factors are associated with different stages of RBIS adoption. These results are also compared with analogous results evaluating the association of these factors with instructors’ time spent lecturing. We offer actionable implications for change agents to provide targeted professional development programming and for institutional leaders to influence the adoption of active learning pedagogies in introductory STEM courses.

6.3 Introduction

Research-based instructional strategies (RBIS; see Dean et al., 2012) and evidence-based instructional practices (EBIPs; see Stains and Vickrey, 2017) are similarly used labels for instructional practices with a basis in educational research, including active learning. Active learning in undergraduate science, technology, engineering, and mathematics (STEM) courses has been demonstrated to enhance student outcomes (Springer et al., 1999; Lorenzo et al., 2006; Haak et al., 2011; Ruiz-Primo et al., 2011; Freeman et al., 2014; Rahman and Lewis, 2019; Theobald et al., 2020). Compared to traditional lecture-based courses, active learning courses are associated with increased achievement across all STEM disciplines (Freeman et al., 2014). Notably, studies also show an increase in achievement outcomes for minoritized populations in active learning courses (Lorenzo et al., 2006; Kogan and Laursen, 2014; Synder et al., 2016; Ballen et al., 2017; Deri et al., 2018; Roberts et al., 2018; Stanich et al., 2018; Theobald et al., 2020). For college students enrolled at two-year institutions and community colleges, active learning has been shown to contribute to increased graduation rates (Riedl et al., 2021) and transfer rates (Wang et al., 2017). In this paper, we refer to these teaching practices as RBIS, which can include think-pair-share, small group work, peer instruction, peer-led team learning, flipped classroom, and just-in-time-teaching (for more comprehensive lists of RBIS, see: Henderson and Dancy, 2009; Borrego et al., 2013; Baker et al., 2014). These strategies have foundations in the education research literature, are contrasted with didactic lecture, and engage
students in the learning process in lieu of passively listening to an instructor (Bonwell and Eison, 1991).

Although active learning pedagogies undoubtedly have established benefits in STEM, lecture-based pedagogical approaches remain dominant (Stains et al., 2018). The research literature suggests the prominence of lecture-oriented pedagogies may be a result of institutional failure to normalize use of RBIS (Henderson and Dancy, 2007; Shadle et al., 2017), failure to implement faculty incentives or rewards for using student-centered pedagogical techniques (Michael, 2007; Brownell and Tanner, 2012; Shadle et al., 2017), and failure to combat student resistance to instructional changes (Henderson and Dancy, 2007; Michael, 2007; Shadle et al., 2017), among the plethora of reasons.

In addition to these barriers, instructors have expressed feeling unprepared for changing the way they teach (Andrews and Lemons, 2015; Bathgate et al., 2019a). Foremost, instructors may have little or no knowledge about and awareness of alternatives to traditional lecturing, i.e., using RBIS (Hativa, 1995; Miller et al., 2000; Luft et al., 2004; Walczyk et al., 2007; Yarnall et al., 2007; Winter et al., 2012). Once aware, though, of RBIS and active learning strategies, instructors may lack opportunities to try out these new strategies (Handelsman et al., 2004; Ebert-May et al., 2011) or may be unconvinced that these strategies are more effective than lecturing (Miller et al., 2000; Yarnall et al., 2007; Winter et al., 2012).

While meta-analytic work by Freeman et al. (2014) and Theobald et al. (2020) have noted the importance of active learning in STEM education, parallel multidisciplinary studies of malleable factors (i.e., something that can be changed and altered) related to the adoption of such active learning pedagogies in postsecondary STEM courses are largely absent from and needed in the research literature (NRC, 2012; AAAS, 2019). To date, only one study at such a level (Yik et al., 2022) details the association of malleable factors with the adoption of active learning in multiple STEM disciplines. In Yik et al. (2022), as with this study, we focus on introductory chemistry, mathematics, and physics, which are high-enrollment courses serving a large number
of students (PCAST, 2012) and are barriers for students not finishing STEM degrees (Seymour and Hewitt, 1997; Koch, 2017; Seymour and Hunter, 2019). In Yik et al. (2022), we evaluated 17 malleable factors that have been reported in the research literature to be associated with percent lecturing (i.e., time not spent using active learning strategies) in gateway STEM courses. However, another measure of active learning is stage of adoption of RBIS. While knowing the amount of time lecturing allows for conclusions about the degree of student-centered learning, it does not provide information about an instructor’s awareness about RBIS, time spent learning about and readiness to tryout RBIS, and adoption of RBIS in their courses (Landrum et al., 2017); instructors’ needs are different based on their awareness, tryout, and adoption of RBIS (Viskupic et al., 2022). This study will add to the research literature by quantifying the association of malleable factors on the adoption of RBIS in introductory chemistry, mathematics, and physics courses. Evaluation of these malleable factors on RBIS adoption allows for comparisons with a previous variable, i.e., percent time lecturing (Yik et al., 2022), on the effects of different aspects of active learning, and thus, yields recommendations for the adoption of RBIS to promote active learning in introductory STEM courses.

In the context of introductory chemistry, mathematics, and physics courses, three research questions guide this study:

1. To what extent are malleable contextual, personal, and teacher thinking factors associated with the stages of RBIS adoption?
2. How do the association of malleable contextual, personal, and teacher thinking factors with the stages of RBIS adoption compare with percent lecturing?

6.4 Conceptual frameworks

Research on barriers and driving forces for the adoption of teaching strategies informed the selection of malleable factors that were previously modeled (see Yik et al., 2022) and are modeled in this study. Researchers (e.g., Gess-Newsome et al., 2003; Dancy and Henderson,
have described a number of contextual, personal, and belief factors that influence instructors' pedagogical decisions. Through a comprehensive review of the literature, (Woodbury and Gess-Newsome, 2002) developed the Teacher-Centered Systemic Reform (TCSR) model to understand how classrooms change due to reform initiatives; this framework was later modified to better suit the higher education system (Gess-Newsome et al., 2003). The TCSR model focuses on teachers' thinking and practices as the origin for change that happen within the classroom inside the context of a larger university system (Gess-Newsome et al., 2003). In a university context, the TCSR framework is comprised of three broad categories: contextual factors (e.g., institution type, discipline, and class size), personal factors (e.g., extent of teacher preparation and teaching-related professional development), and teacher thinking factors (e.g., knowledge about teaching and dissatisfaction with current teaching practices); these categories were used to situate the malleable factors modeled in our prior study (Yik et al., 2022). Our current study also aims to evaluate the association of these same malleable factors with the adoption of RBIS when institutional and disciplinary differences are accounted for. Non-malleable factors (e.g., race/ethnicity) are not included because they do not provide actionable implications. A summary of the malleable factors, logic for inclusion in this study, and relevant literature citations are given in Table 6.1.

Dormant’s (2011) Chocolate Model of Change, also known as the CACAO model, has been used to conceptualize institutional change in STEM education (Marker et al., 2015; Landrum et al., 2017; Shadle et al., 2017; Earl et al., 2020; Pilgrim et al., 2020; Salomone et al., 2020; McAlpin et al., 2022; Viskupic et al., 2022). The CACAO model is organized around four dimensions: (1) Change, (2) Adopters, (3) Change Agents, and (4) Organization (Dormant, 2011). Change is a new idea, process, or system that you want a group of people to accept (e.g., adoption of RBIS). Adopters are the group of people that is targeted to adopt the change (e.g., instructors of introductory STEM courses). Change agents are the people and team trying to enact
<table>
<thead>
<tr>
<th>Table 6.1. Malleable factors included within this study situated within the TCSR framework and hypotheses about how factors impact the adoption of RBIS with relevant citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malleable Factor</td>
</tr>
<tr>
<td><strong>Department characteristics</strong></td>
</tr>
<tr>
<td>Discipline</td>
</tr>
<tr>
<td>Highest degree awarded</td>
</tr>
<tr>
<td><strong>Department appointment expectations</strong></td>
</tr>
<tr>
<td>Teaching load</td>
</tr>
<tr>
<td>Tenure status</td>
</tr>
<tr>
<td>Student e of teaching</td>
</tr>
<tr>
<td>Assessment of teaching performance</td>
</tr>
<tr>
<td><strong>Classroom contextual</strong></td>
</tr>
<tr>
<td>Class size</td>
</tr>
<tr>
<td>Classroom setup</td>
</tr>
<tr>
<td>Class size × classroom setup</td>
</tr>
<tr>
<td>Decision making</td>
</tr>
<tr>
<td>Personal factors</td>
</tr>
</tbody>
</table>
Teaching-focused coursework

Doctoral and postdoctoral training and coursework cover a variety of topics, such as learning theory, effective practices, and instructional design, which provide instructors a foundation for informed teaching decisions. (Windschitl and Sahl, 2002; Lotter et al., 2007; Southerland et al., 2011a; Hora, 2012)

Teaching-related workshops

Workshops disseminate new teaching pedagogies and can give instructors first-hand experience with RBIS. (Landis et al., 1998; Clark et al., 2002; Peace et al., 2002; Burke et al., 2004; Lotter et al., 2007; Manduca et al., 2010; Murray et al., 2011; Lund and Stains, 2015; Fukawa-Connelly et al., 2016; Stegall et al., 2016; Manduca et al., 2017; Stains et al., 2018; Viskupic et al., 2019; Houseknecht et al., 2020; Riihimaki and Viskupic, 2020; Viskupic et al., 2022)

Teaching-related new faculty experiences

Experiences and workshops for new faculty members spread awareness and advocate for the adoption of RBIS. (Wood and Gentile, 2003; Handelsman et al., 2004; Henderson, 2008; Ebert-May et al., 2011; Henderson et al., 2012; Baker et al., 2014; Ebert-May et al., 2015; Stains et al., 2015; Derling et al., 2016; Beane et al., 2020; Emery et al., 2021)

Teacher thinking

Growth mindset

Instructors holding a growth mindset have been reported to use more student-centered approaches. (Rattan et al., 2012; Aragón et al., 2018; Johnson et al., 2018; Bathgate et al., 2019a; Canning et al., 2019; Ferrare, 2019; Yik et al., 2022)

Satisfaction with student learning

Dissatisfaction with current student learning or belief that students may learn better with alternative pedagogies may spur a revision of teaching and result in the adoption of new teaching strategies. (Feldman, 2000; Windschitl and Sahl, 2002; Gess-Newsome et al., 2003; Lotter et al., 2007; Southerland et al., 2011a; Southerland et al., 2011b; Bauer et al., 2013; Andrews and Lemons, 2015; Gibbons et al., 2018; Erdmann et al., 2020; Riihimaki and Viskupic, 2020)

change (e.g., administrators, educational policy makers, educational researchers, instructional designers). The organization encompasses the change, adopters, and change agents. Organizational contexts and influences affect how change agents lead change initiatives in hopes that adopters accept or implement the change. In a higher education setting, departments, colleges/schools, and the institution all have influence on individuals’ beliefs and behaviors, which can influence social norms and thus change in an organization.

In operationalizing the dimensions of the CACAO model, we investigate change as the state of teaching transformations through the uptake of RBIS, adopters as the STEM instructors that amend their instruction to include using RBIS in their teaching practices, change agents as the individuals advocating for adoption of RBIS, and the organization as the members of higher education institutions. For adopters, Dormant (2011) outlines five stages of adoption: (1) awareness, (2) curiosity, (3) mental tryout, (4) hands-on tryout, and (5) adoption.

The TCSR framework and the CACAO model work in tandem to understand institutional change. The CACAO model is designed to aid organizational (i.e., institutional) change agents (i.e., change practitioners and leaders) to understand dimensions of change (Dormant, 2011); in
this study, we focus on malleable factors that influence the dimensions of change. The TCSR framework focuses on teachers’ beliefs, which influence teacher practices, as the center for change that occurs within the institution (Gess-Newsome et al., 2003); contextual factors, personal factors, and teacher thinking factors are described as components in this larger institutional change context. Factors modeled in this study are situated within the components TCSR framework and influence the adopter’s (i.e., instructor’s) uptake of RBIS (i.e., the change) within the organization (i.e., institution). Evaluation of these malleable factors allows for insights and recommendations for change agents to provide opportunities that meet the needs of adopters in the different stages of adoption. Therefore, the TCSR and CACAO models are congruent and complementary, and together, situate our study.

6.5 Methods

6.5.1 Respondents

Target courses are general chemistry, single-variable calculus, and quantitative-based introductory physics. Target institutions are two-year associate-degree granting institutions in the United States that offer all three of these target courses, and four-year bachelor’s and/or graduate degree-granting institutions that have conferred at least one bachelor’s degree in all three disciplines (i.e., chemistry, mathematics, and physics) between 2011–2016 as recorded by the National Center for Education Statistics’ Integrated Postsecondary Education Data System. Target participants are primary instructors for one of the three target courses that was not taught exclusively online in the 2017–18 or 2018–19 academic years in the United States.

A database of the target instructors was assembled through stratified random sampling centered around target institution type; the objective was to construct a representative sample of two-year institutions, four-year institutions, and universities to capture the different types of degree-granting institutions (i.e., associate, bachelor’s, and graduate, respectively). The database was constructed by the American Institute of Physics Statistical Research Center using
publicly available online information and by contacting department chairs at the target institutions. The database contains 18,337 instructors that have met these criteria and is comprised of 8,933 instructors at two-year associate-degree granting institutions and 9,404 instructors at four-year bachelor’s and/or graduate degree-granting institutions.

6.5.2 Data collection

Previous large-scale studies in postsecondary chemistry (Gibbons et al., 2018; Stains et al., 2018), mathematics (Johnson et al., 2018; Apkarian et al., 2019), and physics (Henderson and Dancy, 2009; Walter et al., 2016; Walter et al., 2021) informed the development of the survey instrument. The survey is comprised of five main elements: (1) course context, (2) instructional practices, (3) awareness and usage of active learning instructional techniques, (4) perceptions, beliefs, and attitudes related to students, learning, and departmental context, and (5) personal demographics and experience. Previous instruments and scales with reliability and validity evidence were used where applicable, e.g., mindset (Dweck et al., 1995) and the EBIP Adoption scale (Landrum et al., 2017). Expert review of single-item survey constructs demonstrates content and face validity.

Survey instrument data were collected by the American Institute of Physics Statistical Research Center between March–May 2019 with approval from the Western Michigan University Institutional Review Board (application no. 17-06-10); informed consent was obtained digitally. Survey respondents included 3,769 instructors (20.5% unit response rate) consisting of 1,244 chemistry, 1,349 mathematics, and 1,176 physics instructors; there were 1,099 instructors at two-year institutions and 2,670 instructors at four-year institutions. A total of 1,466 respondents were removed from this analysis due to incomplete responses for all the survey items used in the construction of the multilevel models. This resulted in the study sample of 2,303 respondents including 768 chemistry, 751 mathematics, and 784 physics instructors from 1,371 departments.
at 741 institutions; of these 2,303 instructors, 599 instructors are at two-year institutions and 1,704 instructors are at four-year institutions.

A full list of survey items used in this study can be found in our previous work (see Yik et al., 2022). The RBIS Adoption Scale (described below) resulted in the binary outcome variables used in the multilevel logistic regression models (described below). Factors used in this survey are described above in Table 1. Discipline, highest degree offered by a department, and tenure status are dummy coded with the reference group as mathematics, associate degree, and instructors with no opportunity to earn tenure, respectively. Ordinal variables include teaching load, student evaluation of teaching, assessment of teaching performance, class size, growth mindset, and satisfaction with student learning; the ordinal scales for each of these variables are described in the Results (Table 4, below), with the exception of student evaluation of teaching (role of student evaluation of teaching in review, promotion, or tenure compared to other measures: 0 = not used, 1 = less weight, 2 = equal weight, 3 = more weight, 4 = only used). Decision making (i.e., respondent has sole decision-making authority/is in collaboration with others to make decisions), RBIS use as student, scholarship of teaching and learning, teaching-focused coursework, teaching-related workshops, and teaching-related new faculty experiences are all binary variables (yes/no; except decision making, previously described). Class size is grand-median centered at 30–39 students. Growth mindset is the average of three items on a six-point Likert scale (Dweck et al., 1995); items were reverse-coded, and values were centered at the middle of the scale. Satisfaction of student learning is a single item on a five-point Likert scale (very satisfied to very satisfied) and values were centered at the middle of the scale.

