Beyond the Surface: Understanding STEM Faculty Engagement with Pedagogical Innovations.

Haleigh Machost

B.S. Chemistry, Emory University, 2019

A Dissertation presented to the
Graduate Faculty of the University of Virginia Department of Chemistry
in Candidacy for the degree of Doctor of Philosophy

January 2025

Committee members:

Dr. Marilyne Stains (Advisor)

Dr. Dean Harman (Chair)

Dr. Kateri DuBay

Dr. Michelle Personick

Dr. Lindsay Wheeler

Abstract

This dissertation investigates faculty practices in higher education, with a particular focus on the adoption of pedagogical innovations by Science, Technology, Engineering and Mathematics (STEM) instructors. Central to this work is an exploration of the cognitive, emotional, and motivational factors that guide faculty decision-making processes in relation to their pedagogical practices. Drawing on the theoretical underpinnings of the Teacher-Centered Systemic Reform (TCSR) model, as well as the Diffusion of Innovations (DOI) framework, this dissertation provides insights into how instructors respond to challenges in their teaching environments, how their emotions influence their pedagogical choices, and how innovations like alternative grading schemes are perceived and implemented in the classroom.

Part 1 of this dissertation concerns a project centered on a specific practice – reflection. Within this first part, the part one introduction presents the construct of reflective practice in higher education; chapter 1 subsequently presents an analysis of reflective writings by STEM faculty, a relatively underexplored area in the literature. The study reveals that novice instructors need guidance in pedagogical innovations, like reflective practice, in order for said innovations to be effective. The results have already informed changes to a national new-faculty workshop, demonstrating the real-world impact of this research on instructor development. Chapter 2 extends the findings of chapter 1 by focusing on a particularly neglected area of STEM faculty thinking – emotions in teaching. This chapter addresses this gap by examining the emotional responses STEM instructors express in their reflective writings. The analysis reveals that emotions such as anger, guilt, anxiety, and happiness are large parts of instructors' experiences despite being rarely discussed in formal pedagogical discourse. The research highlights the importance of

acknowledging these emotions, both for instructors' self-reflection and for informing effective teacher training programs.

Part 2 of this dissertation is centered on the implementation of an alternative grading scheme. The part 2 introduction explores the novel grading scheme of specifications grading. This is followed by chapter 3 wherein instructors' motivations for adopting the innovative teaching practice of specifications grading are examined. Through the analysis of semi-structured interviews, this chapter explores faculty perceptions of the relative advantages of specifications grading compared to traditional grading methods. The results reveal that many instructors view specifications grading as a way to provide more flexibility for their students. Chapter 4 delves into the characteristics of the implementation of specifications grading across a range of chemistry course types through a thorough examination of course artifacts. The study reveals the implementation of specifications grading is quite varied, necessitating careful consideration when attributing outcomes to the general umbrella of specifications grading. While much of the literature focuses on instructors' accounts of student benefits under specifications grading, chapter 5 investigates how students perceive the impact of specifications grading on their own learning experiences. The findings suggest that students have mixed opinions, with some praising the clarity of the grading system and its focus on learning outcomes, while others express concerns about the increased effort required to meet the criteria. This chapter calls for further research to explore the differential impacts of innovative grading systems on diverse student populations and highlights the importance of evaluating pedagogical innovations over time, to better align research with the stages of innovation adoption.

This dissertation makes several key contributions to the literature on teaching and learning in higher education. By examining the emotional, cognitive, and motivational dimensions of

faculty decision-making, as well as the practical and student-centered outcomes of pedagogical innovations like specifications grading, the research furthers our understanding of the factors that influence the adoption and effectiveness of pedagogical innovations in STEM disciplines. The findings have broad implications for faculty development programs, the design of grading systems, and the future of pedagogical practices in higher education, offering valuable insights for educators, administrators, and researchers. Finally, the common themes across both parts of this dissertation are explored, which provide beneficial information in furtherance of STEM instructor processional development.

Table of Contents

Abstract ***2***

Table of Contents ***5***

List of Figures ***10***

List of Tables ***11***

Acknowledgements ***15***

A glimpse into the current landscape of STEM higher education and an overview of this dissertation ***17***

Importance and concerns of STEM higher education **17**

The response of education researchers **19**

Evidence-based instructional and assessment practices 20

Alternative grading 22

Reflective practices 24

Current knowledge of STEM instructors' practices **26**

Adoption of evidence-based instructional and assessment practices 26

Adoption of alternative grading 28

Adoption of reflective practices 30

The importance of and frameworks for studying instructors **31**

Teacher-Centered Systemic Reform model 32

Diffusion of Innovations model 33

Overview of Dissertation **37**

Part 1 Introduction: An overview of reflection in higher education ***39***

Reflection in practice **40**

What are the different types of reflection? **43**

Time-dependent 43

Depth of reflections 44

Content of reflections 46

Scaffoldings for Reflection **47**

Aim of Part 1 ..... **49**

Chapter 1. Exploring the Nature of STEM Instructors' Written Reflections ***51***

Introduction **51**

Reflective practice 54

Analytical frameworks for reflections 56

Methods **59**

Reflection scaffold ... 59
Participants ... 60
Scaffold analysis ... 61
Trustworthiness ... 64

Results and Discussion ... 65
Nature of critical incidents ... 65
Content of reflections ... 67
Depth of reflections ... 70
Plans to address similar situations in the future ... 74
Relationship between the nature of the critical incident and depth of reflections ... 76
Relationship between the content and depth of reflections ... 77
Relationship between the plans outlined and depth of reflections ... 79

Implications ... 81

Limitations ... 83

Conclusion ... 83

Chapter 2. Uncovering the complexity of emotions experienced by physics faculty when reflecting on a past teaching experience: Shining a light on an often-overlooked aspect of post-secondary instruction. ... 84

Introduction ... 84
What are emotions? ... 85
Why is it important to characterize instructors' emotions? ... 86
What tools have been used to characterize emotions? ... 90
Research Aim ... 92

Methods ... 93
Participants ... 93
Survey Design ... 95
Survey Analysis ... 95
Qualitative Coding ... 96
SEANCE ... 97
Trustworthiness ... 99

Results and Discussion ... 100
ANEW ... 100
Emotions present ... 103
Reasons for emotions ... 107

Implications ... 111
Implications for professional development ... 112
Implications for education researchers ... 113

Limitations ... 114

Conclusion ... 115

Part 1 Conclusions and Future Directions ... 116

Conclusions ... 116

Future Directions **116**

Part 2 Introduction: A brief overview of specifications grading ***119***

The path to alternative grading **119**

Specifications Grading **121**

Diffusion of specifications grading **124**

Previous studies on specifications grading **125**

Aim of Part 2 ..... **126**

Chapter 3. Benefits and challenges of specifications grading: Perceptions from chemistry instructors who have adopted this grading scheme ***127***

Introduction **127**

Specifications Grading 129

Theoretical Framework 132

Methods **134**

Participants and data collection 134

Interview Protocol 136

Data analysis 137

Trustworthiness 138

Results **139**

RQ 1: What are the benefits of specifications grading perceived by chemistry instructors who adopt it? 139

RQ2: What are the challenges that chemistry instructors anticipated before they implemented specifications grading? 149

Discussion **152**

Implications **159**

Professional development 159

Research agenda 160

Limitations **160**

Conclusion **161**

Chapter 4. An investigation into the implementation of specifications grading in chemistry courses ***162***

Introduction **162**

Methods **169**

Data Collection 169

Data Analysis 170

Results and Discussion **170**

Threshold for meeting specifications 171

Levels of specifications in grading scheme 173

Marks ascribed to levels of specifications met 176

Reattempts and revisions 178
Explicit alignment with learning objectives 181
Final exams 183
Final grade determination 185
Trends within course types 187
Revisions and reattempts 188
Final grade determination 189
Common implementations by course type 190

Limitations 193

Conclusion and Implications 193

Chapter 5: Students' perceptions of specifications grading: development and evaluation of the Perceptions of Grading Schemes (PGS) instrument 196

Introduction 196
Alternative Grading 198
Specifications Grading 199
Study goals 202

Methods 203
Research Context 203
Data collection 204
Data analysis 205

Results and Discussion 208
Development of the Perceptions of Grading Schemes instrument 208

Limitations 224

Implications 225

Conclusions 226

Part 2 Conclusions and Future Directions 227

Conclusions 227

Future Directions 227

Overarching Themes of Dissertation 230

Appendix A. Part 1 IRB and instrument documents 233

Appendix A.1: Distributed Online Survey 233

Appendix A.2: Recruitment Email for Collection of Written Reflections 240

Appendix A.3: Informed Consent Agreement for Collection of Written Reflections 241

Appendix B. Supplementary information for chapter 1 244

Appendix B.1: Inter-rater reliability process 244

Appendix B.2: Data analysis 245

Appendix C. Supplemental information for chapter 2 ..... *248*
Appendix C1. Participant Pool **248**
Appendix C2. Data Analysis **249**
SEANCE Analysis: 253
Appendix C3: Additional Results **259**
Appendix D. Part 2 IRB and Instrument documents *260*
Appendix D1. Instructor documents *260*
Appendix D1.1: Pre-Interview Survey **260**
Appendix D1.2: Interview Protocol **266**
Appendix D1.3: Post-Interview Survey **270**
Appendix D1.4: Instructor Recruitment Email for Specification Grading Project **273**
Appendix D1.5: Instructor Informed Consent Agreement for Specifications Grading Study **275**
Appendix D2. Student documents *278*
Appendix D2.1. Focus Group Recruitment **278**
Appendix D2.2. Focus group consent **280**
Appendix D2.3. Focus group protocol **282**
Appendix D2.4. Student recruitment **283**
Appendix D2.5. Student consent **285**
Appendix D2.6. PGS instrument survey **287**
Appendix E. Supplementary information for chapter 3 ..... *296*
Appendix E1. Additional Participant Data **296**
Table **297**
Appendix E2. Data Analysis **298**
Appendix F. Supplementary information for chapter 4 ..... *306*
Appendix G. Supplementary information for chapter 5 ..... *309*
Appendix G1. Additional participant data **309**
Appendix G2. Data Gathering and Analysis **311**
References *321*

List of Figures

Figure 1. TCSR Model. *33*

Figure 2. Depiction of Rogers' DOI theory. *35*

Figure 3. Stages of innovation adoption by a population *37*

Figure 4. Distribution of physics and astronomy instructors' reflections across the different levels of depth of reflection based on Larrivee's (2008a) model. *72*

Figure 5. Overlay of content and depth of instructors' reflective writings. *78*

Figure 6. Combinations of plans among the four levels of reflection *81*

Figure 7. Outline of response cleaning and analytic processes *96*

Figure 8. Abundances of ANEW scores *101*

Figure 9. Correlation graphs between ANEW valence and dominance metrics *103*

Figure 10. Depiction of various bundling schemes in specifications grading *123*

Figure 11. Growth of specifications grading *125*

Figure 12. Instructors' perceived relative advantages of specifications grading *153*

Figure 13. Approaches to revisions and reattempts by course type *188*

Figure 14. Methods of final grade determination by course type *190*

Figure 15. Implementations of specifications grading by course type *193*

Figure 16. Examples of bundling strategy in specifications grading. *200*

Figure 17. Theorized outcomes of specifications grading (Nilson, 2015) *202*

Figure 18 Exemplar summary of PGS instrument modifications *212*

Figure 19. CFA model of Spring 2023 GC2 Lab with standardized factor loadings *218*

Figure 20. Student perceptions in the Spring 2023 GC2 Lab (n = 324). *223*

Figure 21. Student perceptions in the Fall 2023 GC1 Lab (n = 1,031). *223*

Figure C3.1. Graphs showing lack of correlation between ANEW metrics *259*

Figure G2.1. Scree plot of the EFA data *316*

Figure G2.2. Parallel analysis of the EFA data *317*

Figure G2.3. CFA model for the Spring 2023 GC2 Lab withs standardized factor loadings. *318*

List of Tables

Table 1. Content of reflections based on Valli (1997) *57*

Table 2. Depth of reflection based on Larrivee's (2008a) model *59*

Table 3. Topic categories, topic codes, and definitions *63*

Table 4. Inter-rater reliability metrics. *65*

Table 5. Distributions of topics discussed in critical incidents *67*

Table 6. Personalistic content subcodes, definitions, exemplary quotes, and distribution of subcodes within the reflections that contained personalistic content *69*

Table 7. Critical content subcodes, definitions, exemplary quotes, and distribution of subcodes within the reflections that contained critical content *70*

Table 8. Types of plans described in the reflections for managing future similarly challenging situations. *75*

Table 9. Distribution of topics discussed across the four levels of depth. *77*

Table 10. Distribution of plans described by instructors across the four levels. *80*

Table 11. University demographics of participants *94*

Table 12. Participant self-identified demographics *94*

Table 13. SEANCE metrics utilized in analysis *98*

Table 14. Common emotions as detected through qualitative analysis *103*

Table 15. Alignment between GALC and qualitative coding *106*

Table 16. Most common reasons associated with emotions *107*

Table 17. Proposed outcomes of specifications grading *124*

Table 18. Instructors' academic ranks, teaching experiences, terms using specifications grading and demographics. *136*

Table 19: Chemistry instructors' perceived benefits of specifications grading *140*

Table 20: Chemistry instructors' perceived challenges of specifications grading *150*

Table 21. Critical components of specifications grading *165*

Table 22. Methods for determining the threshold of specifications on individual assignments *173*

Table 23. Number of levels in specifications grading schemes *176*

Table 24. Nomenclature for multi-level systems in specifications grading schemes 178

Table 25. Approaches to assignment revisions and reattempts 180

Table 26. Evidenced and explicit alignment between learning objectives and aspects of specifications grading 183

Table 27. Approaches to final exams in specifications graded courses 184

Table 28. General methods of final grade determination 186

Table 29. Methods of final grade determination with a plus/minus system 187

Table 30. Participant demographics 205

Table 31. PGS factor definitions 214

Table 32. EFA pattern loadings 215

Table 33. EFA factor correlations 215

Table 34. Factors and items for the PGS instrument 217

Table 35. McDonald's ω coefficients for factors in the PGS instrument 219

Table 36. Course measurement invariance fit information and model comparisons. 221

Table B1.1. Summary of iterative inter-rater reliability for codebook creation and validation..... 244

Table B2.1. Full codebook for plans described in instructors' reflections 245

Table B2.2: Types of plan developed by instructor based on level of reflection. 247

Table C2.1. Codebook used in secondary coding of emotions expressed by instructors 249

Table C2.2. Codebook used in secondary coding of reasons for emotions as detailed by instructors. 251

Table C2.3. GALC categories 256

Table E1.1. Number of potential participants identified through each method 296

Table E1.2. Carnegie classifications of participants' institutions. 296

Table E1.3. Course characteristics of participants. 297

Table E2.1. Summary of changes to the codebook during the interrater reliability process 298

Table E2.2. Codebook describing perceived advantages of specifications grading that lead to their implementation 299

Table E2.3. Codebook describing instructors' dissatisfaction with traditional grading *302*

Table E2.4. Codebook describing perceived challenges of specifications grading.... *304*

Table F.1. Detailed levels of specifications *306*

Table F.2. Marks nomenclature of two-level systems..... *306*

Table F.3. Marks nomenclature of multi-level systems *307*

Table F.4. Revision Systems *308*

Table F.5. Approaches to Final Exam *308*

Table F.6. Determination of Final Letter Grade *308*

Table G1.1. Participant demographics *309*

Table G2.1. Pilot version of the Perceptions of Grading Schemes instrument..... *311*

Table G2.2. Descriptive statistics for the pilot version of the Perceptions of Grading Schemes *312*

Instrument..... *312*

Table G2.3. Item correlations for the pilot version of the Perceptions of Grading Schemes *313*

Instrument..... *313*

Table G2.4 Continued. Item correlations for the pilot version of the Perceptions of Grading..... *314*

Schemes instrument..... *314*

Table G2.5. Final version of the Perceptions of Grading Schemes instrument *315*

Table G2.6. CFA fit information for individual, single factor congeneric measurement models for the Spring 2023 GC2 Lab *319*

Table G2.7. CFA fit information for individual, single factor tau equivalent measurement models for the Spring 2023 GC2 Lab *319*

Table G2.8. CFA fit information for individual, single factor congeneric measurement models for the Fall 2023 GC1 Lab *319*

Table G2.9. CFA fit information for individual, single factor tau equivalent measurement models for the Fall 2023 GC1 Lab *319*

Table G2.10. Student perceptions in the Spring 2023 GC2 Lab (n = 324)..... *320*

Table G2.11. Student perceptions in the Fall 2023 GC1 Lab (n = 1,031)..... *320*

Acknowledgements

First and foremost, I would like to express my deepest gratitude to those who have supported the academic side of my journey. I am incredibly fortunate to have had Marilyne as my principal investigator. Her mentorship has been invaluable, and I am eternally grateful that she took a chance on me when I made the decision to transition out of synthetic chemistry. Her guidance, patience, and unwavering belief in me have made this journey not only possible but immensely rewarding.