6.5.3 Research-based instructional strategies adoption scale

The RBIS Adoption Scale is an adaption of the EBIP Adoption Scale (Landrum et al., 2017); the sole difference is in wording of the instrument where “EBIP” is replaced by “RBIS” (Table 6.2). This instrument contains six items to be used as a Guttman scale with yes/no
Table 6.2. RBIS Adoption Scale with associated CACAO stage of adoption

<table>
<thead>
<tr>
<th>RBIS Adoption Scale Item</th>
<th>Score (number of “yes” responses)</th>
<th>CACAO Change Model (Dormant, 2011)</th>
<th>Modified CACAO Change Model (this work)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior to this survey, I already knew about RBIS.</td>
<td>0</td>
<td>Awareness</td>
<td>Awareness</td>
</tr>
<tr>
<td>I have thought about how to implement RBIS in my courses.</td>
<td>1</td>
<td>Awareness</td>
<td>Awareness</td>
</tr>
<tr>
<td>I’ve spent time learning about RBIS and I am prepared to use them.</td>
<td>2</td>
<td>Mental Tryout</td>
<td>Tryout</td>
</tr>
<tr>
<td>I consistently use RBIS in my courses.</td>
<td>3</td>
<td>Hands-on Tryout</td>
<td>Tryout</td>
</tr>
<tr>
<td>I consistently use RBIS and I continue to learn about and experiment with new RBIS.</td>
<td>4</td>
<td>Adoption</td>
<td>Adoption</td>
</tr>
<tr>
<td>I have evidence that my teaching has improved since I started using RBIS.</td>
<td>5</td>
<td>Adoption</td>
<td>Adoption</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Adoption</td>
<td>Adoption</td>
</tr>
</tbody>
</table>

responses. Guttman scales are unidimensional (i.e., is a measure of a single construct: degree of RBIS adoption), ordinal (i.e., items are ordered from the “least agreement” statement to the “most agreement” statement), and deterministic (i.e., results are analyzed based on the last statement the respondent agreed with) (Guttman, 1944).

Guttman scales are self-scoring and is indicated when the response pattern changes from yes to no. The coefficient of reproducibility (CR) is a measure of the reliability of a Guttman scale (Guttman, 1944). That is, items that do not yield the desired outcome are removed from the instrument and items that improve the desired outcome are kept; this process is similar to systematically removing items in a pool of items to increase inter-item reliability via Cronbach’s $\alpha$ in Likert subscales. One issue with the CR is its sensitivity to extreme marginal distributions; in other words, if there are extreme patterns of responses or if respondents respond with extreme patterns, such as answering all scale items with yes. The coefficient of scalability (CS) is a measure of the predictability of the scale, or the proportion of responses that can be correctly predicted from the row and column marginals, and was introduced to combat artificially high CR (Menzel, 1953; Guest, 2000). For evidence of unidimensionality in a Guttman scale, a CR > .90
(Guest, 2000; Aiken and Groth-Marnat, 2006; Abdi, 2010) and a CS > .60 are recommended standards (Menzel, 1953; Guest, 2000).

To provide evidence for unidimensionality, CR and CS values are calculated for the survey sample and for each discipline (i.e., chemistry, mathematics, and physics). For all data that is used in this study, every respondent provided full responses for the RBIS Adoption Scale (i.e., no items were left blank). Responses were ordered, and scale errors and marginal errors were calculated (Guest, 2000; Aiken and Groth-Marnat, 2006) to compute CR and CS values. The RBIS Adoption Scale has acceptable reliability for the study sample and the three STEM disciplines included in this study—sample: CR = 0.974, CS = 0.960; chemistry: CR = 0.992, CS = 0.988; mathematics: CR = 0.974, CS = 0.960; physics: CR = 0.973, CS = 0.965.

### 6.5.4 Multilevel models

Models were constructed using the melogit package in Stata version 17 (StataCorp, 2021) using mean and variance adaptive Gauss–Hermite quadrature (mvaghermite) integration with seven integration points. Predictor variables used in these models have been previously reported (Yik et al., 2022) through a review of the research literature on malleable factors that affect the uptake of active learning strategies from a variety of STEM disciplines.

Three-level models were used to evaluate the association of malleable factors on the stages of RBIS adoption in introductory chemistry, mathematics, and physics. Multilevel models are advantageous when considering that participants are not independent from one another as they can be grouped by department or institutions, and thus, violate the assumption that observations are independent (Raudenbush and Bryk, 2002; Snijders and Bosker, 2012); instructors (level 1) are nested within departments (level 2) that are nested within institutions (level 3). Other studies (e.g., Porter and Umbach, 2001; Smart and Umbach, 2007; Yik et al., 2022) use and advocate for these three-level models in similar contexts.
Two multilevel models are reported herein: the Tryout Model and Adoption Model. The Tryout Model includes sample of 1,079 respondents including 360 chemistry, 430 mathematics, and 289 physics instructors from 826 departments at 579 institutions; of these 1,079 instructors, 327 instructors are at two-year institutions and 752 instructors are at four-year institutions. The Adoption Model includes sample of 1,757 respondents including 591 chemistry, 501 mathematics, and 665 physics instructors from 1,118 departments at 663 institutions; of these 1,757 instructors, 419 instructors are at two-year institutions and 1,338 instructors are at four-year institutions.

6.6 Results

6.6.1 Multilevel models

We report the results of a national survey on the stages of RBIS adoption in introductory chemistry, mathematics, and physics in the United States. Data were collected and are modeled using three-level regression models based on the nested nature of the instructors (level 1) within departments (level 2) at institutions (level 3).

Multilevel regression models are used to explain two outcomes: RBIS tryout (i.e., Tryout Model) and RBIS adoption (i.e., Adoption Model). In the EBIP Adoption Scale by (Landrum et al., 2017), each item is mapped onto one of the CACAO adoption stages (Dormant, 2011); in this study, we combine the mental tryout and hands-on tryout stages into single stage, tryout (see Table 6.2 above). Strategies suggested by the CACAO model for mental tryout include demonstrating examples of change and highlighting success, and strategies suggested for hands-on tryout include providing training, information, and resources (Dormant, 2011). We combine the two tryout stages because strategies are commonly employed together in teaching-related workshops and experiences (Henderson, 2008; Ebert-May et al., 2011; Baker et al., 2014), and we also combine the three distinct adoption stages because instructors in these later stages have been reported to share similar characteristics and teaching practices (Viskupic et al., 2022). The
distribution of RBIS awareness, tryout, and adoption of the respondents in this study is given in Table 6.3.

Two multilevel logistic regression models are used to evaluate the association of the 17 malleable factors with RBIS tryout and RBIS adoption. Two models are used to distinguish between awareness and tryout (i.e., Tryout Model), and also, tryout and adoption (i.e., Adoption Model). We differentiate the two multilevel logistic regression models into what could be a single multinomial logistic regression model; however, this would result in a set of two different regression coefficients from a reference stage of adoption. A single multilevel ordinal logistic regression model could also be used to model the data, but this model assumes the odds between each of the adoption stages are equivalent and proportional. Previous work has demonstrated that the needs of instructors are different depending on what stage of adoption they are at (Viskupic et al., 2022). For more parsimonious and interpretable regression coefficients from which we can provide recommendations to change agents, we therefore model two different outcomes (i.e., tryout and adoption) using two multilevel logistic regression models.

The intraclass correlation coefficient (ICC) is an index of the proportion of variance in the outcome variable that is explained by groups or clusters and is the ratio of the between-group variance and the total variance (Raudenbush and Bryk, 2002; Snijders and Bosker, 2012). Calculations for ICC values were performed as outlined in Liu (2015). The unconditional Tryout Model has an ICC of 0.11 for level 2 (department) and an ICC of 0.08 for level 3 (institution), meaning roughly, 11% of the variation in the outcome variable (i.e., stage of adoption: awareness or tryout) is accounted for by nesting instructors within departments and 8% of the variation in the outcome variable is accounted for by nesting departments within institutions. The unconditional
Adoption Model has an ICC of 0.08 for level 2 and 0.04 for level 3. Small ICC values suggest that a two-level model would be appropriate; however, model misspecification results in less accurate fixed effect and standard error estimates, and inflation of lower-level variance estimates (Chen, 2012). Therefore, we specify the data using the more conceptually appropriate three-level model.

Seventeen factors, comprised of ten contextual, five personal, and two teacher thinking factors, are categorized using the TCSR model (Woodbury and Gess-Newsome, 2002; Gess-Newsome et al., 2003) in previous work (Yik et al., 2022). We report the results of two multilevel logistic regressions (i.e., RBIS Tryout and RBIS Adoption) using odd ratios (OR) which indicate the strength of the association between a predictor variable with the outcome variable (i.e., RBIS Tryout or Adoption) when all other predictor variables are accounted for in the model and held constant (Table 6.4).

Odds ratios are interpreted as the number of times higher when a variable is not zero when all other variables are held constant. For example, in the Tryout Model, academic discipline (i.e., chemistry, mathematics, physics) is evaluated using mathematics as the reference. When all other variables are held constant, the odds of RBIS tryout versus RBIS awareness are not statistically different for instructors in chemistry departments (OR = 0.92, p > .05) than instructors in mathematics departments. However, the odds of RBIS tryout versus RBIS awareness are 2.13 times higher for instructors in physics departments than for instructors in mathematics departments when all other variables are held constant. Odds ratios for the Adoption Model can be interpreted analogously where the odds ratio now represents the odds of RBIS adoption versus RBIS tryout.

The Tryout Model and the Adoption Model both share overlap in the statistically significant factors, but the models are also different. Differences in the models further demonstrate (and corroborate the assumption) that two multilevel models better represent the data than a single multilevel ordered logistic regression model (i.e., proportional odds model). For the Tryout Model, statistically significant malleable factors include physics, class size, the interaction effect between
class size and classroom setup, RBIS use as a student, teaching-related workshops, and satisfaction with student learning. For the Adoption Model, statistically significant malleable factors include physics, assessment of teaching performance, class size, classroom setup, RBIS use as a student, scholarship of teaching and learning, teaching-related new faculty experiences, growth mindset, and satisfaction with student learning.

### 6.6.2 Association of stages of RBIS adoption with percent lecturing

Our previous work (Yik et al., 2022) detailed the association of these malleable factors with percent lecturing. Stage of RBIS adoption and percent lecturing are two measures of active learning and comparison of these outcomes can provide further insight into the similarities and differences between these models. Table 6.5 presents the association of the stage of RBIS adoption with percent lecturing. A Kendall’s tau-b $(\tau_b)$ correlation was run to determine the relationship between the stage of RBIS adoption and percent lecturing among the 2,266 respondents that provided usable data for all survey items used in this study including the RBIS Adoption Scale and percent lecturing. There is a medium-sized, negative association between stage of RBIS adoption and percent lecturing ($\tau_b = -0.335, p < 0.001$); increasing adoption of RBIS is associated with a decrease in percent lecturing.

Visualizations comparing the magnitude of the each of the instructor-level malleable factors between the Tryout Model and Percent Lecturing Model (Figure 6.1A) and the Adoption Model and Percent Lecturing Model (Figure 6.1B) are provided. In summary, nine malleable factors are statistically significantly associated with percent time lecturing: class size; classroom setup; the interaction effect between class size and classroom setup; decision making; RBIS use as a student; scholarship of teaching and learning; teaching-related workshops, new faculty experiences, and coursework; and growth mindset (Yik et al., 2022).
## Table 6.4. Factors associated with RBIS Tryout and Adoption

<table>
<thead>
<tr>
<th>Factor</th>
<th>RBIS Tryout</th>
<th></th>
<th>p</th>
<th>RBIS Adoption</th>
<th></th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry</td>
<td>0.92</td>
<td>0.20</td>
<td>0.708</td>
<td>1.28</td>
<td>0.19</td>
<td>0.112</td>
</tr>
<tr>
<td>Physics</td>
<td>2.13</td>
<td>0.47</td>
<td>0.001</td>
<td>2.06</td>
<td>0.31</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Bachelor’s program</td>
<td>1.02</td>
<td>0.26</td>
<td>0.950</td>
<td>1.14</td>
<td>0.19</td>
<td>0.436</td>
</tr>
<tr>
<td>Graduate program</td>
<td>0.47</td>
<td>0.14</td>
<td>0.009</td>
<td>1.21</td>
<td>0.24</td>
<td>0.327</td>
</tr>
<tr>
<td>Teaching load</td>
<td>1.16</td>
<td>0.09</td>
<td>0.069</td>
<td>0.94</td>
<td>0.05</td>
<td>0.300</td>
</tr>
<tr>
<td>Tenured faculty</td>
<td>1.17</td>
<td>0.24</td>
<td>0.445</td>
<td>0.81</td>
<td>0.11</td>
<td>0.125</td>
</tr>
<tr>
<td>Tenure-track faculty</td>
<td>0.92</td>
<td>0.26</td>
<td>0.761</td>
<td>0.96</td>
<td>0.18</td>
<td>0.814</td>
</tr>
<tr>
<td>Student evaluation of teaching</td>
<td>1.06</td>
<td>0.09</td>
<td>0.457</td>
<td>0.93</td>
<td>0.06</td>
<td>0.243</td>
</tr>
<tr>
<td>Assessment of teaching performance</td>
<td>1.07</td>
<td>0.11</td>
<td>0.510</td>
<td>1.23</td>
<td>0.09</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>+1 (not influential)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+2 (somewhat influential)</td>
<td></td>
<td></td>
<td></td>
<td>1.51</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>+3 (influential)</td>
<td></td>
<td></td>
<td></td>
<td>1.85</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>+4 (very influential)</td>
<td></td>
<td></td>
<td></td>
<td>2.28</td>
<td>0.16</td>
</tr>
<tr>
<td>Class size</td>
<td>0.34</td>
<td>0.03</td>
<td>0.67</td>
<td>0.04</td>
<td>0.82</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>+1 (20–29 students)</td>
<td></td>
<td>0.59</td>
<td>0.05</td>
<td>1.22</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>+2 (40–99 students)</td>
<td></td>
<td>2.92</td>
<td>0.25</td>
<td>1.50</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>+3 (100+ students)</td>
<td></td>
<td>4.98</td>
<td>0.42</td>
<td>1.84</td>
<td>0.11</td>
</tr>
<tr>
<td>Classroom setup</td>
<td>0.95</td>
<td>0.24</td>
<td>0.839</td>
<td>2.29</td>
<td>0.53</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Class size × classroom setup</td>
<td>0.97</td>
<td>0.08</td>
<td>0.685</td>
<td>0.04</td>
<td>0.82</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>+1 (20–29 students)</td>
<td></td>
<td>0.40</td>
<td>0.05</td>
<td>1.22</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>+2 (40–99 students)</td>
<td></td>
<td>5.25</td>
<td>0.63</td>
<td>1.84</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>+3 (100+ students)</td>
<td></td>
<td>12.26</td>
<td>1.45</td>
<td>1.84</td>
<td>0.11</td>
</tr>
<tr>
<td>Decision making</td>
<td>1.59</td>
<td>0.41</td>
<td>0.070</td>
<td>0.87</td>
<td>0.11</td>
<td>0.135</td>
</tr>
<tr>
<td>RBIS use as a student</td>
<td>6.34</td>
<td>2.72</td>
<td>&lt; 0.001</td>
<td>2.15</td>
<td>0.38</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Scholarship of teaching and learning</td>
<td>1.01</td>
<td>0.20</td>
<td>0.937</td>
<td>2.25</td>
<td>0.28</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Teaching-focused coursework</td>
<td>0.88</td>
<td>0.16</td>
<td>0.458</td>
<td>1.21</td>
<td>0.15</td>
<td>0.128</td>
</tr>
<tr>
<td>Teaching-related workshops</td>
<td>4.85</td>
<td>1.33</td>
<td>&lt; 0.001</td>
<td>1.46</td>
<td>0.34</td>
<td>0.105</td>
</tr>
<tr>
<td>Teaching-related new faculty experiences</td>
<td>1.07</td>
<td>0.19</td>
<td>0.709</td>
<td>1.51</td>
<td>0.19</td>
<td>0.001</td>
</tr>
<tr>
<td>Growth mindset</td>
<td>1.15</td>
<td>0.08</td>
<td>0.055</td>
<td>0.51</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+1.5 (moderate fixed mindset)</td>
<td></td>
<td>0.66</td>
<td>0.04</td>
<td>1.15</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>+0.5 (slight fixed mindset)</td>
<td></td>
<td>0.87</td>
<td>0.05</td>
<td>1.50</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>+1.5 (slight growth mindset)</td>
<td></td>
<td>1.97</td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+2.5 (strong growth mindset)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfaction with student learning</td>
<td>1.32</td>
<td>0.13</td>
<td>0.58</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.15</td>
<td>0.11</td>
<td>0.76</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.87</td>
<td>0.09</td>
<td>1.31</td>
<td>0.09</td>
<td>&lt; 0.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.76</td>
<td>0.08</td>
<td>1.73</td>
<td>0.12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a* Ordinal variable; *p*-values are omitted for additional ordinal levels.  
*b* Not statistically significant variable; odds ratios are omitted for ordinal levels.  
*c* Variable with interaction effect; variable should only be interpreted if the other variable is held constant; the interaction effect odds ratio accounts for the interaction effect and the individual fixed effects.
Table 6.5. Association of RBIS stages of adoption with percent lecturing

<table>
<thead>
<tr>
<th>Stage of Adoption</th>
<th>Sample (n = 2,266)</th>
<th>Chemistry (n = 761)</th>
<th>Mathematics (n = 732)</th>
<th>Physics (n = 773)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Awareness</td>
<td>68.79</td>
<td>22.52</td>
<td>69.65</td>
<td>22.58</td>
</tr>
<tr>
<td>Tryout</td>
<td>65.70</td>
<td>21.66</td>
<td>68.09</td>
<td>20.74</td>
</tr>
<tr>
<td>Adoption</td>
<td>46.56</td>
<td>22.99</td>
<td>50.51</td>
<td>23.62</td>
</tr>
</tbody>
</table>

In comparing the Tryout Model and Percent Lecturing Model (Figure 6.1A), there are fewer instructor-level malleable factors associated with RBIS tryout than percent lecturing. Similarly, across the models, experience with RBIS as a student and participation in teaching-related workshops are associated with RBIS tryout and a decrease in percent lecturing. Inversely, larger class sizes are associated with RBIS tryout, but an increase in percent lecturing. Additionally, the interaction effect of larger class sizes in rooms that allow for group work and dissatisfaction with student learning are associated with RBIS tryout but are not significantly associated with percent lecturing.

In comparing the Adoption Model and Percent Lecturing Model (Figure 6.1B), there are nearly an equal number of instructor-level malleable factors associated with RBIS tryout as percent lecturing, however, there are differences. Classroom allowing for group work is the factor most strongly associated with RBIS adoption and decrease in percent lecturing. Other strongly associated factors in both models include engagement in SOTL/DBER, experience with RBIS as a student, participation in teaching-related workshops and new faculty experiences, and holding a growth mindset. Similarly, larger class sizes are associated with adoption of RBIS and increase in percent lecturing. However, satisfaction with student learning is associated with RBIS adoption, but is not significantly associated with percent lecturing. Two factors, shared decision making and previous teaching-focused coursework, are associated with a decrease in percent lecturing, but are not significantly associated with RBIS adoption.
6.7 Discussion

Multiple factors are associated with an increase in tryout and adoption of RBIS at the instructor level, when all other factors are held constant: assessment of teaching performance, class size, classroom setup, the interaction effect between class size and classroom setup, RBIS use as a student, participation in the scholarship of teaching and learning or discipline-based education research, teaching-related workshops, teaching-related new faculty experiences, holding a growth mindset, and satisfaction with student learning. Factors unique to either the Tryout Model or the Adoption Model allow for tangible recommendations for instructors in each group (i.e., those seeking to tryout an RBIS or those seeking to formally adopt an RBIS). Department-level factors (e.g., discipline and highest degree awarded) are unchangeable and thus lack formal implications in the context of this study; however, instructor-level factors can lead to tangible and practical implications for change efforts, and therefore, we will focus our discussion on these malleable instructor-level factors.

6.7.1 Importance of classroom spaces

Instructors continually note that teaching large classes in large fixed-seating classrooms (i.e., auditorium-style lecture halls) make it difficult to promote student engagement and use RBIS (Henderson and Dancy, 2007; Hora, 2012; Lund and Stains, 2015; Sturtevant and Wheeler, 2019). Additionally, classrooms that can accommodate for group work, such as active learning classrooms or rooms with movable seats, desks, or tables, provide environments that are conducive to fostering student-student and student-instructor interactions (Beichner et al., 2007; Beichner, 2008; Cotner et al., 2013; Lund and Stains, 2015; Foote et al., 2016; Knaub et al., 2016). Our results indicate that instructors in the tryout and adoption stages are more likely to use RBIS as the class size becomes larger when compared to instructors in the awareness and tryout stages, respectively. However, instructors in the tryout stage are more likely than instructors in the awareness stage to implement RBIS regardless of the classroom space, but instructors in the
adopter stage need (and potentially require or demand) classrooms that allow for group work. While class sizes have been reported to ease the facilitation of student engagement (Bressoud et al., 2015), our results indicate that instructors are less likely to use RBIS for smaller (< 30 students) class sizes.

Active learning, though, can occur in any classroom environment. Increasing odds of an instructor using RBIS when teaching larger courses suggest that uptake of RBIS is a means to make class sizes seem smaller by encouraging discussion and collaboration (Robert et al., 2016; Beane et al., 2020; Raker et al., 2021). For example, when large auditorium-style lecture halls are fitted with swivel chairs, students have been demonstrated to outperform their counterparts in a fixed-seat lecture hall (Ogilvie, 2008; Condon et al., 2016), which may be attributed to the room layout promoting discussion and collaboration. It has also been reported in the literature that various levels of student–student interactions can still be implemented with large class sizes in auditorium-style rooms (Lund et al., 2015).

Teaching large enrollment courses in classrooms that allow for group work is critical for instructors trying out RBIS. Classroom spaces designed to accommodate flexible teaching spaces can help aid in reducing barriers to implementing active learning (Ellis et al., 2016). These spaces may also help sustain the adoption of active learning practices due to the efforts it takes to learn about and support the implementation of these RBIS (Knaub et al., 2016). Instructors that have adopted RBIS may request to teach in such spaces because the space facilitates the use of active learning pedagogies, which can lead to further and sustained adoption (Foote et al., 2016).

6.7.2 Perceived value of assessment of teaching performance

Instructors’ perceived value of how their department or institution values the assessment of teaching performance is influential in their pedagogies. Incentives guide instructors’ professional decisions: for example, departments and institutions may require instructors to adopt RBIS in their teaching as a part of review, promotion, or tenure packages (Lund and Stains, 2015).
Figure 6.1. Association of instructor-level malleable factors between (A) RBIS Tryout and Percent Lecturing Models and (B) RBIS Adoption and Percent Lecturing Models. *p < 0.05, **p < 0.01, ***p < 0.001. Scaled magnitudes of odds ratios are shown on the left vertical axis (RBIS Tryout & Adoption) and scaled magnitudes of percentages are shown on the right vertical axis (Percent Lecturing). Additional levels for ordinal variables are not shown. The gray horizontal line indicates an odds ratio of one (RBIS Tryout & Adoption) and zero percent change (Percent Lecturing). Values above the horizontal line indicate increased odds of (A) RBIS Tryout or (B) RBIS Adoption and decreased percent time lecturing; values below the horizontal line indicate increased odds of (A) RBIS Awareness or (B) RBIS Tryout and increased percent time lecturing. Connecting lines indicate association between the same significant factor in both models, or between a factor that is significant in one model and not the other model.
Alternatively, if there are no structures in place to evaluate and reward instructors’ teaching, then there can be little external incentive to adopt RBIS in their classes (Hativa, 1995; Walczyk et al., 2007; Brownell and Tanner, 2012; Elrod and Kezar, 2017; Shadle et al., 2017; Johnson et al., 2018). Our results suggest that greater the perceived influence of assessment performance on instructors’ review, promotion, or tenure, the greater the odds of adoption of RBIS.

While instructors may be knowledgeable about and tryout RBIS, active learning strategies are most effective when instructors are committed to the pedagogy and are provided with ongoing support (Bressoud et al., 2015). For instructors to perceive emphasis on teaching, departmental (e.g., faculty and chairs) and institutional support (e.g., college, provost, dean) are needed for sustained change (Henderson and Dancy, 2007; Shadle et al., 2017; Carney et al., 2021; Dunnigan and Halcrow, 2021; Mingus and Koelling, 2021). Support of these endeavors can also come from other departments, centers for teaching and learning, and professional organizations (Mingus and Koelling, 2021). Departments and institutions can showcase value of teaching by rewarding instructors for their efforts in learning about, trying out, and adopting active learning strategies (Fairweather, 2008; Seymour et al., 2011; Wieman, 2015). Institutions can also support emphasis on adopting RBIS by providing travel support to external teaching-focused professional development programs and workshops (Reinholz and Apkarian, 2018; Carney et al., 2021), or small stipends or service credits to engage in institutional programs (Lotter et al., 2007; Foote et al., 2016; Herman et al., 2018; Reinholz and Apkarian, 2018).

6.7.3 Experience as a student in a course using research-based instructional strategies

The attitudes, beliefs, and intentions of instructors are partly a result of their own undergraduate and graduate education shaping the way that instructors currently enact pedagogies (Oleson and Hora, 2014; Lund and Stains, 2015; Fukawa-Connelly et al., 2016). Many instructors may have experienced more traditional, lecture-based instructional modalities as students, and thus, they may imitate these teaching styles in their own teaching practice.
(Adamson et al., 2003). However, instructors who experienced active learning as a student have been reported to be more likely to implement active learning in their classes as an instructor (Lund and Stains, 2015; Yik et al., 2022). Our results support this idea; instructors who experienced RBIS as a student are more likely to tryout and adopt RBIS as an instructor. Experience with RBIS as a student is the highest influencing factor on RBIS tryout in our model; it is the among the highest for RBIS adoption, alongside class size and classroom effects.

The next generation of instructors, thus, will hopefully be greatly influenced by how we teach today. Our results point to the notion that “we teach the way we were taught” (Mazur, 2009). However, it is possible that instructors who use RBIS are more likely to recall the teaching strategies they experienced as a student, and those instructors who do not use RBIS are less likely to recall RBIS experiences when they were a student. Regardless, for lasting sustainable change in adopting active learning practices in introductory STEM courses, current instructors must adopt RBIS in their classes to influence the thinking of future instructors (i.e., current undergraduate and graduate students). Therefore, instructors must be incentivized to participate in teaching-related professional development and the scholarship of teaching and learning to assist in the adoption of RBIS.

6.7.4 Engagement in teaching-related professional development

Teaching-related workshops have many different forms from being broader, such as how to manage a classroom and use the learning management system, to being more specific, such as on how to implement active learning strategies (e.g., Aebersold, 2019; Miller et al., 2021). Workshops have largely placed emphasis on develop–disseminate efforts to change individual instructor’s teaching practices (Beach et al., 2012; Borrego and Henderson, 2014), which stems from the belief that teaching is an individualistic effort (Tanner and Allen, 2006; Lane et al., 2019). Some studies suggest that these models for instructional change are generally ineffective (Henderson et al., 2011), and therefore, may not result in the desired widespread changes
(Fairweather, 2008; Austin, 2011; Kezar, 2011; Froyd et al., 2017). However, our results demonstrate that participation in workshops aids instructors in trying out RBIS. Additionally, while participation in teaching-related new faculty experiences may not initially play a role in RBIS tryout, the effects of these experiences are prolonged by exhibiting a significant role in RBIS adoption.

Change can take place at many different levels, ranging from an individual instructor (Steinert et al., 2006), to departments (AACU, 2014), and entire institutions (Elrod and Kezar, 2016). While change strategies may target a certain level, we recommend that change agents at all levels work together and support one another to achieve desired changes (AAAS, 2019). Instructors should be active participants in their teaching roles and in learning about and adopting new teaching pedagogies (Fairweather, 2008; Smith et al., 2014; Quan et al., 2019). However, there can be a lack of incentives for instructors to participate in teaching-related professional development (Walczyk et al., 2007; Brownell and Tanner, 2012). External motivators to attend professional development events can greatly contribute to influencing instructors. Institutions and departments need to support, highly incentivize, and reward participation in these opportunities (Seymour et al., 2011; AAAS, 2019; Bathgate et al., 2019b), and even stipends can act as a token of appreciation for instructors’ time and giving priority classroom preferences may sustain the adoption of RBIS (Soto and Marzocchi, 2021).

Continuing professional development is critical for the sustained implementation of active learning approaches (Speer and Wagner, 2009; Camburn, 2010; Quan et al., 2019; Pilgrim et al., 2021). Change in instructors' beliefs and practices occur slowly (Derting et al., 2016). To facilitate successful change, interventions (e.g., teaching-related professional development) should last a semester or longer (Henderson et al., 2011). As time progresses, instructors may regress back to old teaching habits. It is necessary for instructors to continually engage with opportunities to learn about and demonstrate new teaching pedagogies (Brownell and Tanner, 2012). These opportunities should also be diverse across a continuum to assist instructors in their incremental
growth in implementing active learning strategies (Soto and Marzocchi, 2021); our results show that instructors at the differing stages of adoption (i.e., awareness, tryout, and adoption) have different needs, and therefore, programs must be designed for instructors at various stages (Austin and Sorcinelli, 2013; Borda et al., 2020). Both teaching-related new faculty experiences and broader workshops are crucial for tryout and adoption of RBIS, and our results suggest that such programs should continue to be funded and operated.

Centers for teaching and learning serve as important resources for instructors to learn about, obtain advice on, and get help to properly implement RBIS. Additionally, these centers can support sustained adoption of RBIS through faculty learning communities (Cox, 2004; Pelletreau et al., 2018; Shadle et al., 2018; Dancy et al., 2019) and communities of practice (Henderson et al., 2017; Tomkin et al., 2019; Benabentos et al., 2021). One method is to incorporate faculty learning communities or communities of practice as a component of an extended (i.e., longer than one semester) new faculty experience (Beane-Katner, 2013). At institutions where centers for teaching and learning do not exist, instructor-organized communities may be effective at facilitating adoption of RBIS (Ma et al., 2019). Instructors can discuss teaching pedagogies through observation and feedback (Gormally et al., 2014; Smith et al., 2014) or successful adopters of RBIS may also engage with instructors trying out RBIS as form of peer-coaching (Desimone and Pak, 2017; Ma et al., 2018).

### 6.7.5 Participation in the scholarship of teaching and learning

Engagement in the scholarship of teaching and learning (SOTL) has been reported to be related with improvements in course transformations (Henderson et al., 2011) and it is believed that practicing the closely related field of discipline-based education research (DBER) yields similar results (Henderson et al., 2012). Work has shown that instructors engaged in SOTL employed instructional practices that were more student-focused (Pelletreau et al., 2018; Dancy et al., 2019; Tomkin et al., 2019; Benabentos et al., 2021; Yik et al., 2022). In our study,
participation in SOTL or DBER was associated with the RBIS adoption, but not RBIS tryout. Engagement in SOTL is an approach to develop reflective educators (Henderson et al., 2011). By agreeing to the statement, “I have evidence that my teaching has improved since I started using RBIS” in the RBIS Adoption Scale (Table 2), instructors are testifying to engaging in SOTL in their teaching. This leads to the ambiguity whether adoption of RBIS leads to SOTL or engaging in SOTL leads to the adoption of RBIS; regardless, instructors are employing active learning strategies and are becoming reflective educators.

Instructors participating in SOTL or DBER can help advance the adoption of RBIS of other instructors. Our previous study (Yik et al., 2022) detailed the association of shared decision on instructional methods with decreased time lecturing, and thus, increased time spent on active learning. Other work (Lane et al., 2020) suggested that instructors predominantly talk to other instructors with similar teaching approaches, i.e., RBIS users talk with other RBIS users. One approach to guide instructors to higher stages of RBIS adoption is through co-teaching (e.g., Henderson et al., 2009) or course coordination (e.g., Apkarian and Kirin, 2017), which can involve instructors at earlier or mid-stages of RBIS adoption (i.e., awareness and tryout) with SOTL, and also support sustained change and continuous course improvement over an extended period of time (Marbach-Ad et al., 2007; Marbach-Ad et al., 2014; Reinholz and Apkarian, 2018; Mingus and Koelling, 2021).

6.7.6 Holding a growth mindset

Instructors’ mindset has been shown to influence pedagogical decisions (Rattan et al., 2012; Aragón et al., 2018; Bathgate et al., 2019a; Canning et al., 2019; Ferrare, 2019; Richardson et al., 2020). Instructors holding a fixed mindset (i.e., student intelligence is fixed and cannot be changed; Dweck, 1999) adopt fewer active learning practices (Rattan et al., 2012; Aragón et al., 2018), and lecture more (Yik et al., 2022); correspondingly, instructors holding a growth mindset (i.e., student intelligence is malleable and can be improved with time and experience; Dweck,
are more willing to consider (Johnson et al., 2018) and adopt more active learning practices (Aragón et al., 2018), and lecture less (Yik et al., 2022).

Our results show that mindset is not significantly associated with trying out RBIS but holding a growth mindset is significant in the adoption of RBIS; this finding mirrors previous findings that instructors’ growth mindset beliefs is associated with a decrease in time spent lecturing (Yik et al., 2022). Fixed mindset beliefs are associated with greater odds of RBIS tryout than adoption, which may explain why instructors leave the innovation-decision process when using RBIS after an initial implementation (Henderson et al., 2012). Alternatively, espousing a growth mindset may increase instructors’ persistence when adopting RBIS (Limeri et al., 2020).

Professional development workshops and experiences are possible avenues to promote instructors’ growth mindset (Pilgrim et al., 2021; Yik et al., 2022). However, instructors have differences in their levels of motivation to participate in teaching-related professional development (Woodbury and Gess-Newsome, 2002; Bouwma-Gearhart, 2012; McCourt et al., 2017). Interventions to promote growth mindset for students have been demonstrated to be effective, generalizable, and replicable (Dweck and Leggett, 1988; Dweck, 1999; Yeager et al., 2016; Bettinger et al., 2018; Yeager et al., 2019), and psychosocial interventions can be leveraged to motivate instructors to participate in professional development that can emphasize growth mindset beliefs (see Limeri et al., 2020). This is particularly imperative because instructors’ perceived mindsets have effects on students; students who perceive their instructors exhibiting growth mindsets are reported to have higher academic success, and positive motivational and psychological outcomes (Rattan et al., 2018; Fuesting et al., 2019; Lou and Noels, 2020; Muenks et al., 2020; LaCrosse et al., 2021), and this perception is strengthened with using active learning practices (Muenks et al., 2021). By holding a growth mindset, instructors are more likely to adopt RBIS and active learning, and resultingly, positively influence student outcomes.
6.7.7 (Dis)satisfaction with student learning

The misalignment of teaching practices with instructional goals and student outcomes can result in instructors’ dissatisfaction with student learning or current instructional methods (Southerland et al., 2011a, 2011b). The result of this disconnect between teacher thinking and practice can consequently lead to the adoption of new teaching strategies, such as RBIS (Feldman, 2000; Windschitl and Sahl, 2002; Gess-Newsome et al., 2003; Andrews and Lemons, 2015; Lund and Stains, 2015; Gibbons et al., 2018).