To all of my lab mates, both present and past, I cannot thank you enough. You have made graduate school more bearable, especially during those moments when imposter syndrome hit hard and when reviewers' comments felt like mountains too steep to climb. Your understanding and encouragement have been a source of strength throughout this process.

I also want to express my sincere gratitude to everyone I worked with during my internship at EHS. The team at EHS was nothing but kind, patient, and supportive as I learned the field, and I am so thankful for the opportunity to grow and learn alongside such a talented group of people. And to be honest, my inability to handle the free time I had with a 9-5 schedule was probably a key factor in ensuring the timely completion of this dissertation.

I would also like to extend a heartfelt thank you to my family for their love, support, and unwavering belief in me. I am not sure how I would have made it through without the safety net I know is always behind me.

To my friends – both those I've known for years before and those I've made along this academic journey – thank you for being my pillars of support. The coffee runs, nature walks, and late-night phone calls have kept me sane, and I truly couldn't have made it without you.

A special thank you goes to my two COVID acquisitions: my cat, Bella, and my husband, Devon. Bella has been my constant companion during late nights of writing, offering me cuddles (and disapproving glares) whenever I needed a break (or it was past her bedtime). No matter how much she might prefer to be with her dad, her presence has been a comfort through the most stressful times.

Devon, words cannot express how grateful I am for you. We have been through so much together, and I feel incredibly lucky that we found each other at the right time. Your love, patience, and constant support have been everything to me. I am so excited to embark on the next chapter of our lives together, and I can't wait to see where the journey takes us.

A glimpse into the current landscape of STEM higher education and an overview of this dissertation

Importance and concerns of STEM higher education

STEM (Science, Technology, Engineering, and Mathematics) higher education is crucial for continued innovation and economic growth in our modern world. By equipping students with advanced knowledge and problem-solving skills, STEM programs prepare students to tackle complex challenges, from developing new technologies, to addressing global issues like climate change and healthcare (National Academies of Sciences, 2021). As industries and economies increasingly rely on technology and data-driven decision-making, STEM graduates are essential for maintaining and advancing societal progress. As such, STEM higher education not only enhances individual career prospects, but also contributes to a more informed society (National Academies of Sciences, 2021). Despite the essential nature of STEM education, there remains persistent deficiencies in American post-secondary STEM programs with an acknowledged need for increased quality of STEM higher education in the United States (National Academies of Sciences, 2021). Indeed, "concerns remain about persistent academic achievement gaps between various demographic groups… and the ability of the U.S. STEM education system to meet domestic demand for STEM labor" (Granovskiy, 2018, p. 1). Changes to the American higher education STEM programs are thus of undeniable importance.

Fundamental to altering the state of STEM higher education is increasing the diversity of STEM graduates (National Academies of Sciences, 2021; Olson & Riordan, 2012). While the proportion of women in STEM occupations has increased since the 1970s, women remain underrepresented with under one-third of STEM professionals identifying as women in 2011 (Granovskiy, 2018). Further highlighting issues of gender representation, the percentage of degrees awarded to women in mathematics and computer science has actually decreased in that same time

frame. In 1966, 33.2% of mathematics and computer science bachelor's degrees were awarded to women; this number reduced to 25.5% as of 2012 (Granovskiy, 2018). Such discrepancies in STEM higher education are also seen among racial and ethnic lines. In 2014, only 12.1% of STEM bachelor's degrees were awarded to Hispanic or Latino/a students (Granovskiy, 2018). More concerningly, the percentage awarded to black/African American students "remained virtually constant at just below 9%" from 2004 to 2014 (Granovskiy, 2018, p. 17). The lack of diversity in STEM has implications both for global competitiveness and for equity and social justice (Granovskiy, 2018; National Academies of Sciences, 2021). The lack of representation of different sub-populations among STEM professionals inherently limits the available talent and varying perspectives which could be contributing to solving societal issues (Granovskiy, 2018). Additionally, lack of representation among STEM professionals can "perpetuate economic gaps that exist in the United States" (Granovskiy, 2018, p. 15) and affect the targets of innovation (National Academies of Sciences, 2021) in addition to perpetuating biases and negatively impacting individual students.

The lack of representation is not due to a lack of interest from historically underrepresented students but rather the inability of the system to retain these students. Indeed, despite some underrepresented students actually being overrepresented when polling for intended majors, they remain underrepresented in terms of degrees awarded (Asai, 2020). A decade ago, fewer than 40% of students who originally enrolled in a STEM degree saw their program through to completion (Olson & Riordan, 2012). Some progress has been made; however, the attrition rates of students in STEM are still only slightly below 50% (Seymour et al., 2019). Additionally, women and other underrepresented groups are more likely to leave a STEM major (Asai, 2020; Seymour et al., 2019; Whitcomb & Singh, 2021). Studies have shown that attrition is most likely to take place during

the first two years of a STEM program (Johri et al., 2017; Ohland et al., 2008; Olson & Riordan, 2012). It is possible that negative interactions and experiences with instructors contribute to students leaving STEM programs (Park et al., 2020). Research has also shown that students' self-efficacy and belief in their own competence in STEM fields affect their retention (Cromley, Perez, & Kaplan, 2016; Hansen, Palakal, & White, 2024) as do their sense of belonging or inclusion (Morris, Hensel, & Dygert, 2019). Another crucial factor connected to student retention is the quality of instruction in introductory STEM courses. A 2019 study revealed that an astonishing 96% of students who left a STEM major cited ineffective teaching as a contributing factor to their decision; this same work also demonstrated that similar sentiments existed within just under 75% of the surveyed students who did persist in their program (Seymour et al., 2019).

Notably, the concern about education quality in STEM programs is not limited to student perceptions; it is supported by education researchers as many studies have shown that students leave introductory STEM courses with a lack of understanding of the discipline's core concepts. The issues contributing to STEM program attrition are thus widely varied. However, efforts have been made to address many of these issues.

The response of education researchers

When attempting to increase the quality of STEM education, Discipline-Based Education Research (DBER) has taken a multifaceted approach. Research has investigated student factors, such as student motivation to learn (Cromley, Perez, & Kaplan, 2016; Hernandez et al., 2013; Kryshko et al., 2022; Simon et al., 2015; Young et al., 2018), attitudes towards learning (Altakhyneh & Abumusa, 2020; Bennett, Braund, & Sharpe, 2014; Unfried et al., 2015; Wu, Deshler, & Fuller, 2018), and self-efficacy (Kryshko et al., 2022; Marshman et al., 2018; Peters, 2013; Rittmayer & Beier, 2008; Syed et al., 2019; Wilson et al., 2015). Additional work has

centered on faculty and their practices as instructional methods directly impact student learning, and faculty beliefs impact student factors such as students' fixed or growth mindset (Canning et al., 2019; Ulug, Ozden, & Eryilmaz, 2011). As such, researchers have developed student-targeted interventions, evidence-based instructional practices (EBIPs), evidence-based assessment practices, alternative grading schemes, and instructor-centered practices, such as reflection. All are ultimately aimed at improving diverse student outcomes.

Evidence-based instructional and assessment practices

There is a shift towards active learning and student-centered practices, and away from the status quo of solely lecture based, or didactic, instruction. Active learning is an umbrella term capturing many practices which all actively engage students in the knowledge creation process (Brame, 2016); this can be accomplished through discussions, problem solving, writing exercises, and other methods which engage students in higher-order thinking. As such, active learning is heavily rooted in the constructivist learning theory, wherein students learn new information through a complex lens of prior knowledge and past experiences (Phillips, 1995). Student-centered pedagogical practices are centered on the ability of students to construct knowledge when able to actively engage with course material (Felder & Brent, 1996; Prince & Felder, 2006). The broad definition of active learning is perhaps best understood by its stark contrast to passive learning, wherein students are the recipients and memorizers of information that is transmitted directly from an instructor (Mahmood, Tariq, & Javed, 2011). As active learning gained traction, specific strategies were developed which incorporate constructivist concepts, encourage higher-order thinking, and have been rigorously tested and shown to improve conceptual understanding and student retention. Examples of such practices include peer instruction (PI) (Crouch & Mazur, 2001), think-pair-share (TPS) (Kothiyal et al., 2013), predict-observe-explain (POE) (James,

Kreager, & LaDue, 2022), peer-led team learning (PLTL) (Snyder et al., 2016), process-oriented guided inquiry learning (POGIL) (Moog & Spencer, 2008), and course-based undergraduate research experiences (CURE) (Corwin, Graham, & Dolan, 2015). Importantly, the incorporation of active learning strategies has been shown to increase learning outcomes among students in STEM courses (Freeman et al., 2014; Schweingruber, Nielsen, & Singer, 2012). Indeed, specific outcomes linked with student-centered active learning include higher scores, reduced DFW rates, and improvement of opportunity gaps among students (Freeman et al., 2014; Theobald et al., 2020; White, Vincent-Layton, & Villarreal, 2021). Due to their continually proven benefits, student-centered and active learning practices remain prevalent in the education literature and are invaluable to higher education.

However, arguments have been made that altering instruction alone is not sufficient to address issues in STEM education. There needs to be an additional focus on how course content is assessed, with aims toward emphasizing greater conceptual understanding (Holme et al., 2010; Laverty et al., 2016) and supporting student learning (Shepard, 2000). Furthermore, assessments must function to provide feedback both to instructors and to students regarding student comprehension (Black, 1998; Council, 2001; Shepard, 2000). The way in which students are assessed is of further importance as evidence has shown that students view assessments as a communication of what specific content is important from a course (Stowe et al., 2021). As such, researchers have developed various assessment tools to target students' conceptual understanding of key concepts, including Concept Inventories (CI) (Laverty et al., 2016) and Three-dimensional Learning (Laverty et al., 2016). While evidence-based assessment types and their associated effects on students are investigated and reported on in the literature (Brownell & Tanner, 2012;

Crouch & Mazur, 2001; Freeman et al., 2014; McAlpin et al., 2022; Turpen & Finkelstein, 2009; Undersander et al., 2017), innovative grading schemes remain relatively unexplored.

Alternative grading

A growing area of interest among educational researchers and practitioners is the grading systems that are used by instructors. Specifically, some educators are moving away from traditional grading systems, such as the familiar A-F scale, and exploring alternative approaches (Clark & Talbert, 2023; Hackerson et al., 2024). The now ubiquitous A-F grading present in academia was a relatively new innovation in the history of formalized education (Williams, 2022). The A-F grading system, which emerged in response to the need for performance comparisons—such as determining eligibility for academic awards or graduate school admissions—became prevalent in the mid-1900s (Clark, 2019; Schneider & Hutt, 2014). In the U.S., this traditional grading typically utilizes a points-based scale to reflect overall student performance based on various assessments (e.g., homework and exams) and behavioral factors (e.g., attendance). This is then translated to a 100-point scale and corresponds to an overall letter grade. (Clark & Talbert, 2023).

Despite its widespread use, traditional grading has notable limitations. First, while these grades serve as evaluations of student performance, they fail to provide meaningful feedback for improvement (Cain et al., 2022). Research indicates that assigning scores or letter grades on student assignments does not meaningfully increase student learning outcomes and offers minimal insight for students regarding how they can improve (Campbell & Cabrera, 2014; Guskey, 2019; Stewart & White, 1976), students report that detailed, actionable comments are the most valuable form of feedback, as opposed to numerical grades (Guskey, 2019).

Even if traditional grades are paired with meaningful feedback, the grades themselves can be unreliable. Traditional grades often reflect factors unrelated to student academic performance,

such as a student's access to resources and prior preparation (Feldman, 2019a, 2019b; Link & Guskey, 2019; Matz et al., 2017; McKay, 2019). Instructors' biases also impact the reliability of grades. For example, underrepresented students may receive lower grades due to societal issues, such as unconscious racial, class, and gender biases (Feldman, 2019b). Additionally, grading can vary significantly between instructors depending on one instructor's standards as compared to another (Cain et al., 2022; Donaldson & Gray, 2012; Herridge & Talanquer, 2020; Herridge, Tashiro, & Talanquer, 2021). Even if students are compared on the same exams with a detailed key used by all graders, difference in course grades can still result from instructor differences as opposed to differences in learning outcomes. Indeed, instructors include different course assignments (e.g., homework, exams, extra credit, etc.) and requirements (e.g., attendance, participation, layout of assignments etc.) in their course grading scheme (Brookhart, 1991; Guskey & Link, 2019; Herridge & Talanquer, 2020; James, 2023; Mutambuki & Fynewever, 2012; Petcovic et al., 2013). As instructors determine whether to consider different parameters into their scheme, and the associated weights each factor has on the final grade, similar learning outcomes can translate to vastly different final letter grades.

Beyond the unreliable measurements of learning, students in traditional grading were found to have increased levels of anxiety and lowered levels of intrinsic motivation (Chamberlin, Yasué, & Chiang, 2018; Lewis, 2020; Pulfrey, Buchs, & Butera, 2011; Schinske & Tanner, 2014). Traditional grading systems can diminish intrinsic motivation, pushing students to focus on grades as external rewards rather than fostering a genuine interest in learning (Chamberlin, Yasué, & Chiang, 2018; Kohn, 2011; Pulfrey, Buchs, & Butera, 2011). Indeed, grades act as an extrinsic motivator due to their role in determining students' ability to earn their desired degree, maintain

positions (e.g., positions on collegiate sports teams), and be awarded different opportunities (e.g., academic fellowships, internships, etc.) (Grant & Green, 2013).

Instructors have thus begun the shift away from traditional grading methods and towards alternative grading schemes. Talbert and Clark (2023) have outlined four key elements of alternative grading: *clearly defined standards, helpful feedback, representative marks, and reattempts without penalty*. These features aim to shift the focus from evaluation to learning. *Clearly defined standards* include specific learning outcomes and detailed rubrics that outline what students need to achieve. Clear expectations when paired with *helpful feedback* provides students with a transparent path towards improving their knowledge, understanding, and performance. After students complete an assignment, they receive marks that clearly reflect their performance, such as "needs revision" or "exceeds expectations;" these are in lieu of the traditional points-based grades. Lastly, students are encouraged to revise their work without incuring a grading penalty. This multi-faceted approach fosters a learning environment where improvement and mastery are prioritized (Clark & Talbert, 2023).

Three primary models of alternative grading have gained traction in STEM fields: ungrading (Ferguson & Bonner, 2024; Newton, 2023; Rapchak, Hands, & Hensley, 2022; Spurlock, 2023; von Renesse & Wegner, 2023), standards-based grading (Beatty, 2013; Boesdorfer, Baldwin, & Lieberum, 2018; Del Carlo & Strauss, 2023; Lewis, 2020), and specifications grading (Ahlberg, 2021; Donato & Marsh, 2023; Evensen, 2022; Nilson, 2015). However, specifications grading is of particular note as it is the most widely represented alternative grading approach in chemistry higher education courses (Hackerson et al., 2024).

Reflective practices

[Note: The following is adapted from Machost, H., & Stains, M. (2023). Reflective practices in education: A primer for practitioners. CBE—Life Sciences Education, 22(2), es2.]