Our findings suggest that dissatisfaction with student learning is associated with the tryout of RBIS and the satisfaction with student learning is associated with the adoption of RBIS. Instructors’ dissatisfaction with student learning may stem from dissatisfaction with the current pedagogy, which can include current RBIS use, and result in a change of teaching practices, and therefore, the trying out using (different) RBIS (Gess-Newsome et al., 2003). It can be postulated that once instructors are satisfied with their students’ learning, it is due to the implemented pedagogies, and this satisfaction can lead to the sustained adoption of RBIS.

Change agents can leverage instructors’ dissatisfaction or satisfaction with students learning in change strategies. For example, dissatisfied instructors are more likely to try out RBIS. These instructors would be prime targets to engage in teaching-related workshops to learn about RBIS, be shown evidence of success using RBIS, and be provided training and resources to implement RBIS (Dormant, 2011). Additionally, satisfied instructors are more likely to adopt RBIS. These instructors can not only benefit from learning about and experimenting with new RBIS, but more importantly, having evidence that teaching has improved since using RBIS (Landrum et al., 2017). Change agents can identify these instructors as participants to engage in the scholarship of teaching and learning; by engaging in SOTL, instructors will have evidence that shows supported student learning, and thus, instructor satisfaction.
6.7.8 Association of malleable factors with RBIS tryout, adoption, and percent lecturing

Malleable factors associated with RBIS tryout and RBIS adoption are similar, but also different, to the malleable factors associated with percent lecturing. In all three regression models, only one factor has an association with higher odds of RBIS tryout and adoption and a decrease percent lecturing: experience in a course using RBIS as a student. Therefore, it is vital we ensure that current instructors adopt RBIS such that our current students (i.e., our future instructors) are more likely to implement and adopt RBIS and active learning strategies in the future.

Tryout and adoption of RBIS and percent lecturing are different measures of active learning. To such an end, the malleable factors used to consider change strategies differ depending on the desire to influence RBIS tryout, RBIS adoption, or percent lecturing. Regardless of the exact malleable factors chosen to inform change, any of the factors significantly associated with any of the three outcomes (i.e., RBIS tryout, RBIS adoption, and percent lecturing) have greater odds of increasing the uptake of active learning strategies.

6.8 Limitations

Findings from this study are constrained by three noteworthy limitations. First, this study is limited to the disciplines of chemistry, mathematics, and physics, and is therefore limited in disciplinary scope; additionally, survey respondents were comprised of instructors from the introductory courses of these disciplines, and is therefore limited in course scope. Other survey-based studies illustrate differences between lower- and upper-level STEM courses (Benabentos et al., 2021) and observation-based studies report instructors are more likely to implement RBIS in introductory courses than in more advanced courses (Lund et al., 2015). Inclusion of additional STEM disciplines and courses would be in productive space for future studies.

Second, higher education is a complex system. The TCSR framework focuses on instructor beliefs and instructors as the primary change agents at an institution (Gess-Newsome et al., 2003). In our conceptualization of the TCSR framework, we situate barriers and motivations
for instructional change into contextual factors, personal factors, and teacher thinking factors; however, it would be unfeasible to model every single possible factor in a regression model. For example, our study does not consider student-level factors. Some students are reported to find active learning classes to be disjointed with an overall feeling of frustration and confusion and many studies note student resistance as a barrier to implement active learning strategies (Deslauriers et al., 2019; Owens et al., 2020).

Lastly, plausible reliability threats may be due to the self-reported nature of our survey items. While we have provided evidence for the reliability of the instrument (i.e., RBIS Adoption Scale) used to obtain the outcome measure (i.e., stage of RBIS adoption), there may be reliability threats to survey items that correspond to respondents’ contextual, personal, and teacher thinking factors. Nonetheless, studies have demonstrated evidence to suggest that self-reported data regarding teaching practices align well with observational data (Durham et al., 2018; Gibbons et al., 2018), and data from observation-based studies would complement data from survey studies.

6.9 Conclusions

We advocate for the sustained adoption of research-based instructional strategies to promote use of active learning in introductory STEM courses. There are boundless paths instructors and change agents can take to that lead to a student-centered classroom which incorporates active learning strategies, and instructors should adopt RBIS that best fits the characteristics of their unique classroom contexts (Budd et al., 2013; Beane et al., 2020). In this journey on instructional change, instructors must be first aware of RBIS, tryout these strategies, and then adopt the pedagogies. The RBIS Adoption Scale is a quick, efficient, and reliable instrument to gauge instructors' stage of RBIS adoption; using this information, change agents can help facilitate instructors' course transformation journeys. We offer recommendations for change agents to provide directed opportunities for instructors.
Recommendations that target instructors aware of RBIS and are working toward trying out RBIS:

- Identify instructors that are dissatisfied with their students’ learning
- Design workshops that provide training and resources to implement RBIS, and highlight the success stories and research-based data of those RBIS
- Encourage instructors that teach large class sizes to seek out classrooms that allow for group work and introduce RBIS that large class sizes appear smaller

Recommendations that target instructors trying out RBIS and are working toward adopting RBIS:

- Demonstrate the need to assess student learning with evidence when RBIS are implemented
- Engage instructors in SOTL so that instructors have the knowledge and resources to obtain evidence of enhanced student learning
- Encourage instructors to teach in classrooms that allow for group work, irrespective of class sizes
- Help foster growth mindset beliefs through interventions or other professional development programs

Recommendations for institutional leaders and policy makers:

- Value and reward teaching and instructors’ efforts in instructional change
- Build classroom spaces that allow for group work
- Encourage and incentivize the participation in teaching-related new faculty experiences and workshops
- Showcase benefits of RBIS and promote the uptake of RBIS use

The results from our national survey of introductory chemistry, mathematics, and physics instructors that evaluates the association of malleable factors with the adoption of active learning inform these recommendations for change agents, institutional leaders, and policy makers. Our
goal is for these recommendations to inform instructor-focused change initiatives that result in meaningful and sustainable change around teaching practices.

6.10 Conflict of interest

The authors declare that they have no competing interests.

6.11 Author contributions

CH, MHD, EJ, JRR and MS obtained financial support and conceived of the study. CH, MHD, EJ, JRR, MS, and NA designed the survey. NA cleaned and compiled the data. BJY, JRR, NA, and MS analyzed the data. BJY carried out the formal analysis of the data. BJY wrote the paper. All authors read, edited, and approved the final manuscript.

6.12 Funding

This material is based upon work supported by the National Science Foundation under Grant Nos. DUE 1726126 (JRR), 2028134 (MS), 1726328 (CH), 1726042 (MHD), and 1726281 (EJ). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

6.13 Acknowledgments

The authors would like to thank Eun Sook Kim (University of South Florida) for aid in the development and interpretation of the multilevel models. We would like to thank all the instructors that participated in this study.

6.14 References


Canning, E. A., Muenks, K., Green, D. J., and Murphy, M. C. (2019). STEM faculty who believe ability is fixed have larger racial achievement gaps and inspire less student motivation in their classes. *Science Advances*, 5:eaau4734. doi: 10.1126/sciadv.aau4734


Marbach-Ad, G., Briken, V., Frauwirth, K., Gao, L.-Y., Hutcheson, S. W., Joseph, S. W., Mosser, D., Parent, B., Shields, P., Song, W., Stein, D. C., Swanson, K., Thompson, K. V., Yuan, R. and Smith, A. C. (2007). A faculty team works to create content linkages among various courses to increase meaningful learning of targeted concepts of microbiology. *CBE—Life Sciences Education*, 6, 155-162. doi: 10.1187/cbe.06-12-0212


264


268


curriculum materials to community college technical courses. *Community College Journal
of Research and Practice*, 31, 583-601. doi: 10.1080/10668920701428881

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E.,
Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K.,
Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., Hahn, P.
national experiment reveals where a growth mindset improves achievement. *Nature*, 573,
364-369. doi: 10.1038/s41586-019-1466-y

S. (2016a). Using design thinking to improve psychological interventions: The case of the
growth mindset during the transition to high school. *Journal of Educational Psychology*,
108, 374-391. doi: 10.1037/edu0000098

Yik, B. J., Raker, J. R., Apkarian, N., Stains, M., Henderson, C., Dancy, M. H., and Johnson, E.
(2022). Evaluating the impact of malleable factors on percent time lecturing in gateway
chemistry, mathematics, and physics courses. *International Journal of STEM Education*,

Education*, 12, 55-76. doi: 10.1080/13562510601102131
Chapter 7

Conclusion

Work presented in this dissertation advances understanding in students’ explanations of reaction mechanisms in organic chemistry and in evaluating the uptake of research-based instructional strategies in introductory chemistry, mathematics, and physics.

The first part of this work (Chapters 3 and 4) is on students’ explanations of reaction mechanisms in organic chemistry. This work started with a constructed response item initially developed by Cooper et al. (2016) that asked students what is happening in an aqueous acid–base proton transfer reaction and why it happens. The constructed-response item was later modified by Dood et al. (2018) and used to develop a computer-based predictive scoring model for use of the Lewis acid–base model. In Chapter 3, the work further builds upon previous work by Dood et al. (2018, 2019) by outlining the development of a generalized computer-based predictive scoring model for the correct use of the Lewis acid–base model. For the development of other predictive scoring models, the work in Chapter 3 can be advanced by broadening the scope of reactions involved (i.e., multi-step reactions versus single-step acid–base reactions) and through the scoring techniques used (i.e., multiclass classification versus binary classification). However, rubrics grounded in the research literature for this multiclass classification must first be developed; Chapter 4 details the development of a generalized rubric for explanation sophistication for nucleophiles.

The second part of this work (Chapters 5 and 6) is on evaluating the uptake of research-based instructional strategies (RBIS) in introductory chemistry, mathematics, and physics. Building upon previous large-scale studies (Henderson and Dancy, 2007; Walter et al., 2016, 2021; Gibbons et al., 2018; Johnson et al., 2018; Stains et al., 2018; Apkarian et al., 2019), a
multi-institution, large-scale survey of postsecondary introductory-level chemistry, mathematics, and physics instructors in the United States was conducted in Spring 2019. Multilevel models were constructed to explain the effects of malleable factors on the uptake of RBIS. Two outcomes are modeled; the first outcome, percent lecturing, is reported in Chapter 5, and the second outcome, stages of RBIS adoption, is reported in Chapter 6. This chapter will summarize the work in this dissertation and the broader implications for instruction, research, and policy.

7.1 Student explanations of reaction mechanisms

7.1.1 Summary of results

Chapter 3 describes the development of a machine learning-based tool to evaluate correct Lewis acid–base model use in written responses to open-ended formative assessment items. This predictive model was developed from constructed-response item that asked students what is happening and why it happens for aqueous proton transfer reactions and reactions that can only be explained using the Lewis acid–base model. In addition, the model was also developed from items that asked students to explain properties of acids and bases: why a molecule can act as an acid, base, or both (i.e., amphoteric). Responses were classified as either correct use or incorrect/non-use of the Lewis acid base model. A support vector machines-based machine learning algorithm was used to train the generalized predictive scoring model. Results indicate that accuracies greater than 84% can be achieved through a variety of validation procedures and is generalizable to new reaction type (i.e., non-aqueous proton transfer reactions) that was included in the training data. These findings are a proof-of-concept that illustrate generalized predictive scoring models can be developed and be used to formatively assess students’ responses to open-ended constructed-response items.

Chapter 4 describes the development of a generalized rubric to evaluate level of explanation sophistication for nucleophiles in written responses to open-ended formative assessment items. These assessment items once again asked students what is happening and
why it happens in a variety of reaction mechanisms (\(N = 85\)) spanning a year-long organic chemistry course. This generalized rubric was synthesized by combining aspects of other frameworks used to understand mechanistic understanding (Sevian and Talanquer, 2014; Cooper et al., 2016; Caspari et al., 2018; Bodé et al., 2019; Crandell et al., 2020; Dood et al., 2020a, 2020b) and is applicable in an assortment of contexts: stages of an organic chemistry course, across nucleophile and reaction types, and across prompt variations. Additionally, a case study illustrates how this rubric can be used as an assessment tool to inform learning and instruction.

7.1.2 Implications for instruction

**Assessments send explicit messages about what is important.**

Assessments convey clear messages to students about what is important in a course and discipline (Crooks, 1988; Entwistle, 1991; Scouller and Prosser, 1994; Scouller, 1998; Momsen et al., 2013). Therefore, instructors should design and choose assessments that demonstrate what they want their students to take away from their courses (Holme et al., 2010). Assessments that allow for student success through memorization or surface-level thinking convey messages that critical thinking or skills that translate beyond the scope of the course are not important (Cooper, 2016; Stowe and Cooper, 2017, 2019).

Deliberately selected assessments can engage students in scientific practices that are essential for learning and doing science (NRC, 2012; NGSS Lead States, 2013). In particular, the Framework for K-12 Science Education and the Next Generation Student Standards emphasize constructing explanations as one of the essential practices. The process of constructing written explanations helps students think deeper and engage more with course material (Rivard, 1994; Bangert-Drowns et al., 2004; Reynolds et al., 2012). Therefore, open-ended written assessment should ask students what is happening and why it happens to promote deeper learning (Bell and Cowie, 2001; Cooper, 2015; Cooper et al., 2016; Caspari et al., 2018; Underwood et al., 2018). For example, the work described in Chapters 3 and 4 suggests that assessments with the scaffold
asking students what is happening in reaction mechanisms and why it happens conveys to students that understanding what and why is important. Constructing explanations is central to chemistry and science, instructors should design learning environments, which includes assessments, that support students in this practice (DeGlopper et al., 2022).

*Instructors should evaluate assessment results through reflective practices that lead to improved instruction.*

Assessments are the means through which instructors gain insight into students’ understanding. This new insight can be used to improve instruction, and therefore, improve student learning (Bell and Cowie, 2001; Fies and Marshall, 2006; MacArthur and Jones, 2008; Gibbons et al., 2022). Instructors should engage in the “scholarship of assessment,” which includes ensuring that students have sufficient formative assessment tasks (Rust, 2007). Two tenets of the scholarship of assessment are that assessment feedback is timely and assessments communicate clear and high expectations (Rust, 2007).

Assessment tools presented in this dissertation include the generalized predictive model for correct use of the Lewis acid–base model and the generalized rubric for levels of sophistication of nucleophiles. These assessment tools exemplify the aforementioned tenets of scholarship of assessment and can be used by instructors to engage in reflective practices (Henderson et al., 2011). Reflective practices can lead to greater awareness about one’s teaching (Osterman, 1990); that is, instructors should be encouraged to consistently use a critical lens on their teaching considering the results of assessments with the goal of learning from these teaching and learning experiences to improve instruction (Kane et al., 2004; McAlpine et al., 2017). For example, in Chapter 4, a case study is provided and demonstrates how assessment results using the nucleophile rubric can be iteratively used to inform learning and instruction.

Interactive web applications can be developed to automate assessment scoring and provide assessment results for reflective practices. For example, the R code for generalized
predictive model (Yik and Raker, 2021) described in Chapter 3 can be integrated into an
interactive web application using the “shiny” package in R (Chang et al., 2022). These applications
can be used to quickly score students’ written responses and provide students with information
about their understanding and provide instructors with data about students’ understanding. These
assessment data provide a mean to be used in reflective practices to consider broader teaching,
learning, and assessment transformations.

7.1.3 Implications for research

Written assessments can uncover students’ understanding.

The copious number of written responses collected from the variety of constructed-
response items exposed how students describe what is happening in reaction mechanisms and
why it is happening. In Chapter 3, the majority of students are able to correctly use the Lewis
acid–base model to explain acid–base reaction mechanisms, and why molecules can act as an
acid, base, or both. When students were asked to explain why reactions occur, the majority of
students were unable to do so at instructor-defined target levels of explanation sophistication for
the concept of nucleophiles in Chapter 4. Differences in the levels of explanation sophistication
were observed for different nucleophile types (e.g., lone pair, sigma bond, and pi bond) and
reaction families (e.g., carbocation intermediate, aromatic reactions, nucleophilic additions, etc.).
These findings suggest that researchers can qualitatively elucidate the differences in
understanding for these different types of nucleophiles in a variety of reaction contexts. Written
assessments using constructed-response items like those presented in this dissertation can be
used to evaluate understanding in for different reaction components (e.g., electrophile, leaving
groups, etc.) using the levels of explanation sophistication framework such that instructors have
more insight into students’ understanding of reaction mechanisms. More broadly, the what and
why constructed-response scaffold can be used beyond the scope of organic chemistry reaction
mechanisms, and students can be asked to explain other chemical or scientific phenomena (e.g., Becker et al., 2016; Noyes and Cooper, 2019; Noyes et al., 2020).