A review by Henderson et al. (2011) concluded that an important first step to change instructional practices is for instructors to understand their practices, beliefs, and values around teaching and to help them problematize their teaching. While this alone is not sufficient and long-term support and cultural change around teaching at the department and institution levels are also required, this step is essential as the dissatisfaction experienced once a problem is identified can be a powerful initiator for change (Andrews & Lemons, 2015). Engaging STEM instructors in reflective teaching practices is a promising strategy to help them problematize their teaching. It is common to conceptualize reflection about teaching situations as a way to help "fix" any problems or issues that present themselves (Brookfield, 2017). However, this view is counterproductive to the overarching goal of reflective practices – to continually improve one's own efficacy and abilities as an educator. As described by Brookfield, reflection can act as a "gyroscope," helping educators stay balanced amidst a changing environment (Brookfield, 2017, p. 81). Through the process of reflection, practitioners focus on what drives them to teach and their guiding principles, which define how they interact with both their students and their peers. Furthermore, reflective practitioners are deliberately cognizant of the reasoning behind their actions, enabling them to act with more confidence when faced with a sudden or difficult situation (Brookfield, 2017). While it is true that reflective practitioners are aware of areas for improvement in their teaching, it is also true that they acknowledge, celebrate, and learn from good things that happen in their classrooms and in their interactions with students and peers. As such, they are more consciously aware of their victories, even if they happen to be small (Brookfield, 2017). In a similar vein, reflective practices can help educators realize when certain expectations or cultural norms are out of their direct ability to address. These potential benefits have resulted in developing reflective practice and reflective practitioners being identified as one of four dominant change strategies in the literature

(Henderson, Beach, & Finkelstein, 2011). Specifically, developing reflective practitioners is identified as a strategy which empowers individual educators to enact change (Henderson, Beach, & Finkelstein, 2011). The practice of reflection gives instructors an opportunity to analyze their teaching practices and to learn from their own views concerning their efficacy and interactions with students (McAlpine & Weston, 2002). The positive impacts of instructors' reflections have been repeatedly reported, particularly in the K-12 literature (Ansarin, Farrokhi, & Rahmani, 2015; Belvis et al., 2012; Fox, Campbell, & Hargrove, 2011; Markkanen et al., 2020; Tajeddin & Aghababazadeh, 2018).

Larrivee (2000) suggested that there is not a prescribed strategy to becoming a reflective practitioner but that there are three practices that are necessary: 1) carving time out for reflection, 2) constantly problem solving, and 3) questioning the status quo. For educators who are new to reflective practices, it is useful to view the method as "transforming what we are already doing, first and foremost by becoming more aware of ourselves, others, and the world within which we live" (Rodgers & Laboskey, 2016, p. 101) rather than as a complete reformation of their current methods.

Current knowledge of STEM instructors' practices

Adoption of evidence-based instructional and assessment practices

Despite advocation for evidence-based practices, traditional methods of instruction remain common among STEM instructors. A landmark study investigating over 700 STEM college-level courses utilized the Classroom Observation Protocol for Undergraduate STEM (COPUS) (Smith et al., 2013) to assess instructional practices (Stains et al., 2018). This method documents the frequency of 13 student behaviors (e.g. listening, answering questions, asking questions) in two-minute intervals; notably, multiple behaviors could coincide with the same time interval. The large

scale observational data revealed that the most common instructor behavior was lecturing, with an average occurrence of approximately 75% of the total 2-min intervals within a given class (Stains et al., 2018). Indeed, just over half of the instructors had a profile where 80% or more of their class time contained lecture. The remaining half of the instructors were split between those who incorporated student-centered strategies (e.g., group work) to varying degrees (Stains et al., 2018). With respect to assessment, work has shown that instructors who design their assessments continue to factor low-level questions, such as recalling information (Davila & Talanquer, 2010; Momsen et al., 2013; Stowe et al., 2021). This does not target students' comprehension of complex concepts, nor does it adequately communicate to students the importance of conceptual understanding in STEM courses.

This low level of uptakes of evidence-based practices has been extensively explored int he literature (Sturtevant & Wheeler, 2019). Common barriers occur at a level above the individual instructors. The departmental or disciplinary culture around teaching, and the balance of teaching and research, has been reported to influence instructional practices (Lund & Stains, 2015; Michael, 2007; Sturtevant & Wheeler, 2019). Indeed, one study investigating STEM instructions across disciplines at a single institution found that such contextual factors positively influenced physics instructor's pedagogical practices yet negatively influenced that of chemistry instructors (Lund & Stains, 2015). An additional study revealed that the adoption of evidence-based practices was viewed by some as misaligned with tenure criteria, that instructors felt more pressure to devote time to research as opposed to teaching, and that instructors did not have incentives to devote time and effort to their teaching practices (Shadle, Marker, & Earl, 2017). Another high-level barrier to the adoption of evidence-based practices is the pressure to cover a vast amount of material in a single course (Andrews & Lemons, 2015; Michael, 2007; Shadle, Marker, & Earl, 2017). Lower-

level factors also can have a profound influence. Several studies have shown a relationship between classroom layout and enrollments and instructors' pedagogical practices (Michael, 2007; Sturtevant & Wheeler, 2019; Yik et al., 2022a, 2022b).

Beyond contextual factors, an instructor's personal experiences can also impact their pedagogical choices. Studies have demonstrated that various instructor-related factors impact the adoption of certain pedagogical practices including: instructors' available time to develop implementations, their training or lack thereof in pedagogical practices, and potential alignments and conflicts between a pedagogical change and their professional identity (Brownell & Tanner, 2012). Furthermore, work has indicated that the perceptions instructors have of pedagogical practices they experienced as a student can impact their adoption of such practices when in the role of an instructor (Kraft et al., 2024; Yik et al., 2022b). Experiences in professional development can be just as crucial and has been shown to correlate to a decrease in the percentage of time instructors spend lecturing (Yik et al., 2022a, 2022b). Interestingly, participation in teaching-focused workshops is also correlated with the adoption of research-based assessment tools (Gibbons et al., 2022).

Adoption of alternative grading

In stark contrast to the well-known, lagging uptake of other pedagogical practices by STEM instructors, specification grading has experienced rapid growth (Ahlberg, 2021; Blackstone & Oldmixon, 2019; Carlisle, 2020; Elkins, 2016; Evensen, 2022; Fernandez et al., 2020; Harrington et al., 2024; Helmke, 2019; Hofmeister et al., 2023; Katzman et al., 2021; Kelz et al., 2023; Kiefer & Earle, 2023; Mendez, 2018a, 2018b; Mirsky, 2018; Roberson, 2018; Tsoi et al., 2019; Williams, 2018), especially among chemistry educators. The rise of specifications grading is of further interest as there remains a lack of evidence concerning its effectiveness and impact on

students. Most publications on specifications grading to date are descriptions of implementations and anecdotal evidence based on personal experiences and instructors' reports of their students' satisfaction with specifications grading (Bunnell et al., 2023; Houseknecht & Bates, 2020; Martin, 2019; McKnelly, Morris, & Mang, 2021). Recently, there has been an emergence of studies that explore the effectiveness of specifications grading in chemistry, though the scope of these studies remains small (Ahlberg, 2021; Bunnell et al., 2023; Closser, Hawker, & Muchalski, 2024; Donato & Marsh, 2023; Howitz, McKnelly, & Link, 2021; Katzman et al., 2021; McKnelly et al., 2023). These studies examine the effectiveness of specifications grading often by comparing final exam scores and overall course grade distributions between specifications-graded and traditionally-graded courses.

There is currently no body of evidence investigating instructors' perceptions of specifications grading which lead to adoption, nor of the different barriers that prevent instructors from adopting the practice. While such aspects are not the focus of studies, a thorough investigation of the literature reveals several factors. First, student-resistance is often reported. Indeed, despite studies which show that students appreciate the opportunity to reattempt assessments (Closser, Hawker, & Muchalski, 2024; Howitz, McKnelly, & Link, 2021; Hunter, Pompano, & Tuchler, 2022), students also report feeling confusion about the criteria on assignments (Hunter, Pompano, & Tuchler, 2022), a lack of understanding of the course requirements (Closser, Hawker, & Muchalski, 2024; Howitz, McKnelly, & Link, 2021), and frustration with the lack of partial credit in the specifications grading scheme (Noell et al., 2023). Notably, the latter was also associated with increased student stress, which is antithetical to one of Nilson's hypothesized outcomes (Nilson, 2015). From an instructor-centered view, the adoption of specifications grading may be hindered by limitations in learning management systems,

specifically challenges with incorporating a non-points-based grading scheme into standard gradebooks (Joseph et al., 2023). Additionally, there is the ingrained usage of points-based grading schemes in academia and the requirement to report a letter grade. This emerges from how integrated grades are with graduate admissions, institutional ranking, and scholarship or fellowship opportunities.

Adoption of reflective practices

In higher education, recent calls for reforms on the evaluation of teaching have recognized the importance of reflections (Accelerating Systemic Change Network, 2023; Bradforth et al., 2015; Dennin et al., 2017; Simonson, Earl, & Frary, 2022; The University of Kansas Center for Teaching Excellence, 2024; Weaver et al., 2020). For instance, practicing reflective teaching is one of the criteria described in the Framework for Assessing Teaching Effectiveness (FATE; Simonson, Earl, & Frary, 2022) and an essential component of the Benchmarks for Teaching Effectiveness developed by the Center for Teaching Excellence at the University of Kansas (2024). Such teaching evaluation frameworks and guidelines are built on the premise that engaging in reflections will lead instructors to engage in instructional growth and the adoption of learner-centered practices. However, the literature on reflective practices in education has demonstrated that reflections can range in quality; this will thus affect any outcomes (Dyment & O'connell, 2010; O'Connell & Dyment, 2011; Ryan, 2013; Spalding & Wilson, 2002).

The teaching evaluation frameworks often describe reflections in broad terms and provide limited scaffolding or examples of effective reflection. FATE specifies that an effective reflection "demonstrates a high level of self-reflection around teaching broadly, objectively describing their strengths and weaknesses, consistent with evidence of teaching practices" (Simonson, Earl, & Frary, 2022, p. 170). Similarly, the Benchmarks for Teaching Effectiveness describes someone

who “regularly adjusts teaching based on reflection on student learning, within or across semesters and examines student performance following adjustments” as an expert in reflection (The University of Kansas Center for Teaching Excellence, 2024).

Unfortunately, few studies have explored the nature and quality of STEM instructors’ reflections, whether as part of the teaching evaluation frameworks previously discussed or as a part of a separate research study. What is known is that the adoption of reflective practices must be done in a way that does not negate its benefits. For example, Galea (2012) highlights the negative effects of routinizing or systematizing this extremely individual and circumstance-based method (e.g., identification of specific areas to focus on, standardized timing and frequency of reflections). In doing so, the systems that purportedly support teachers using reflection remove their ability to think of creative solutions, limit their ability to develop as teachers, and can prevent an adequate response to how the students are functioning in the learning environment (Tan, 2008). Effective reflection can be stifled when reflections are part of educators’ evaluations for contract renewal, funding opportunities, and promotions and tenure. Reflective practices are inherently vulnerable as they involve both being critical of oneself and taking responsibility for our personal actions (Larrivee, 2008b). Being open about areas for improvement is extremely difficult when it has such potential negative impacts one’s career. However, embarking on honest reflection privately, or with trusted peers and mentors, can be done separately from what is presented for evaluation.

The importance of and frameworks for studying instructors

The necessity of understanding and studying faculty practices and motivations in higher education cannot be overstated. Instructors are the keystones of pedagogical improvement. In addition to teaching their students course content, instructors are often the designers of both

curriculum and assessments while also serving as mentors. As such, instructors influence the quality of education, the institutional culture, and the success of their students. By researching the various dimensions of instructors' implementation of pedagogical innovations and practices, it is possible to support faculty professional development, aid instructors with improving student outcomes, and inform instructors of various methods for enhancing STEM education. The primary step that researchers can bring to instructors is a third party which provides evidence for or against different strategies, thus enabling instructors to make more informed decisions. Additionally, through recognizing the practices that work for instructors in various contexts, researchers can specifically target those who will most benefit them. Indeed, when striving to continually improve STEM education, it is necessary to understand and aid those who are in the classrooms and actually teaching the students. In this work, we rely on two different frameworks: the Teacher-Centered Systemic Reform model and the Diffusion of Innovations model.

Teacher-Centered Systemic Reform model

One common model used for framing educational research focused on educators is the Teacher-Centered Systemic Reform (TCSR) Model. The TCSR model was emergent from a thorough review of secondary educational reform literature and highlights several factors that contribute to pedagogical decisions (Woodbury & Gess-Newsome, 2002). In particular, the TCSR model shows the interconnected nature of personal factors, teacher thinking factors, contextual factors, and instructional practices. Personal factors are those that are specific to individual educators. They include characteristics such as an instructor's age, race, gender, and culture as well as aspects such as teaching experience and experiences as a student. Teacher thinking factors are representations of the knowledge and belief individual educators have about the nature of teaching and learning. Finally, contextual factors encompass components of the structural and cultural

environments surrounding educators. These factors are widely varied and include aspects such as required textbooks, guidance from professional organizations, the physical layout of a classroom, and the cultural norms of an institution or department. These three categories of factors influence each other and heavily contribute to instructors' educational practices. Furthermore, this relationship is not uni-directional; rather, the TCSR model depicts a complex system wherein educational practices can also affect different personal, teacher thinking, and contextual factors. While the TCSR model is effective for framing instructor-centered research, an additional framework is useful for examining the process of adopting specific practices.

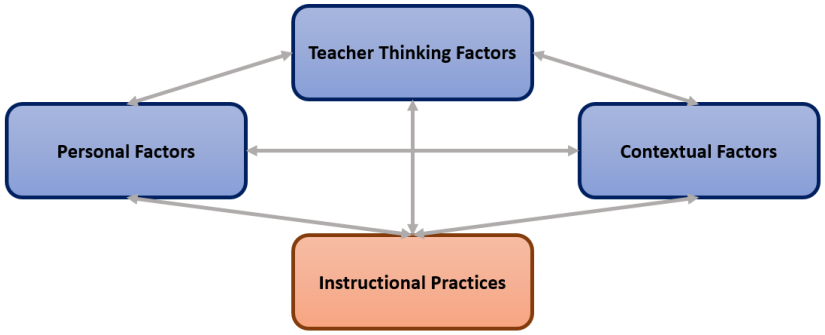

Figure 1. TCSR Model.

Diffusion of Innovations model

Rogers' Diffusion of Innovations (DOI) theory (Rogers, 2003) is ideal for educational settings due to its proven applicability and comprehensive insight into the factors influencing the adoption and dissemination of pedagogical practices (Andrews & Lemons, 2015; Genné-Bacon, Wilks, & Bascom-Slack, 2020; Henderson, Dancy, & Niewiadomska-Bugaj, 2012; Kraft et al., 2024; Lund & Stains, 2015; McConnell, Montplaisir, & Offerdahl, 2020).

The DOI theory has five primary stages that describe an instructor's decision to adopt a practice (Figure 1). In the first stage, an instructor gathers knowledge about a pedagogical practice that is new to them. During the persuasion stage, the instructor then considers their particular context and different features of the innovation to form an opinion about the innovation and its fit for their context. Following this stage, the instructor decides whether to adopt the innovation or not. If the decision to adopt is made, the instructor then follows through with testing the innovation - implementation stage. If the implementation is deemed successful, the instructor will integrate the innovation in their practice either as is or with modifications to fit their needs. Notably, while these stages do follow a logical progression, they are interrelated. Thus, there is not a strict, linear progression from the first to the last stage, and, according to the nature of the confirmation stage, the process is necessarily cyclical in parts. Indeed, Rogers specified that the first three stages in particular are not a strictly set path that individuals follow (Rogers, 2003).