Effects of constructed-response prompts on students’ written responses is a possible area of exploration. In Chapter 4, two different constructed-response prompts (e.g., Original and More Cued) were used to elicit students’ responses about reaction mechanisms. Future work can investigate how different amounts of cueing affect students’ level of explanation sophistication. Cueing students may activate resources about deeper-level chemical phenomena, such as electrostatic interactions, or scaffold responses, such that prompts ask students to identify components of a chemical reaction before writing about why reactions happen. Cueing effects may cause students to think more deeply about chemical phenomena beyond surface-level explanations. For example, a very cued prompt can involve providing examples of roles of reactants/intermediates (e.g., electrophile, nucleophile, and leaving group). In addition, in lieu of asking students to “explain why each of the reactants and intermediates interact,” the prompt can suggest using electronic properties (e.g., electronegativity and polarity) to explain reactivity. It can be hypothesized that these cues provoke deeper mechanistic reasoning.

Constructed-response prompts can also vary from the what and why scaffold. In Chapters 3 and 4, the what and why scaffold was used to encourage deeper understand and reasoning about reaction mechanisms. However, other researchers have used different prompts to develop students’ scientific argumentation skills through written assessments by using contrasting cases (Caspari et al., 2018; Graulich and Schween, 2018; Caspari and Graulich, 2019), likelihood of similar mechanistic steps (Bodé et al., 2019), and scaffolds with alternative reaction pathways (Lieber and Graulich, 2022; Lieber et al., 2022a, 2022b). For example, Caspari and Graulich (2019) asked students to predict the reaction has a lower activation energy from two very similar reactions with slight differences (e.g., leaving group moiety and degree of substitution for a carbocation); this difference encourages students to think more deeply about the phenomena beyond the surface-level difference. These other prompts and question types can be developed
into endless possibilities of constructed-response items that can be used to uncover students’ understanding.

Generalized predictive models can be developed and used to evaluate students’ written responses.

Constructed-response items from a variety of prompts relating to acid–base chemistry were used to develop a generalized predictive scoring model. Results presented in Chapter 3 show overall high predictive scoring accuracies across many unique reactions and molecules. Once generalized predictive scoring models are developed, they can be used to efficiently score new sets of student response data across a broad topic. The high accuracies for the generalized predictive model for correct use of the Lewis acid–base suggest promising avenues for the further development of more generalized predictive models that efficiently and accurately evaluate students’ understanding of other chemical concepts. For example, nucleophiles and electrophiles (Anzovino and Bretz, 2015, 2016), and leaving groups (Popova and Bretz, 2018) are possible pieces and parts of reaction mechanisms that can be explored and generalized predictive models developed. In particular, the levels of explanation sophistication rubric for nucleophiles described in Chapter 4 can serve as the classification outcomes for a generalized predictive model for students’ understanding of nucleophiles. Chemistry is not the limit; generalized predictive models can also be developed for other scientific concepts and phenomena building upon other predictive models developed for a single prompt (Haudek et al., 2011; Haudek et al., 2012; Prevost et al., 2013; Dood et al., 2020a; Noyes et al., 2020). Predictive models have the potential to change how instructors assess students and their progress over time, and these tools can give the ability for students to have greater control over their own learning.
7.2 Uptake of research-based instructional strategies

7.2.1 Summary of results

Chapter 5 describes how malleable contextual, personal, and teacher thinking factors influence instructors’ pedagogical decisions. These malleable factors were used as variables to model the percent time lecturing in introductory chemistry, mathematics, and physics courses. Multilevel modeling analyses indicate that several malleable factors are statistically associated with percent time lecturing when controlling for all other factors included in the study: class size, classroom setup, shared decision making, previous RBIS experience as a student, engagement in SOTL, taking teaching-focused coursework, participation in teaching-focused workshops and new faculty experiences, and holding a growth mindset. Findings provide tangible recommendations for instructors and administrators to influence the adoption of more active learning strategies.

Chapter 6 describes a parallel study on how these same malleable factors influence a different outcome variable: stages of RBIS adoption. Three stages of RBIS adoption are described: awareness, tryout, and adoption. Multilevel modeling analyses are also used to evaluate these malleable factors by modeling RBIS tryout and RBIS adoption. Results indicate that several malleable factors are associated with RBIS tryout when controlling for all other factors included in the study: class size, the interaction effect between class size and classroom setup, previous RBIS experience as a student, participation in teaching-focused workshops, and dissatisfaction with student learning. In addition, results also indicate that several and similar malleable factors are associated with RBIS adoption when controlling for all other factors included in the study: perceived influential importance of assessment of teaching performance in annual review, promotion, or tenure; class size; classroom setup; previous RIBS experience as a student; engagement in SOTL; participation in teaching-related new faculty experiences; holding a growth mindset; and satisfaction with student learning. The RBIS Tryout and RBIS Adoption Models were also compared with the Percent Lecturing Model; these comparisons show that malleable factors
associated with all three models are similar, though distinct. Therefore, change agents should keep in mind the desired outcome of that change when considering malleable factors used to inform change strategies.

7.2.2 Implications for research

Quantitative, empirical support for the Teacher-Centered Systemic Reform (TCSR) framework provides a foundation for qualitatively investigating factors influencing enacted instructional practices.

Work presented in Chapters 5 and 6 on evaluating the uptake of RBIS in introductory STEM courses provides quantitative, empirical support for the TCSR framework as a model of instructional change. In particular, multilevel models demonstrate that factors within each of the three broad categories (i.e., contextual factors, personal factors, and teacher thinking factors) can be used to understand influences on instructors’ pedagogical decisions. While work in Chapters 5 and 6 are quantitative descriptions of how malleable factors are associated with active learning practices, these associations provide a rich area of research for broader understanding of how and why these factors influence instruction.

Other areas of exploration can include the interactions between the broad categories in the TCSR framework. One limitation of the work presented in this dissertation is the lack of empirical evidence to support these interactions. For example, the interaction between teacher thinking factors and personal factors. More specifically, findings in Chapters 5 and 6 indicate that instructors’ growth mindset about students are associated with decreases in percent time lecturing and adoption of RBIS, but how is this growth mindset instilled? One possible means is through professional development opportunities such as teaching-focused coursework and teaching-related workshops and new faculty experiences. For example, a future study could involve interviewing instructors that hold growth mindsets to investigate why they have these mindsets and the roots of these beliefs.
7.2.3 Implications for policy

Professional development opportunities are needed if instructors are to learn about and implement research-based instructional strategies.

Professional development programs can be comprised of teaching-related new faculty experiences and workshops, and teaching-focused coursework. Dissemination has been historically focused on develop–disseminate efforts to increase individual instructors’ awareness of RBIS and inclusion of these strategies in instructors’ teaching practices (Henderson, 2008; Beach et al., 2012; Borrego and Henderson, 2014; Ebert-May et al., 2015; Stains et al., 2015). Results in Chapters 5 and 6 indicate that these teaching-related or teaching-focused professional development opportunities are associated with either a reduction in percent time lecturing or the tryout or adoption of RBIS. Instructional change can take place at the individual, departmental, and institutional levels (Steinert et al., 2006; AACU, 2014; Elrod and Kezar, 2016), and therefore, change agents at all levels of the higher educational system must work together to develop programming to not only spread awareness of RBIS, but also to encourage the tryout and strive toward the sustained adoption of RBIS. Thus, national programs and new faculty experiences to spread RBIS awareness should be continued to be funded, and departments and institutions should urge and incentivize instructors to participate in these professional development programs.

Classroom spaces are needed to support the uptake of research-based instructional strategies.

Studies report that smaller class sizes in flexible learning spaces allow for the implementation of student-centered instructional practices (Cotner et al., 2013; Lund and Stains, 2015; Shadle et al., 2017), which contrast large class sizes in fixed classroom layouts that impede the implementation of RBIS (Gess-Newsome et al., 2003; Henderson and Dancy, 2007; Hora, 2012; Lund and Stains, 2015; Sturtevant and Wheeler, 2019). Findings in Chapters 5 and 6 indicate the uptake of RBIS is supported by teaching large introductory courses in classroom
spaces that allow for student group work. Therefore, purposeful design of classroom spaces should be advocated for.

Active learning can happen in any classroom environment, but is better enacted in more flexible classroom environments. For more facile and widespread adoption of RBIS, traditional classrooms should be reimagined. In the classroom design process, for large lecture auditoriums, the switch from forward-facing fixed seats to swivel seating has the potential to increase student engagement and enhance student outcomes (Ogilvie, 2008). For medium-sized classroom spaces, active learning classrooms have been effective at transforming learning (Beichner et al., 2007; Brooks, 2011, 2012; Van Horne et al., 2012; Knaub et al., 2016; Copridge et al., 2021); active learning classrooms promote student collaboration through the use of round tables, whiteboard space, and monitors or computers throughout the room. For smaller classrooms, moveable furniture (e.g., modern mobile chairs and tables paired with chairs on casters) promote student–student interactions and offer instructors flexibility to try out new instructional practices (Harvey and Kenyon, 2013). Additionally, studies show that students, irrespective of academic performance, report positive effects on creativity and innovation (Chiu and Cheng, 2017). While instructors believed that teaching in flexible classroom spaces improved student learning, administrators (e.g., department chairs, deans, provosts) are likely unaware of enhanced student outcomes and benefits of these rooms to promote adoption of active learning environments (Van Horne and Murniati, 2016). Students matter, instructors matter, and administrators must be made aware about how classroom spaces influence instruction and student learning.

7.3 Overall implications

The future of science education should strive toward the development of research-based, open-access assessment tools.

Studies in this dissertation describe the development of assessment tools: a generalized predictive model and generalized rubric for level of explanation sophistication. The R code for the
generalized predictive model reported in Chapter 3 is openly available for instructors to run their own analyses and for researchers to develop their own predictive models (Yik and Raker, 2021). Other efforts at developing research-based, open access assessment tools include the Automated Analysis of Constructed Response (AACR) group (beyonddimultiplechoice.org) which includes a database of constructed-response items; these items can be scored with automated predictive scoring models and summary reports are provided to instructors. While many constructed-response items are in this database, a limitation is that these predictive models are item-specific, and thus, generalized predictive models may be a promising route.

Assessment tools, such as rubrics, can help reduce the ambiguity present in the literature on the mechanistic reasoning (Dood and Watts, 2022) that is expected of students in organic chemistry reaction mechanisms. An advantage of generalized rubric is the provided operational definitions that is applicable to both researchers and instructors to aid in bridging the gap between research and practice. Rubrics have the ability to evaluate courses and departmental curricula due to its adaptability and generalizability to a variety of assessments and courses. Therefore, research-based, open-access assessment tools such as predictive models and rubrics should be developed.

*There exists a need for better characterization of instructional practices in upper-division STEM courses.*

Instructional practices have been widely studied primarily in introductory STEM courses (Lund and Stains, 2015; Matz et al., 2018; Stains et al., 2018; Gavassa et al., 2019; Harris et al., 2020), which have been identified as barriers to the completion of undergraduate STEM degrees. Chapters 5 and 6 work toward better characterizing instructional practices in these lower-division courses. However, upper-division courses have been continually neglected, even in these works and the data set from which Chapters 5 and 6 originated. Some studies demonstrate differences in instruction practices between lower- and upper-division courses (Lund et al., 2015; Benabentos
et al., 2021). Characterizing instructional practices in upper-division courses and increasing the adoption of RBIS across the entire STEM curriculum can therefore enhance student outcomes and retention throughout the undergraduate career (Freeman et al., 2014). While most studies target introductory courses, it is important to characterize and monitor the degree of instructional change throughout the undergraduate curriculum to ensure that instructional practices are continuously aligned in students’ undergraduate experiences. Thus, there is an opportunity to apply research methodologies used in introductory courses to characterize the progress of instructional change in upper-division courses. Findings from these studies can inform better change agents how to best facilitate RBIS adoption.

7.4 Summary

The work presented in this dissertation advances assessment tools related to students’ explanations of reaction mechanisms in organic chemistry and in understanding factors associated with the uptake of RBIS in introductory chemistry, mathematics, and physics. The study in Chapter 3 builds upon previously developed constructed-response items (Cooper et al., 2016; Dood et al., 2018; Dood et al., 2019) and develops a generalized predictive model to determine correct use of the Lewis acid–base model in students’ written responses. The study in Chapter 4 constructs and applies a generalized rubric for levels of explanation sophistication for nucleophiles in a variety of contexts in organic chemistry reaction mechanisms; this generalized rubric can serve as a potential classification scheme for a generalized predictive model for understanding of nucleophiles. The study in Chapter 5 evaluates the influence of a collective set of malleable factors on the instructors’ percent time lecturing in introductory STEM courses. Finally, the study in Chapter 6 evaluates this same set of factors on the tryout and adoption of RBIS, and compares these findings to the findings in Chapter 5. It is the hope that future researchers use the work presented in this dissertation to develop more research-informed assessment tools for organic chemistry reaction mechanisms and change agents use this work
to provide better guided opportunities and recommendations for more widespread adoption of RBIS.

7.5 References


286


Cooper, M. M. (2016). It is time to say what we mean. *Journal of Chemical Education* 93, 799-800. doi: 10.1021/acs.jchemed.6b00227


Crandell, O. M., Lockhart, M. A. and Cooper, M. M. (2020). Arrows on the page are not a good gauge: Evidence for the importance of causal mechanistic explanations about nucleophilic substitution in organic chemistry. *Journal of Chemical Education* 97, 313-327. doi: 10.1021/acs.jchemed.9b00815


292


Appendix A
Supporting Information for Chapter 3

A.1 Constructed response items used in the training, cross-validation, and split-validation sets

All mechanism questions were given the following prompt:

Part A: Describe in full what you think is happening on the molecular level for this reaction. Be sure to discuss the role of each reactant and intermediate.

Part B: Using a molecular level explanation, explain why this reaction occurs. Be sure to discuss why the reactants form the products shown.

[pt-HCl]: Consider the mechanism for the following acid–base reaction between water and hydrochloric acid to form hydronium and chloride ion.

\[
\begin{align*}
\text{H}_2\text{O} & \quad \text{H}^+\text{Cl}^- \\
\rightarrow & \\
\text{H}_3\text{O}^+ & \quad \text{Cl}^-
\end{align*}
\]

[pt-HBr]: Consider the mechanism for the following acid–base reaction between water and hydrobromic acid to form hydronium and bromide ion.

\[
\begin{align*}
\text{H}_2\text{O} & \quad \text{H}^+\text{Br}^- \\
\rightarrow & \\
\text{H}_3\text{O}^+ & \quad \text{Br}^-
\end{align*}
\]
[pt-HI]: Consider the mechanism for the following acid–base reaction between water and hydroiodic acid to form hydronium and iodide ion.

![Mechanism for the acid-base reaction between water and hydroiodic acid](image)

[lew-carbocation]: Consider the mechanism for the reaction between ethanol and 2-methyl-2-propanylium to form tert-butyl(ethyl)oxonium.

![Mechanism for the reaction between ethanol and 2-methyl-2-propanylium](image)

[lew-BMe₃]: Consider the mechanism for the reaction between bromide and trimethylborane to form bromotrimethylborate.

![Mechanism for the reaction between bromide and trimethylborane](image)

[lew-BF₃]: Consider the mechanism for the reaction between ammonia and trifluoroborane to form the ammonia-trifluoroborane adduct.

![Mechanism for the reaction between ammonia and trifluoroborane](image)

[lew-AlCl₃]: Consider the mechanism for the reaction between acetone and aluminum trichloride to form the acetone-aluminum trichloride adduct.

![Mechanism for the reaction between acetone and aluminum trichloride](image)
[ampho-EtOH]: Explain why ethanol, CH₃CH₂OH, can act as both an acid and a base.

[ampho-IPA]: Explain why isopropanol, (CH₃)₂CHOH, can act as both an acid and a base.

[acid-AlCl₃]: Explain why aluminum trichloride, AlCl₃, can act as an acid.

[acid-BH₃]: Explain why borane, BH₃, can act as an acid.

[acid-carbocation]: Explain why methylium, CH₃⁺, can act as an acid.

[base-py]: Explain why pyridine, C₅H₅N, can act as a base.

[base-NEt₃]: Explain why triethylamine, N(CH₂CH₃)₃, can act as a base.

[base-PPh₃]: Explain why triphenylphosphine, P(C₆H₅)₃, can act as a base.
A.2 Constructed response items used in the external validation set

All mechanism questions were given the following prompt:
Part A: Describe in full what you think is happening on the molecular level for this reaction. Be sure to discuss the role of each reactant and intermediate.
Part B: Using a molecular level explanation, explain why this reaction occurs. Be sure to discuss why the reactants form the products shown.

[pt-acetone]: Consider the mechanism for the reaction between diisopropylamide and acetone to form diisopropylamine and acetone enolate.

[pt-BzOH]: Consider the mechanism for the reaction between ethanethiolate and benzoic acid to form ethanethiol and benzoate.

[pt-ammonium]: Consider the mechanism for the reaction between methoxide and propanaminium to form methanol and propanamine.

[lew-BCl₃]: Consider the mechanism for the reaction between imidazole and trichloroborane to form the imidazole-trichloroborane adduct.
[lew-AcCl]: Consider the mechanism for the reaction between ethanamine and acetyl chloride to form chloro(ethylammonio)ethanolate.

![Mechanism for the reaction between ethanamine and acetyl chloride to form chloro(ethylammonio)ethanolate.]

[lew-CO2]: Consider the mechanism for the reaction between water and carbon dioxide to form carbonic acid.