Rogers (2003) described in his model four factors that affect the rate of adoption of an innovation: (1) prior conditions and the context of an individual before they begin the process at the knowledge stage, (2) the personal characteristics of an individual who is involved in the process, (3) the attributes an individual perceives an innovation to have, and (4) the communication channels that are used to inform and propagate the innovation (Figure 2).

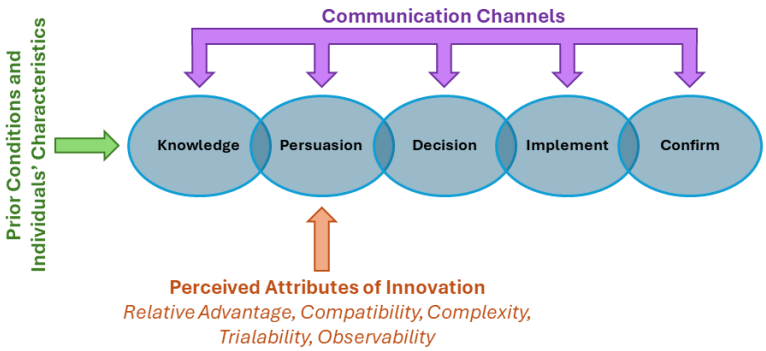

Figure 2. Depiction of Rogers' DOI theory.
Adapted from Kraft et al. 2024.

Additionally, Rogers identified five attributes of an innovation which account for the majority of the variation in rate of adoption (Rogers, 2003): *relative advantage, compatibility, complexity, trialability,* and *observability*. *Relative advantage* describes "the degree to which an innovation is perceived as being better than the idea it supersedes" (Rogers, 2003, p. 212). *Compatibility* refers to how well the innovation is perceived to align with an individual's personal values, past experiences, and situational needs. *Complexity* describes how "difficult to understand and use" an innovation is (Rogers, 2003, p. 242). Finally, *trialability* describes "the degree to which an innovation may be experimented with on a limited basis" (Rogers, 2003, p. 243), and *observability* "is the degree to which the results of an innovation are visible to others" (Rogers, 2003, p. 244).

Rogers further posits five sequential waves of adoptees according to their willingness to adopt innovations (Figure 3): innovators (2.5%), early adopters (13.5%), early majority adopters (34%), late majority adopters (34%), and laggards (16%) (Rogers & Shoemaker, 1971). Innovators are the first to adopt an innovation and are risk-takers. They are eager to try innovations and have

a high level of comfort with both uncertainty and handling potential backlash. Early adopters are more likely to be somewhat comfortable with taking risks and handling potential adversity. However, they still only adopt innovations after careful consideration. Early adopters are crucial for the speed of innovation propagation, and they are often sources of advice or information for others. The early majority group adopts an innovation before the majority of those in their profession but after the early adopters. This group expresses more skepticism and is often persuaded by evidence produced by the early adopters. Thus, they maintain an open, yet cautious, approach to innovations, and serve as a bridge between risk-taking and risk-averse groups. Those in the late majority group tend to adopt an innovation after most of their peers have done so. Notably, the motivations ascribed to this group center on social pressure, practical necessity, or a desire to conform. Thus, they may adopt an innovation they view positively only after their social circle has adopted it, or they may adopt an innovation they are skeptical about due to pressure from their peers. Laggards are the last group to adopt an innovation. They may have either a commitment to traditional methods and/or a deep skepticism of new practices. Thus, laggards are those who require the most substantial proof of the benefits of an innovation before adoption (Rogers & Shoemaker, 1971).

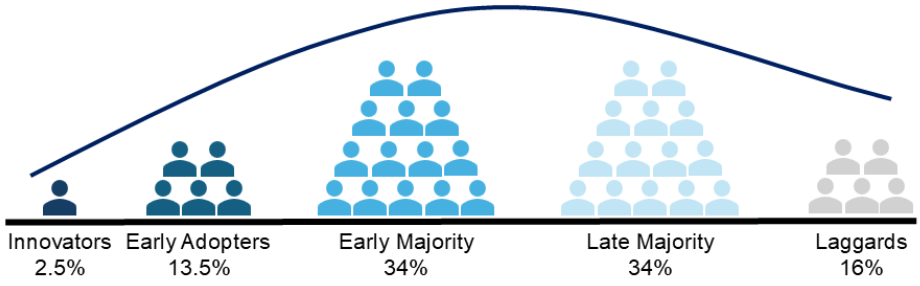

Figure 3. Stages of innovation adoption by a population

Overview of Dissertation

This dissertation highlights several works which have contributed to the understanding of faculty practices and their adoption of innovations in higher education. Guided by the foundations of the TCSR model, work in the field which has sought to understand faculty pedagogical decisions, and the DOI framework, the research described herein explores both aspects of instructors' thinking as well as the different components and states of innovation adoption.

Chapter 1 investigates the reflective writings of STEM faculty. This work is one of very few in the literature which captures the nature and nuances of STEM instructor's reflective writing. The content, depth of analysis, and different factors considered in the reflective writings are analyzed and examined for trends; the results afford an in-depth look at how instructors respond to and rationalize critical incidents they encounter while teaching. Further, as the focus lies in novice instructors, the insights gained provide valuable information for professional development and change agents. Indeed, the results have already impacted the design of a long-standing national new-faculty workshop.

Chapter 2 is emergent from the same dataset as chapter 1. However, chapter 2 focuses on the unexpected, yet vitally important, emotional responses that are discussed in instructors' reflective writings. This work provides valuable insight into an often over-looked aspect of education – the emotions instructors experience while teaching. Little work has investigated what STEM instructors feel while teaching; thus, the research in chapter four aims to both destigmatize emotion in the classroom and provide valuable information on the emotions commonly felt and the associated causes in the classroom. As research has demonstrated a link between instructor emotions and teaching practice, chapter four is invaluable to instructor training and classroom preparation.

Chapter 3 investigates another aspect of instructor thinking; however, instead of investigating emotional responses and causes in the classroom, it examines instructors' motivation for introducing an innovation into their course. The use of alternative grading schemes is of growing interest and prominence in STEM higher education, with specifications grading the most common among chemistry instructors. Chapter 3 of this dissertation investigates why instructors chose to adopt specifications grading in their courses through exploring the instructors' perception of specifications grading's relative advantage over traditional grading. The incites provided by this work have direct impacts for the propagation of not only specifications grading, but also other beneficial pedagogical innovations among STEM instructors.

Chapter 4 expands upon the investigation into specifications grading. Specifically, the details of implementation are reported from a variety of chemistry higher education courses. Through this work, common components of specifications grading are identified and trends are examined across course types. The complex picture of specifications grading implementation presented by this work showcases a need to correlate further research with specific implementations to observed outcomes and student effects, rather than attributing benefits or consequences to specifications grading as a whole. Additionally, the different methods of implementation that are detailed can act as a guide or starting point for instructors who are considering the adoption of specifications grading.

Chapter 5 follows by examining student perceptions of specifications grading. The results indicate students have mixed opinions about the efficacy of specifications grading, with both traditional grading and specifications grading being favorable in regards to different factors tested. Notably, this work provides an avenue for empirically testing the student-centered outcomes of specifications grading that were hypothesized by Nilson. Additionally, the instrument presented

can be adapted to a variety of alternative grading schemes, enabling both research and instructors to investigate how students perceive different innovative grading schemes.

In its totality, this dissertation is representative of the multifaceted approach needed when researching instructors and innovations in STEM higher education.

Part 1 Introduction: An overview of reflection in higher education

The following is adapted from: Machost, H., & Stains, M. (2023). Reflective practices in education: A primer for practitioners. *CBE—Life Sciences Education*, 22(2), es2. which is available via open-access publication with copyright retained by the authors.

The origin of reflective practices does not lie in academia, but in professional training. It is often traced back to Donald Schön's instrumental 1983 work "The Reflective Practitioner," which targeted non-academic professionals (Munby & Russell, 1989).