![Mechanism for the reaction between water and carbon dioxide to form carbonic acid.]

[acid-BF3]: Explain why trifluoroborane, BF₃, can act as an acid.

![Structure of trifluoroborane, BF₃.]

[acid-ammonium]: Explain why cyclohexanaminium, (C₆H₁₁)NH₃⁺, can act as an acid.

![Structure of cyclohexanaminium, (C₆H₁₁)NH₃⁺.]

[base-HNEt₂]: Explain why diethylamine, (CH₃CH₂)₂NH, can act as a base.

![Structure of diethylamine, (CH₃CH₂)₂NH.]

[base-propoxide]: Explain why isopropoxide, (CH₃)₂CHO⁻, can act as a base.

![Structure of isopropoxide, (CH₃)₂CHO⁻.]

[ampho-fBuOH]: Explain why tert-butanol, (CH₃)₃COH, can act as both an acid and a base.

![Structure of tert-butanol, (CH₃)₃COH.]

301
## Appendix B

**Supporting Information for Chapter 4**

### B.1 Nucleophiles rubric

<table>
<thead>
<tr>
<th>Level</th>
<th>Description of level</th>
<th>Key features of level</th>
<th>Exemplar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absent</td>
<td>• No response</td>
<td>• The nucleophile is not identified</td>
<td>&quot;The first step of the S_N2 reaction is a nucleophilic attack followed by the loss of a leaving group. The interaction happens at the same time.&quot;</td>
</tr>
<tr>
<td></td>
<td>• Non-normative</td>
<td>• The nucleophile–electrophile reaction step is missing from the explanation</td>
<td></td>
</tr>
<tr>
<td>Descriptive</td>
<td>• Describes the nucleophile engaging in bond forming processes</td>
<td>• The nucleophilic molecule or atom is identified</td>
<td>&quot;The iodine is acting as a leaving group and detaching from the 1-iodopropane. At the same time, the cyanide is attacking the molecule and substituting the leaving group. The reactants and intermediates interact because this is a substitution reaction. Iodine is a good leaving group so therefore this triggers the substitution reaction. Prompting the cyanide to replace it.&quot;</td>
</tr>
<tr>
<td></td>
<td>• Simplistic description of bond forming processes</td>
<td>• Electrons are not used to describe nucleophilic behavior</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Nucleophilic behavior is described at a surface, atomic level</td>
<td>• Bond forming processes are described (e.g., &quot;nucleophile attacks/forms a bond with electrophile&quot;)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Describes the nucleophile engaging in bond forming processes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foundational</td>
<td>• Nucleophilic behavior is described at a surface, electronic level</td>
<td>• Electrons are central to nucleophilic behavior and arrows represent the movement of electrons</td>
<td>&quot;The I is a good leaving group and it is leaving while the electron pair on the carbon connected to the triple bond is attacking the carbon that the leaving group is leaving from. They all interact in the same step simultaneously.&quot;</td>
</tr>
<tr>
<td></td>
<td>• Explicit features are mentioned</td>
<td>• Example descriptors of electrons: sigma electrons/bond, pi electrons/bond, lone pair (and not alkene or double bond)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Implicit features may be mentioned, but not fully explained</td>
<td>• Bond forming processes are described using electrons (e.g., &quot;the lone pair on the nucleophile attacks/forms a bond with electrophile&quot;)</td>
<td></td>
</tr>
<tr>
<td>Complex</td>
<td>• Describes why the nucleophile is involved in bond forming processes</td>
<td>• Implicit electronic features are used to describe nucleophilic behavior</td>
<td>&quot;The cyanide is acting as a nucleophile attacking the partially positive alpha carbon kicking off the iodine leaving group, thus forming the product. The electron rich cyanide is attracted to the alpha carbon because of its partial positive charge.&quot;</td>
</tr>
<tr>
<td></td>
<td>• Nucleophilic behavior is described at a deeper, electronic level</td>
<td>• Examples of electronic properties: electron density, electronegativity, partial charges, molecular orbital descriptions (e.g., HOMO/LUMO)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Explicit features are used to infer implicit features that are sufficiently explained</td>
<td>• Bond forming processes are described using electrons and electronic properties (e.g., &quot;the high electron density on the nucleophile is attracted to the area of low electron density on the electrophile&quot;)</td>
<td></td>
</tr>
</tbody>
</table>
B.2 Reaction mechanisms

B.2.1 Alkyl halide

B.2.1.1 Bimolecular substitution (Sn2)

Consider the mechanism for the Sn2 reaction between (R)-sec-butyl 4-methylbenzenesulfonate and azide to form (S)-2-azidobutane.

Consider the mechanism for the Sn2 reaction between (S)-pentan-2-yl-4-methylbenzenesulfonate and benzenethiolate to form (R)-pentan-2-yl(phenyl)sulfane.

Consider the mechanism for the Sn2 reaction between 1-iodopropane and cyanide to form butyronitrile.
Consider the mechanism for the S$_{N}$2 reaction between isobutyl 4-methylbenzenesulfonate and ethoxide to form 1-ethoxy-2-methylpropane.

![S$_{N}$2 reaction mechanism](image)

B.2.1.2 Unimolecular substitution (S$_{N}$1)

Consider the mechanism for the S$_{N}$1 reaction between $t$-butyl bromide and ethanol to form ethyl $t$-butyl ether.

![S$_{N}$1 reaction mechanism](image)

Consider the mechanism for the S$_{N}$1 reaction between $t$-butyl iodide and ethanol to form ethyl $t$-butyl ether.

![S$_{N}$1 reaction mechanism](image)
Consider the mechanism for the $S_N1$ reaction between $t$-butyl chloride and ethanol to form ethyl $t$-butyl ether.

Consider the mechanism for the $S_N1$ reaction between 1-bromo-1-methylcyclohexane and methanol to form 1-methoxy-1-methylcyclohexane.

Consider the mechanism for the $S_N1$ reaction between (S)-3-bromo-2,3-dimethylpentane and methanol to form (S)-3-methoxy-2,3-dimethylpentane.
Consider the mechanism for the $S_{N}1$ reaction between ($R$)-3-chloro-3-methyl-1-pentene and ethanol to form ($R$)-3-ethoxy-3-methyl-1-pentene.

\[
\begin{align*}
\text{H}_3\text{C}\text{CH}_2\text{C} & \xrightarrow{\text{O}^-} \text{O}^-\text{CH}_2\text{CH}_3 \\
\text{+ enantiomer} & \xrightarrow{\text{H}} \text{OCH}_2\text{CH}_3 \\
\end{align*}
\]

Consider the mechanism for the $S_{N}1$ reaction between 1,1-difluoro-1-iodopropane and methanol to form 1,1-difluoro-1-methoxypropane.

\[
\begin{align*}
\text{CH}_3 & \xrightarrow{\text{O}^-} \text{O}^-\text{CH}_3 \\
\text{F} & \xrightarrow{\text{H}} \text{OCH}_3 \\
\end{align*}
\]

Consider the mechanism for the $S_{N}1$ reaction between dichloro(phenyl)methyl 4-methylbenzenesulfonate and isopropanol to form (dichloro(isopropoxy)methyl)benzene.

\[
\begin{align*}
\text{Ts} & \xrightarrow{\text{Cl}^-} \text{O}^-\text{CH}_3 \\
\text{Cl} & \xrightarrow{\text{H}} \text{OCH}_3 \\
\end{align*}
\]
**B.2.2 Alkene**

**B.2.2.1 Halogenation/hydrohalogenation**

Consider the mechanism for the halogenation of cyclohexene with bromine to form \((1R,2R)-1,2\)-dibromocyclohexane.

\[
\begin{align*}
\text{Cyclohexene} & \quad \text{Br} \quad \text{Br} \quad & \rightarrow & \quad \text{Cyclohexene} & \quad \text{Br} \quad \Theta \quad & \rightarrow & \quad \text{Cyclohexene} & \quad \text{Br} \quad \Theta \\
& & & & & + \text{enantiomer} \\
& & & & & & & + \text{enantiomer}
\end{align*}
\]

Consider the mechanism for the halogenation of \((Z)-2\)-pentene with chlorine to form \((2R,3R)-2,3\)-dichloropentane.

\[
\begin{align*}
\text{2-Pentene} & \quad \text{Cl} \quad \text{Cl} \quad & \rightarrow & \quad \text{2-Pentene} & \quad \text{Cl} \quad \Theta \quad & \rightarrow & \quad \text{2-Pentene} & \quad \text{Cl} \quad \Theta \\
& & & & & + \text{enantiomer} \\
& & & & & & & + \text{enantiomer}
\end{align*}
\]

Consider the mechanism for the hydrohalogenation of 2-methyl-2-butene with hydrobromic acid to form 2-bromo-2-methylbutane.

\[
\begin{align*}
\text{2-Methyl-2-butene} & \quad \text{H} \quad \text{Br} \quad & \rightarrow & \quad \text{2-Methyl-2-butene} & \quad \text{H} \quad \Theta \quad & \rightarrow & \quad \text{2-Methyl-2-butene} & \quad \text{Br} \quad \Theta \\
& & & & & + \text{HBr}
\end{align*}
\]

Consider the mechanism for the halogenation of 2-methylpropene with bromine to form 1,2-dibromo-2-methylpropane.

\[
\begin{align*}
\text{2-Methylpropene} & \quad \text{Br} \quad \text{Br} \quad & \rightarrow & \quad \text{2-Methylpropene} & \quad \text{Br} \quad \Theta \quad & \rightarrow & \quad \text{2-Methylpropene} & \quad \text{Br} \quad \Theta \\
& & & & & + \text{Br}
\end{align*}
\]

Consider the mechanism for the hydrohalogenation of 1,1-dichloroethane with hydrochloric acid to form 1,1,1-trichloroethane.

\[
\begin{align*}
\text{1,1-Dichloroethane} & \quad \text{H} \quad \text{Cl} \quad & \rightarrow & \quad \text{1,1-Dichloroethane} & \quad \text{H} \quad \Theta \quad & \rightarrow & \quad \text{1,1-Dichloroethane} & \quad \text{H} \quad \Theta \\
& & & & & + \text{HCl}
\end{align*}
\]
B.2.2.2 Halohydrin formation

Consider the mechanism for the formation of a halohydrin from cyclopentene and chlorine in the presence of water to form \((1R,2R)-2\)-chlorocyclopentanol.

\[
\begin{align*}
\text{Cyclopentene} & \xrightarrow{\text{Cl}} \text{Cyclopentyl chloride} \\
\text{Cyclopentyl chloride} & \xrightarrow{\text{H}_2\text{O}} \text{Cyclopentanol} + \text{enantiomer}
\end{align*}
\]

Consider the mechanism for the formation of a halohydrin from \((Z)-4\)-methyl-2-pentene and bromine in the presence of water to form \((2S,3S)-3\)-bromo-4-methyl-2-pentanol.

\[
\begin{align*}
\text{4-Methyl-2-pentene} & \xrightarrow{\text{Br}} \text{4-Methyl-2-pentenyl bromide} \\
\text{4-Methyl-2-pentenyl bromide} & \xrightarrow{\text{H}_2\text{O}} \text{4-Methyl-2-pentanol} + \text{enantiomer}
\end{align*}
\]
Consider the mechanism for the formation of a halohydrin from methylcyclopentene and bromine in the presence of water to form \((1R,2R)-2\text{-bromo-1-methylcyclo-1-pentanol.}\)

\[\text{Consider the mechanism between } (Z)-5\text{-methyl-4-heptenol and hydronium ion to form the cyclic ether product.}\]

\[\text{Consider the mechanism for the acid-catalyzed hydration of 2-methylbutene to form 2-methyl-2-butanol.}\]
Consider the mechanism for the acid-catalyzed hydration of 2-methylpropene with hydronium to form tert-butanol.

Consider the mechanism for the acid-catalyzed hydration of methylenecyclohexane to form 1-methylcyclo-1-hexanol.

**B.2.3 Alkyne**

**B.2.3.1 Alkylation**

Consider the mechanism for the reaction of 3-methyl-1-butyne with amide and then iodoethane to form 2-methyl-3-hexyne.
B.2.4 Alcohol

B.2.4.1 Formation of a good leaving group then $S_N2$

Consider the mechanism for the reaction of ethanol and phosphorus tribromide to form bromoethane.

Consider the mechanism for the reaction between 2-methylpropanol and hydrobromic acid to form 1-bromo-2-methylpropane.

Consider the mechanism for the reaction between 2-propanol and phosphorus tribromide to form 2-bromopropane.
Consider the mechanism for the reaction between 1-propanol and thionyl chloride to form 1-chloropropane.

\[
\begin{align*}
\text{OH} \quad &\quad \text{Cl} : \text{SO} : \text{Cl} \\
\rightarrow &\quad \text{OH} \quad \text{Cl} : \text{SO} : \text{Cl} \\
\rightarrow &\quad \text{Cl} \quad \text{SO} : \text{Cl} \\
\rightarrow &\quad \text{Cl} \quad + \text{SO}_2 \quad + \text{2} \quad \boxed{\text{PhCl}}
\end{align*}
\]

B.2.4.2 Epoxidation/ring-opening

Consider the mechanism for the epoxidation of \((1R,2R)-2\)-chlorocyclohexanol with hydroxide and its subsequent reaction with thiophenolate to form \((1S,2S)-2\)-(phenylthiol)cyclohexanol.

**Step 1:**

\[
\begin{align*}
\text{Cl} \quad \text{H} \quad \text{O} \quad &\quad \text{O} \quad \text{Cl} \\
\rightarrow &\quad \text{Cl} \quad \text{H} \quad \text{O} \\
\rightarrow &\quad \text{H} \quad \text{O}
\end{align*}
\]

**Step 2:**

\[
\begin{align*}
\text{Cl} \quad \text{SPh} \quad &\quad \text{OH} \quad \text{H} \quad \text{O} \\
\rightarrow &\quad \text{Cl} \quad \text{SPh} \quad \text{OH} \\
\rightarrow &\quad \text{Cl} \quad \text{SPh} \quad \text{OH} \\
\rightarrow &\quad \text{Cl} \quad \text{SPh} \quad \text{OH} \\
\rightarrow &\quad \text{Cl} \quad \text{SPh} \quad \text{OH} \quad \text{enantiomer} \quad \text{enantiomer}
\end{align*}
\]
Consider the mechanism for the epoxidation of (2S,3S)-3-bromo-4,4-dimethyl-2-pentanol with hydroxide and its subsequent reaction with cyanide to form (2S,3S)-3-hydroxy-2,4,4-trimethylpentanenitrile.

**Step 1:**

**Step 2:**

Consider the mechanism for the epoxidation of (1R,2R)-2-chlorophenylpropanol with hydroxide and its subsequent reaction with hydrobromic acid to form (1R,2R)-2-bromophenylpropanol.

**Step 1:**

**Step 2:**
Consider the mechanism for the epoxidation of \((1S,2S)-2\text{-bromo-1-methyl-1-cyclopentanol}\) and its subsequent reaction with methoxide to form \((1S,2S)-2\text{-methoxy-1-methyl-1-cyclopentanol}\).

**Step 1:**

\[ \text{Br} \quad \Theta \quad \text{OH} \quad \rightarrow \quad \text{Br} \quad \Theta \quad \rightarrow \quad \text{O} \quad \Theta \]

**Step 2:**

\[ \Theta \quad \text{OCH}_3 \quad \rightarrow \quad \Theta \quad \text{OH} \quad \rightarrow \quad \text{OH} \quad \Theta \]

\[ + \text{enantiomer} \quad + \text{enantiomer} \]

Consider the mechanism for the epoxidation of \((2R,3R)-3\text{-chloro-2-butanol}\) and its subsequent reaction with methylmagnesium bromide to form \((R)-3\text{-methyl-2-butanol}\).

**Step 1:**

\[ \text{Cl} \quad \Theta \quad \text{OH} \quad \rightarrow \quad \text{Cl} \quad \Theta \quad \rightarrow \quad \text{O} \quad \Theta \]

**Step 2:**

\[ \text{Mg} \quad \Theta \quad \text{Br} \quad \rightarrow \quad \text{Mg} \quad \Theta \quad \rightarrow \quad \text{OH} \quad \Theta \]

\[ \text{CH}_3 \quad \text{OH} \]
Consider the mechanism for the epoxidation of \((1R,2R)-2\text{-chloro-1,2-diphenyl-1-ethanol}\) and its subsequent reaction with lithium aluminum hydride to form \((S)-1,2\text{-diphenyl-1-ethanol}\).

**Step 1:**

![Chemical structure diagram]

**Step 2:**

![Chemical structure diagram]

**B.2.5 Aromatic ring**

**B.2.5.1 Acylation**

Consider the mechanism for the Friedel-Crafts acylation of benzene to form acetophenone.

![Chemical structure diagram]
B.2.5.2 Addition–elimination

Consider the mechanism for the addition–elimination reaction of chlorobenzene with amide to form aniline.

\[
\begin{align*}
\text{Cl:} & \quad \text{H} \quad \text{N-H} \\
\text{Cl} & \quad \text{H} \quad \text{N-H} \\
\text{Cl} & \quad \text{H} \quad \text{N-H} \\
\end{align*}
\]

B.2.5.3 Alkylolation

Consider the mechanism for the Friedel-Crafts alkylation of benzene to form isopropylbenzene.

\[
\begin{align*}
\text{Cl} & \quad \text{Al-Cl} \\
\text{Cl} & \quad \text{Al-Cl} \\
\text{Cl} & \quad \text{Al-Cl} \\
\text{Cl} & \quad \text{Al-Cl} \\
\text{CH}_3 & \quad \text{Cl} \\
\text{CH}_3 & \quad \text{Cl} \\
\text{CH}_3 & \quad \text{Cl} \\
\text{H} & \quad \text{Cl} \\
\text{CH}_3 & \quad \text{Cl} \\
\end{align*}
\]
B.2.5.4 Azo coupling

Consider the mechanism for the azo coupling reaction between 4-nitrobenzenediazonium and resorcinol to form azo violet.

B.2.5.5 Electrophilic aromatic substitution

Consider the mechanism for the electrophilic aromatic bromination of benzene to form bromobenzene.
Consider the mechanism for the electrophilic aromatic chlorination of benzene to form chlorobenzene.

\[
\begin{align*}
\text{Cl} & \quad \text{Cl} & \quad \text{AlCl}_3 \\
\rightarrow & \quad \text{Cl} & \quad \text{Cl} & \quad \text{AlCl}_3 \\
\text{H} & \quad \text{Cl} & \quad \text{Cl} & \quad \text{AlCl}_3 \\
\rightarrow & \quad \text{Cl} & \quad \text{Cl} & \quad \text{AlCl}_3 \\
\rightarrow & \quad \text{HCl} & \quad \text{AlCl}_3 \\
\end{align*}
\]

Consider the mechanism for the electrophilic aromatic nitration of benzene to form nitrobenzene.

\[
\begin{align*}
\text{H} & \quad \text{O} & \quad \text{O} & \quad \text{H} \\
\rightarrow & \quad \text{H} & \quad \text{O} & \quad \text{O} & \quad \text{H} \\
\rightarrow & \quad \text{H} & \quad \text{O} & \quad \text{O} & \quad \text{H} \\
\rightarrow & \quad \text{H} & \quad \text{O} & \quad \text{O} & \quad \text{H} \\
\end{align*}
\]

\[
\begin{align*}
\text{H} & \quad \text{O} & \quad \text{O} & \quad \text{H} \\
\rightarrow & \quad \text{H} & \quad \text{O} & \quad \text{O} & \quad \text{H} \\
\rightarrow & \quad \text{H} & \quad \text{O} & \quad \text{O} & \quad \text{H} \\
\rightarrow & \quad \text{H} & \quad \text{O} & \quad \text{O} & \quad \text{H} \\
\end{align*}
\]
Consider the mechanism for the electrophilic aromatic sulfonation of benzene to form benzenesulfonic acid.

\[
\begin{align*}
\text{Benzene} & \xrightarrow{\text{HSO}_3^-} \text{Benzenesulfonic acid} \\
\text{H}_2\text{SO}_3 & \xrightarrow{\text{H}_2\text{O}^-} \text{Benzenesulfonic acid}
\end{align*}
\]

Consider the mechanism for electrophilic aromatic bromination of pyridine to form 3-bromopyridine.

\[
\begin{align*}
\text{Pyridine} & \xrightarrow{\text{Br}^-} \text{3-Bromopyridine} \\
\text{Pyridine} & \xrightarrow{\text{Br}^-} \text{3-Bromopyridine}
\end{align*}
\]

Consider the mechanism for the electrophilic aromatic sulfonation of anisole with hydroxide to form 4-methoxybenzenesulfonic acid.

\[
\begin{align*}
\text{Anisole} & \xrightarrow{\text{H}_2\text{SO}_3^-} \text{4-Methoxybenzenesulfonic acid} \\
\text{H}_2\text{SO}_3 & \xrightarrow{\text{H}_2\text{O}^-} \text{4-Methoxybenzenesulfonic acid}
\end{align*}
\]
B.2.5.6 Nucleophilic aromatic substitution

Consider the mechanism for the nucleophilic aromatic substitution of 1-bromo-4-nitrobenzene with hydroxide to form 4-nitrophenol.

Consider the mechanism for the nucleophilic aromatic substitution of 1-chloro-2-nitrobenzene with hydroxide to form 2-nitrophenol.
B.2.6 Carbonyl

B.2.6.1 Acyl substitution

Consider the mechanism for the acid-catalyzed nucleophilic acyl substitution of benzoic 4-methylbenzenesulfonic anhydride with ethanol to form ethyl benzoate.

Consider the mechanism for the base-catalyzed nucleophilic acyl substitution of benzoyl chloride with methoxide to form methyl benzoate.
Consider the mechanism for the nucleophilic acyl substitution of benzoyl chloride with ammonia to form benzamide.

\[
\begin{align*}
\text{Benzoic acid} & \quad \text{Cl}^- & \quad \text{NH}_3 \\
\text{Benzamide} & \quad \text{NH}_2^- & \quad \text{NH}_3 \\
\text{Ammonium chloride} & \quad \text{Cl}^- & \quad \text{NH}_4^+
\end{align*}
\]

B.2.6.2 Aldol addition/substitution

Consider the mechanism for the aldol addition of propanal to itself to form 3-hydroxy-2-methylpentanal.

\[
\begin{align*}
\text{Propanal} & \quad \text{OH}^- & \quad \text{OH} \\
\text{Aldehyde} & \quad \text{OH}^- & \quad \text{OH} \\
\text{Water} & \quad \text{OH}^- & \quad \text{OH}
\end{align*}
\]
Consider the mechanism for the aldol condensation of 2-butanone to form (Z)-3,4-dimethyl-3-hexen-2-one.

B.2.6.3 Alpha-halogenation

Consider the mechanism for the acid-catalyzed bromination of cyclopentanone to form 2-bromocyclopentanone.
B.2.6.4 Condensation of an ester/diester

Consider the mechanism for the Claisen condensation of methyl propanoate to form methyl 2-methyl-3-oxopentanoate.

**Step 1:**

**Step 2:**

H\textsubscript{3}CO

\[ \begin{align*}
\text{H}_3\text{CO} & \text{O} \rightarrow \text{H}_3\text{CO} \text{O} \left( \text{H}_3\text{CO} \right) \text{O} \left( \text{H}_3\text{CO} \right) \\
\text{H}_3\text{CO} & \text{O} \rightarrow \text{H}_3\text{CO} \text{O} \left( \text{H}_3\text{CO} \right) \text{O} \left( \text{H}_3\text{CO} \right) \\
\end{align*} \]
Consider the mechanism for the Dieckmann cyclization of diethyl adipate to form ethyl cyclopentanone-2-carboxylate.

**Step 1:**

**Step 2:**
B.2.6.5 Conjugate addition

Consider the mechanism for the conjugate addition of diethyl malonate with methyl vinyl ketone to form diethyl 2-(3-oxobutyl)malonate.

Step 1:

Step 2:

Step 3:
Consider the mechanism for the conjugate addition of cyclo-2-hexen-1-one with lithium dimethylcuprate and subsequently with 1-iodopropane to form 2-ethyl-3-methylcyclo-1-hexanone.

**Step 1:**

![Step 1 reaction diagram]

**Step 2:**

![Step 2 reaction diagram]

B.2.6.6 Enamine/imine synthesis

Consider the mechanism for the acid-catalyzed nucleophilic addition of phenylamine to benzophenone to form N-(diphenylmethylene)phenylamine.

![Enamine/imine synthesis reaction diagram]
Consider the mechanism for the enamine synthesis between cyclopentanone and pyrrolidine to form 1-(cyclopent-1-en-1-yl)pyrrolidine.

B.2.6.7 Esterification

Consider the mechanism for the acid-catalyzed Fischer esterification of acetic acid with ethanol to form ethyl acetate.
Consider the mechanism for the Grignard reaction between phenylmagnesium bromide and cyclopentanone to form phenylcyclopentanol.

**Step 1:**

Consider the mechanism for the Grignard reaction between ethylmagnesium bromide and cyclopentanecarbaldehyde to form 1-cyclopentyl-1-propanol.

**Step 1:**

**Step 2:**
Consider the mechanism for the Grignard reaction between phenylmagnesium bromide and carbon dioxide to form benzoic acid.

\[
\begin{align*}
\text{Ph} & \quad \text{MgBr} \quad \text{CO}_2 \\
\text{Ph} & \quad \text{MgBr} \quad \text{CO}_2 \\
\text{Ph} & \quad \text{MgBr} \quad \text{CO}_2
\end{align*}
\]

Consider the mechanism for the Grignard reaction between methyl benzoate and methylmagnesium bromide to form 2-phenyl-2-propanol.

\[
\begin{align*}
\text{Ph} & \quad \text{MgBr} \quad \text{O} \quad \text{H}_3\text{C} \\
\text{Ph} & \quad \text{MgBr} \quad \text{O} \quad \text{H}_3\text{C} \\
\text{Ph} & \quad \text{MgBr} \quad \text{O} \quad \text{H}_3\text{C}
\end{align*}
\]

Consider the mechanism for the Grignard reaction between isobutyraldehyde and methylmagnesium bromide to form 3-methyl-2-butanol.

**Step 1:**

\[
\begin{align*}
\text{CH}_3 & \quad \text{MgBr} \\
\text{CH}_3 & \quad \text{MgBr} \\
\text{CH}_3 & \quad \text{MgBr}
\end{align*}
\]

**Step 2:**

\[
\begin{align*}
\text{H}_3\text{C} & \quad \text{OH} \\
\text{H}_3\text{C} & \quad \text{OH} \\
\text{H}_3\text{C} & \quad \text{OH}
\end{align*}
\]
Consider the mechanism for the Grignard reaction between 2-pentanone and ethylmagnesium bromide to form 3-methyl-3-hexanol.

**Step 1:**

**Step 2:**

B.2.6.9 Hemiacetal/acetal formation

Consider the mechanism for the hemiacetal formation reaction between acetone, hydronium ion, and methanol to form 2-methoxypropan-2-ol.
Consider the mechanism for the acid-catalyzed nucleophilic addition of phenol to benzophenone to form diphenoxypiphenylmethane.

Consider the mechanism for the acid-catalyzed hemiacetal formation from acetone and ethanol to form 2-ethoxy-2-propanol.
Consider the mechanism for the base-catalyzed hemiacetal formation from acetone and ethoxide to form 2-ethoxy-2-propanol.

B.2.6.10 Hydration/dehydration

Consider the mechanism for the base-catalyzed hydration of cyclopropanecarbaldehyde to form cyclopropylmethanediol.

Consider the mechanism for the dehydration of propionamide with thionyl chloride to form propionitrile.
B.2.6.11 Hydrolysis

Consider the mechanism for the base-catalyzed hydrolysis of acetamide to give acetate and ammonia.

\[
\text{H}_3\text{C} \begin{array}{c} \text{O} \\ \text{NH}_2 \end{array} \rightleftharpoons \text{OH} \rightarrow \text{H}_3\text{C} \begin{array}{c} \text{O} \\ \text{NH}_2 \end{array} + \text{enantiomer}
\]

B.2.6.12 Reduction

Consider the mechanism for the reduction of \(\alpha\)-methyl-\(\gamma\)-butyrolactone with lithium aluminum hydride to form (S)-2-methylbutane-1,4-diol.

**Step 1:**

\[
\text{H}_3\text{C} \begin{array}{c} \text{O} \\ \text{H} \end{array} \rightleftharpoons \text{H}_3\text{C} \begin{array}{c} \text{O} \\ \text{H} \end{array} \text{Li}^+ \rightarrow \text{H}_3\text{C} \begin{array}{c} \text{O} \\ \text{H} \end{array} \text{Li}^+ 
\]

**Step 2:**

\[
\Theta \rightarrow \text{[H}^+] \rightarrow \text{OH} 
\]

334
Consider the mechanism for the reduction of ethyl 3-oxobutanoate with sodium borohydride to form ethyl 3-hydroxybutanoate.

**Step 1:**

**Step 2:**
Consider the mechanism for the reduction of 2-butane with lithium aluminum hydride to form 2-butanol.

**Step 1:**

\[
\text{O} \quad \text{Li}^+ \quad \text{Al} \quad 
\begin{array}{c}
\text{H} \\
\text{H} \\
\end{array} 
\quad \rightarrow 
\begin{array}{c}
\text{H} \\
\text{H} \\
\end{array} + 
\begin{array}{c}
\text{H} \\
\text{H} \\
\end{array}
\]

**Step 2:**

\[
\text{O} \quad \text{H} \\
\text{H} \\
\quad \text{O} \quad \text{H} \\
\text{H} \\
\quad \text{O} \quad \text{H} \\
\text{H} \\
\quad \text{O} \quad \text{H} \\
\text{H} \\
\]

Consider the mechanism for the reduction of 2-methyl-3-pentanone with sodium borohydride to form 2-methyl-3-pentanol.

**Step 1:**

\[
\text{H}_3\text{CCH}_2 \quad \text{B} \quad \text{Na}^+ 
\quad \rightarrow 
\begin{array}{c}
\text{H}_3\text{CCH}_2 \\
\text{H} \\
\end{array} + 
\begin{array}{c}
\text{H} \\
\text{H} \\
\end{array}
\]

**Step 2:**

\[
\text{H}_3\text{CCH}_2 \quad \text{O} \quad \text{H} \\
\text{H} \\
\quad \text{O} \quad \text{H} \\
\text{H} \\
\quad \text{O} \quad \text{H} \\
\text{H} \\
\quad \text{O} \quad \text{H} \\
\text{H} \\
\]

336
Consider the mechanism for the reduction of 2-azidopropane with lithium aluminum hydride to form isopropylamine.

**Step 1:**

\[
\begin{align*}
\text{H} & \quad \text{H} \\
\text{N} & \quad \text{N} \\
\text{N} & \quad \text{N} \\
\text{Al} & \quad \text{Li} \\
\end{align*}
\]

**Step 2:**

\[
\begin{align*}
\text{H} & \quad \text{H} \\
\text{N} & \quad \text{N} \\
\end{align*}
\]

Consider the mechanism for the reduction of acetophenone with lithium aluminum hydride to form 1-phenyl-1-ethanol.

**Step 1:**

\[
\begin{align*}
\text{Ph} & \quad \text{CH}_3 \\
\text{C} & \quad \text{O} \\
\text{Al} & \quad \text{Li} \\
\end{align*}
\]

**Step 2:**

\[
\begin{align*}
\text{Ph} & \quad \text{CH}_3 \\
\text{H} & \quad \text{OH} \\
\end{align*}
\]
Consider the mechanism for the reduction of butyraldehyde with sodium borohydride to form 1-butanol.

**Step 1:**

\[
\text{CH}_3\text{CHCH}_2\text{CHO} + \text{NaBH}_4 \rightarrow \text{CH}_3\text{CHCH}_2\text{CH}_2\text{OH}
\]

**Step 2:**

\[
\text{CH}_3\text{CHCH}_2\text{OH} + \text{H}_2\text{O} \rightarrow \text{CH}_3\text{CH}_2\text{OH}
\]

Consider the mechanism for the reduction of acetaldehyde with lithium aluminum hydride to form ethanol.

**Step 1:**

\[
\text{CH}_3\text{CHO} + \text{LiAlH}_4 \rightarrow \text{CH}_3\text{CH}_2\text{OH}
\]

**Step 2:**

\[
\text{CH}_3\text{CH}_2\text{OH} + \text{H}_2\text{O} \rightarrow \text{CH}_3\text{CH}_2\text{OH}
\]

Consider the mechanism for the reduction of acetone with sodium borohydride to form isopropanol.

**Step 1:**

\[
\text{CH}_3\text{C}(:\text{O})\text{CH}_3 + \text{NaBH}_4 \rightarrow \text{CH}_3\text{CH}_2\text{CH}_2\text{OH}
\]

**Step 2:**

\[
\text{CH}_3\text{CH}_2\text{CH}_2\text{OH} + \text{H}_2\text{O} \rightarrow \text{CH}_3\text{CH}_2\text{OH}
\]
B.2.6.13 Saponification

Consider the mechanism for the saponification of ethyl acetate with hydroxide to form acetic acid.

**Step 1:**

```
\begin{align*}
\text{Et} & \quad \text{OH} \\
\text{O} & \quad \text{OEt} \\
\text{O} & \quad \text{Et} \\
\end{align*}
```

**Step 2:**

```
\begin{align*}
\text{H} & \quad \text{O} \\
\text{O} & \quad \text{H} \\
\text{O} & \quad \text{H} \\
\end{align*}
```

\[ \rightarrow \text{Et} \]

B.2.7 Conjugated diene

B.2.7.1 Electrophilic addition

Consider the mechanism for the addition of a hydrogen halide to a conjugated alkene reaction between 1,3-butadiene and hydrobromic acid to form 3-bromo-1-butene and 1-bromo-2-butene.

```
\begin{align*}
\text{H} & \quad \text{Br} \\
\text{H} & \quad \text{Br} \\
\end{align*}
```

```
\begin{align*}
\text{H} & \quad \text{Br} \\
\text{H} & \quad \text{Br} \\
\end{align*}
```

\[ \rightarrow \text{Br} \]

\[ \rightarrow \text{Br} \]
Consider the mechanism for the addition of a hydrogen halide to an alkene reaction between 1-methyl-1,3-cyclohexene and hydrobromic acid to form 3-bromo-1-methyl-1-cyclohexene and 3-bromo-3-methyl-1-cyclohexene.
Appendix C
Supporting Information for Chapter 5

C.1 Survey Items

<table>
<thead>
<tr>
<th>Factor</th>
<th>Item Text</th>
<th>Item Choices and Coding*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LECTURE</td>
<td>During a typical week, what proportion of time during regular class meetings (i.e., lecture sections) do students spend doing the following?</td>
<td>A. Working individually (Dropdown: 0-100, intervals of 5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B. Working in small groups (Dropdown: 0-100, intervals of 5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C. Participating in whole class discussion (Dropdown: 0-100, intervals of 5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D. Listening to the instructor lecture or solve problems (Dropdown: 0-100, intervals of 5) (LECTURE)</td>
</tr>
<tr>
<td><strong>Department characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discipline</td>
<td>[Embedded data based on stratified sampling strategy]</td>
<td>Chemistry (CHEM), Mathematics (MATH), Physics (PHYS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[MATH is reference]</td>
</tr>
<tr>
<td>Highest degree awarded</td>
<td>[Embedded data based on stratified sampling strategy]</td>
<td>ASSOC, BACH, GRAD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[ASSOC is reference]</td>
</tr>
<tr>
<td><strong>Department appointment expectations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOAD</td>
<td>What is your typical teaching load (i.e., how many course sections do you teach) during a single term?</td>
<td>{Dropdown: 1} (0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>{Dropdown: 2} (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>{Dropdown: 3} (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>{Dropdown: 4} (3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>{Dropdown: 5+} (4)</td>
</tr>
<tr>
<td>Tenure status</td>
<td>What is your tenure status at this institution?</td>
<td>Tenured (TENURED); On tenure track, but not tenured (TENURETRACK); Not on tenure track, but this institution has a tenure system (NOTTENURETRACK); No tenure system at this institution (NOTTENURETRACK)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[NOTTENURETRACK is reference]</td>
</tr>
<tr>
<td>SET</td>
<td>What is the role of student evaluations of teaching (SET) in evaluating teaching performance in decisions of review, promotion, or tenure?</td>
<td>SET is the only measure used to evaluate teaching performance (1); SET are used and given more weight as compared to other measures (2); SET are used and given equal weight as compared to other measures (3); SET are used and given less weight compared to other measures (4); SET are not used to evaluate teaching performance (5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Reverse coded with ‘SET are not used to evaluate teaching performance’ as reference (0) and increasing count of one thereafter for item choices]</td>
</tr>
<tr>
<td>APT</td>
<td>How much does the overall assessment of teaching performance matter in decision of review, promotion, or tenure for someone in your role?</td>
<td>It is not considered (1); Somewhat influential (2); Influential (3); Very influential (4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>['It is not considered' coded as reference (0) with increasing count of one thereafter for item choices]</td>
</tr>
<tr>
<td><strong>Classroom contextual</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td><strong>SIZE</strong></td>
<td>What was the approximate enrollment in a typical lecture section?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>{Text entry}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Numerical responses were binned into six size categories: 2-19, 20-29, 30-39, 40-59, 60-99, 100+ and grand-mean-centered at the 30-39 bin as reference (0) For responses given as ranges, the average of the range was taken.]</td>
<td></td>
</tr>
<tr>
<td><strong>ROOM</strong></td>
<td>Which of the following best describes the set-up in your classroom?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Classroom with fixed seats (0); Classroom that accommodates group work (1); Other {Text entry}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[All written responses were adjudicated as either 0 or 1]</td>
<td></td>
</tr>
<tr>
<td><strong>DECISION</strong></td>
<td>What are the primary decision makers for the following?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Instructional methods you use</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Myself (0); myself and others (1); One or more other people (1); Does not apply (0)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Personal factors</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RBIS</strong></td>
<td>Have you ever been a student in a course taught using RBIS?</td>
</tr>
<tr>
<td></td>
<td>Yes (1), no (0), I don’t know (0)</td>
</tr>
<tr>
<td><strong>SOTL</strong></td>
<td>Do you conduct STEM education research and/or participate in the scholarship of teaching and learning?</td>
</tr>
<tr>
<td></td>
<td>Yes (1), no (0), I don’t know (0)</td>
</tr>
<tr>
<td><strong>TFC</strong></td>
<td>How many academic courses focused on learning how to teach have you taken at the undergraduate, graduate, and postdoctoral levels?</td>
</tr>
<tr>
<td></td>
<td>{Dropdown: 0} (0)</td>
</tr>
<tr>
<td></td>
<td>{Dropdown: 1, 2, 3, 4+} (1)</td>
</tr>
<tr>
<td><strong>WKSP</strong></td>
<td>Have you ever participated in any of the following types of teaching-related professional development?</td>
</tr>
<tr>
<td></td>
<td>Half-day workshop(s)</td>
</tr>
<tr>
<td></td>
<td>Full-day or longer workshop(s)</td>
</tr>
<tr>
<td></td>
<td>Attending a teaching-focused conference</td>
</tr>
<tr>
<td></td>
<td>No, yes</td>
</tr>
<tr>
<td></td>
<td>[Coded as (1) if ‘yes’ to any of the three items, or (0) if ‘no’ to all three items]</td>
</tr>
<tr>
<td><strong>NFE</strong></td>
<td>Have you ever participated in any of the following types of teaching-related professional development?</td>
</tr>
<tr>
<td></td>
<td>New faculty experience at my institution</td>
</tr>
<tr>
<td></td>
<td>New faculty workshop external to my institution</td>
</tr>
<tr>
<td></td>
<td>No, yes</td>
</tr>
<tr>
<td></td>
<td>[Coded as (1) if ‘yes’ to either item, or (0) if ‘no’ to both items]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Teacher thinking</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GROWTH</strong></td>
<td>Average of three items on a six-point Likert scale from 1 to 6 that describe fixed mindset (Dweck et al., 1995)</td>
</tr>
<tr>
<td></td>
<td>Strongly disagree (1); Disagree (2); Slightly disagree (3); Slightly agree (4); Agree (5); Strongly agree (6)</td>
</tr>
<tr>
<td></td>
<td>[Items were reverse coded, centered at the middle of the scale, and the average value was used]</td>
</tr>
<tr>
<td><strong>SATISFACTION</strong></td>
<td>How satisfied are you with your students’ learning in your course?</td>
</tr>
<tr>
<td></td>
<td>Very dissatisfied (1); Dissatisfied (2); Neither dissatisfied or satisfied (3); Satisfied (4); Very satisfied (5)</td>
</tr>
<tr>
<td></td>
<td>[Values were centered at the middle of the scale]</td>
</tr>
</tbody>
</table>

*Curly brackets indicate item dropdown or text entry. Parentheses with non-bolded values indicate non-numerical values assigned to item choices. Parentheses with bolded values indicate item codes. Brackets indicate item coding information.*
References


C.2 Descriptive statistics for level 1 variables (instructors; \( n = 2,382 \))

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent lecturing</td>
<td>56.17</td>
<td>24.95</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Chemistry</td>
<td>0.33</td>
<td>0.47</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Physics</td>
<td>0.34</td>
<td>0.47</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Bachelor program</td>
<td>0.34</td>
<td>0.47</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Graduate program</td>
<td>0.40</td>
<td>0.49</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Class size</td>
<td>0.44</td>
<td>1.62</td>
<td>-2</td>
<td>3</td>
</tr>
<tr>
<td>Classroom setup</td>
<td>0.52</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Decision making</td>
<td>0.13</td>
<td>0.34</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Teaching load</td>
<td>1.88</td>
<td>1.17</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Tenured faculty</td>
<td>0.54</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Tenure-track faculty</td>
<td>0.17</td>
<td>0.37</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Student evaluation of teaching</td>
<td>2.26</td>
<td>0.99</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Assessment of teaching performance</td>
<td>2.03</td>
<td>0.89</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>RBIS use as a student</td>
<td>0.15</td>
<td>0.35</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Scholarship of teaching and learning</td>
<td>0.39</td>
<td>0.49</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Teaching-focused coursework</td>
<td>0.47</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Teaching-related workshops</td>
<td>0.91</td>
<td>0.29</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Teaching-related new faculty experiences</td>
<td>0.21</td>
<td>0.41</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Growth mindset</td>
<td>1.23</td>
<td>1.13</td>
<td>-2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Satisfaction with student learning</td>
<td>0.65</td>
<td>0.83</td>
<td>-2</td>
<td>2</td>
</tr>
</tbody>
</table>

C.3 Descriptive statistics for level 2 variables (department; \( n = 1,405 \))

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry</td>
<td>0.33</td>
<td>0.47</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Physics</td>
<td>0.32</td>
<td>0.47</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Bachelor program</td>
<td>0.35</td>
<td>0.48</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Graduate program</td>
<td>0.33</td>
<td>0.47</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
### C.4 Correlations among the variables at level 1 (instructors; \( n = 2,382 \))

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Percent lecturing</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Class size</td>
<td>0.18***</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Classroom setup</td>
<td>-0.30***</td>
<td>-0.49***</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Decision making</td>
<td>-0.06***</td>
<td>0.12***</td>
<td>-0.04</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Teaching load</td>
<td>-0.07***</td>
<td>-0.35***</td>
<td>0.18***</td>
<td>-0.09***</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Tenured faculty</td>
<td>0.03</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>-0.08***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Tenure-track faculty</td>
<td>-0.05</td>
<td>0.00</td>
<td>-0.03</td>
<td>0.00</td>
<td>-0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Student evaluation of teaching</td>
<td>0.04*</td>
<td>0.15***</td>
<td>-0.13***</td>
<td>0.00</td>
<td>-0.14***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Assessment of teaching performance</td>
<td>-0.07***</td>
<td>-0.09***</td>
<td>0.10***</td>
<td>-0.02</td>
<td>0.12***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. RBIS use as a student</td>
<td>-0.14***</td>
<td>-0.03</td>
<td>0.07***</td>
<td>-0.03</td>
<td>0.07***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Scholarship of teaching and learning</td>
<td>-0.22***</td>
<td>0.07***</td>
<td>0.03</td>
<td>0.05</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Teaching-focused coursework</td>
<td>-0.15***</td>
<td>-0.11***</td>
<td>0.08***</td>
<td>0.00</td>
<td>0.16***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Teaching-related workshops</td>
<td>-0.19***</td>
<td>-0.09***</td>
<td>0.09***</td>
<td>0.00</td>
<td>0.10***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Teaching-related new faculty experiences</td>
<td>-0.16***</td>
<td>-0.06***</td>
<td>0.05</td>
<td>0.02</td>
<td>-0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Growth mindset</td>
<td>-0.21***</td>
<td>0.01</td>
<td>0.08***</td>
<td>0.00</td>
<td>0.00***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Satisfaction with student learning</td>
<td>-0.14***</td>
<td>-0.14***</td>
<td>0.12***</td>
<td>-0.02</td>
<td>0.00***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>11</td>
<td>12</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>----</td>
<td>-----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Percent lecturing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Class size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Classroom setup</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Decision making</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Teaching load</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Tenured faculty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Tenure-track faculty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Student evaluation of teaching</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Assessment of teaching performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. RBIS use as a student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Scholarship of teaching and learning</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Teaching-focused coursework</td>
<td>0.08***</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Teaching-related workshops</td>
<td>0.15***</td>
<td>0.16***</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Teaching-related new faculty experiences</td>
<td>0.11***</td>
<td>0.07***</td>
<td>0.13***</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Growth mindset</td>
<td>0.12***</td>
<td>0.06**</td>
<td>0.08***</td>
<td>0.06***</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Satisfaction with student learning</td>
<td>-0.02</td>
<td>0.04</td>
<td>0.04</td>
<td>0.02</td>
<td>0.06**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. *p < 0.05, **p < 0.01, ***p < 0.001

**C.5 Correlations among the variables at level 2 (department; n = 1,405)**

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Chemistry</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Physics</td>
<td>-0.49***</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Bachelor program</td>
<td>-0.03</td>
<td>0.08''</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>4. Graduate program</td>
<td>-0.01</td>
<td>-0.02</td>
<td>-0.51***</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note. *p < 0.05, **p < 0.01, ***p < 0.001
### C.6 Variance inflation factors (VIF) on the standardized model

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry</td>
<td>1.63</td>
</tr>
<tr>
<td>Physics</td>
<td>1.51</td>
</tr>
<tr>
<td>Bachelor program</td>
<td>1.81</td>
</tr>
<tr>
<td>Graduate program</td>
<td>2.42</td>
</tr>
<tr>
<td>Class size</td>
<td>2.72</td>
</tr>
<tr>
<td>Classroom setup</td>
<td>1.34</td>
</tr>
<tr>
<td>Decision making</td>
<td>1.02</td>
</tr>
<tr>
<td>Teaching load</td>
<td>1.28</td>
</tr>
<tr>
<td>Tenured faculty</td>
<td>1.50</td>
</tr>
<tr>
<td>Tenure-track faculty</td>
<td>1.45</td>
</tr>
<tr>
<td>Student evaluation of teaching</td>
<td>1.06</td>
</tr>
<tr>
<td>Assessment of teaching performance</td>
<td>1.14</td>
</tr>
<tr>
<td>RBIS use as a student</td>
<td>1.08</td>
</tr>
<tr>
<td>Scholarship of teaching and learning</td>
<td>1.09</td>
</tr>
<tr>
<td>Teaching-focused coursework</td>
<td>1.17</td>
</tr>
<tr>
<td>Teaching-related workshops</td>
<td>1.08</td>
</tr>
<tr>
<td>Teaching-related new faculty experiences</td>
<td>1.09</td>
</tr>
<tr>
<td>Growth mindset</td>
<td>1.05</td>
</tr>
<tr>
<td>Satisfaction with student learning</td>
<td>1.03</td>
</tr>
</tbody>
</table>
Appendix D

Institutional Review Board Approval

7/11/2017

Jeffrey Raker, PhD
CITRUS - Center for the Improvement of Teaching and Research in Undergraduate STEM
Education
4202 East Fowler Avenue
CHE205
Tampa, FL 33620

RE: Not Human Subjects Research Determination
IRB#: Pro00031458
Title: Collaborative Research: Evaluating the Uptake of Research-Based Instructional Strategies in Undergraduate Chemistry, Mathematics, & Physics (NSF Proposal Number 1726126)

Dear Dr. Raker:

The Institutional Review Board (IRB) has reviewed your application. The activities presented in the application involve methods of program evaluation, quality improvement, and/or needs analysis. While potentially informative to others outside of the university community, study results would not appear to contribute to generalizable knowledge. As such, the activities do not meet the definition of research under USF IRB policy, and USF IRB approval and oversight are therefore not required.

While not requiring USF IRB approval and oversight, your study activities should be conducted in a manner that is consistent with the ethical principles of your profession. If the scope of your project changes in the future, please contact the IRB for further guidance.

If you will be obtaining consent to conduct your study activities, please remove any references to "research" and do not include the assigned Protocol Number or USF IRB contact information.

If your study activities involve collection or use of health information, please note that there may be requirements under the HIPAA Privacy Rule that apply. For further information, please contact a HIPAA Program administrator at (813) 974-5638.

Sincerely,

[Signature]

John A. Schumaker, Ph.D.
Appendix E
Publisher Permissions Documentation

E.1 Chapter 3 (Chemistry Education Research and Practice)

Development of a machine learning-based tool to evaluate correct Lewis acid-base model use in written responses to open-ended formative assessment items

B. J. Yik, A. J. Dood, D. Cruz-Ramírez de Arellano, K. B. Fields and J. R. Raker, 

To request permission to reproduce material from this article, please go to the Copyright Clearance Center request page.

If you are an author contributing to an RSC publication, you do not need to request permission provided correct acknowledgement is given.

If you are the author of this article, you do not need to request permission to reproduce figures and diagrams provided correct acknowledgement is given. If you want to reproduce the whole article in a third-party publication (excluding your thesis/dissertation for which permission is not required) please go to the Copyright Clearance Center request page.

Read more about how to correctly acknowledge RSC content.
E.2 Chapter 4 (Chemistry Education Research and Practice)

Generalized rubric for level of explanation sophistication for nucleophiles in organic chemistry reaction mechanisms


To request permission to reproduce material from this article, please go to the Copyright Clearance Center request page.

If you are an author contributing to an RSC publication, you do not need to request permission provided correct acknowledgement is given.

If you are the author of this article, you do not need to request permission to reproduce figures and diagrams provided correct acknowledgement is given. If you want to reproduce the whole article in a third-party publication (excluding your thesis/dissertation for which permission is not required) please go to the Copyright Clearance Center request page.

Read more about how to correctly acknowledge RSC content.
Evaluating the impact of malleable factors on percent time lecturing in gateway chemistry, mathematics, and physics courses

Author: Brandon J. Yik et al
Publication: International Journal of STEM Education
Publisher: Springer Nature
Date: Feb 10, 2022

Creative Commons
This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.
To request permissions for a type of use not listed, please contact Springer Nature

© 2022 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Data Security and Privacy
| For California Residents | Terms and Conditions/Comments? We would like to hear from you. E-mail us at customercare@copyright.com