> "In the varied topography of professional practice, there is a high, hard ground where practitioners can make effective use of research-based theory and technique, and there is a swampy lowland where situations are confusing 'messes' incapable of technical solution. The difficulty is that the problems of the high ground, however great their technical interest, are often relatively unimportant to clients or to the larger society, while in the swamp are the problems of greatest human concern" (Schön, 1983, p. 42)

Schön's work on the education of professionals gained traction as he diverged from common norms of the time. In particular, he disagreed with separating knowledge and research from practice, and methods from results (Newman, 1999; Schön, 1983). In doing so, he advocated for professionals to develop greater competency in various real-world situations. This ideology became foundational in teaching reflective practices within education (Munby & Russell, 1989).

John Dewey, a psychologist and philosopher who was heavily influential in educational reform, provides a clear description which inspires the works in this dissertation. Reflection is "the active, persistent and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it and the further conclusions to which it tends" (Dewey, 1933, p. 9). The act of reflection in this context is meant to indicate a *process*, with Dewey highlighting the necessity of active thinking when encountering obstacles and problems. In less philosophical phrasing, reflection entails considering past or present experiences, learning from the outcomes observed, and planning how to better approach similar situations in the future. Consequently, Dewey suggests that educators embark on a journey of continual improvement when engaging in reflective practices. Ensuing work on reflection in education focused on gaining evidence for the effectiveness of reflective practices (Dervent, 2015; Zahid & Khanam, 2019) and understanding the obstacles that can prevent their adoption (Davis, 2003; Sturtevant & Wheeler, 2019). Despite the evidenced interest, many educators only engage in reflection when completing documents or tasks relevant to professional advancement, such as yearly evaluations or tenure applications. Such instances of reflective practice may be ineffective due to the perception of judgment preventing authenticity (Brookfield, 2017). This concern is further complicated by the process of reflection being limited to isolated instances. As such, the current state of reflective practices in higher education is lacking compared to its original conception. Thus, it is necessary to advocate for reflective practices as a continual exercise while simultaneously working to understand how instructors reflect.

Reflection in practice

Many choose to be educators to help inform, mentor, or guide others. With such a broad aim can come numerous challenges, and reflective practices can help educators navigate these

challenges. It is common to conceptualize reflection as a way to help "fix" any problems or issues that present themselves in educational environments (Brookfield, 2017). However, this view can be counterproductive to the overarching goal of reflective practices as there is always room for improvement, even among experienced educators. Classrooms are dynamic environments, with changing students, evolving technology, and emerging curricula and pedagogical norms. Each of these can alter the modes of instruction or the concepts and skills being taught. Brookfield describes reflection as a "gyroscope," helping educators stay balanced in this changing environment (Brookfield, 2017, p. 81). Through reflective practices, educators focus on their main motivations, aims, and guiding principles. This can enhance how they interact with both their students and their peers. Furthermore, reflective educators are mindful of the reasoning behind their actions, which can foster confidence when faced with a sudden or difficult situations (Brookfield, 2017). In this way, reflection can help guide educators through the challenging times they may experience in their career.

Indeed, reflective practitioners often analyze critical incidents when engaging in effective reflection. Critical incidents are used because meaningful reflection is often a result of experiencing a problem or some form of cognitive dissonance concerning teaching practices and approaches to their students (Lee, 2005). As such, it is most effective to combine techniques, which are outlined later in this section, with a critical incident to force practitioners into a new and difficult position relating to education. Larrivee details that a sense of "uncertainty, dissonance, dilemma, problem, or conflict" is extremely valuable to personal reflection and growth (Larrivee, 2008b, p. 93). Thus, unsettling experiences encourage changes to action far more than reflecting on typical teaching/learning interactions. This is an inherently uncomfortable experience for the practitioner as feelings of self-doubt, uncertainty, anger, and self- or peer-rejection can come to the

surface (Larrivee, 2008b). Yet, it is when educators are in an uncomfortable position that they are best able to challenge their learned assertions about what they are teaching and how they are supporting their student learning. This requires a conscious effort on the part of the educator. Humans tend to function automatically based on their past experiences and ingrained beliefs. This results in certain aspects of events being ignored while others become the driving force behind reactions. In a sense, humans have a "filter system" which can unconsciously eliminate the most effective course of action; this results in humans functioning in a cycle where current, unquestioned beliefs determine which data and experiences are given attention (Larrivee, 2000, p. 295).

Critical incidents highlight any dissonance present in one's actions, enabling practitioners to tackle social, ethical, political, and pedagogical issues, which may be systemic to their department, their field, or their culture. Critical incidents foster critical reflection (under the depth and content-based models) even in novice teachers (Griffin, 2003; Pultorak, 1996). It is *because* of the difficulty and uncertainty posed by critical incidents that they are widely promoted as an invaluable aspect of reflective practices in education. Therefore, the analysis of critical incidents, whether they are case studies or theoretical examples, have been used in educating both pre-service (Griffin, 2003; Harrison & Lee, 2011) and current educators (Benoit, 2013).

Important parts of reflective practice, which are often overlooked, are the evaluation of successes in the classroom and of factors outside of an instructor's control. Educators often grapple with imposter syndrome, or the sense that their work is insufficient despite their accomplishments (Brems et al., 1994; Collins et al., 2020; Parkman, 2016). Reflection encourages educators to acknowledge, celebrate, and learn from good things that happen in their classrooms and in their interactions with students and peers. Thus, reflective practices can serve as an important reminder

to instructors of their victories, even if they are small (Brookfield, 2017). Additionally, reflection can also help educators recognize when certain expectations or cultural norms are outside of their control (Brookfield, 2017). For example, systemic issues such as racism, sexism, and ableism cannot be solved by individual educators. While a single person can impact their classroom's dynamics to a point, institutions must complement educators' efforts through systems that address inequality. Thus, reflective practices can prevent educators from feeling guilt or blame for large, multifaceted issues. Indeed, reflection on positive experiences and the extent of one's influence can be vital for the success of educators.

What are the different types of reflection?

Reflective practices have been described based on their timing, depth, and content. Notably, reflective practices must span all types of reflection in order to be the most effective for educators (Griffiths & Tann, 1992).

Time-dependent

Schön laid the groundwork for the time-dependent reflections (Schön, 1983). He defines two concepts: '*reflection-in-action'* and *'reflection-on-action,'* which are differentiated based on when the reflection takes place. *Reflection-in-action* is when educators reflect on the actions they are currently taking. Contrastingly, r*eflection-on-action* takes place after the situation being reflected on has taken place. During *reflection-on-action,* educators analyze different influencing factors and carefully consider the outcomes resultant from their actions. However, *reflection-in-action* is perceived as more difficult since instructors must simultaneously analyze the situation they are in and act accordingly. Later work built on this initial description of time-dependent reflections. Loughran renamed the original two timings in order to be more intuitive and added one time point (Loughran, 2002). The three categories include: *anticipatory, contemporaneous*,

and *retrospective. Anticipatory* reflection describes when instructors use their experience to prepare for potential situations they may encounter in their classrooms. *Contemporaneous* and *retrospective* reflection mirrors Schön's *reflection-in-action* and *reflection-on-action,* respectively. Practicing reflection at different points in time can be highly beneficial to instructors, and the two time-dependent models function in tandem with the depth- or content-based understandings of reflections which are described below.

Depth of reflections

Conceptualizing reflection in terms of depth has a long history in the literature. A comprehensive depth-classification model was developed by Larrivee (2008a) after an extensive review of the literature. This classification includes a progression in reflective practices across four levels: *pre-reflection*, *surface*, *pedagogical*, and *critical reflection*.

During the *pre-reflection* stage, educators do not engage in meaningful reflection. They are functioning in "survival mode" (Campoy, 2010, p. 17; Larrivee, 2008a, p. 350), reacting automatically to situations without considering alternatives or potential impacts on the students (Campoy, 2010; Larrivee, 2008a). At this stage, educators may feel little agency, consider themselves the victim of circumstance, or do not recognize their role while simultaneously blaming others. (Campoy, 2010; Larrivee, 2008a). They are unlikely to question the status quo, thereby failing to consider and adapt to the needs of the various learners in their classroom (Campoy, 2010; Larrivee, 2008a). While the description of educators at this level is non-ideal, educators at the pre-reflection level are not ill-intended. The pre-reflective level is present among practitioners as evidenced in a 2015 study investigating 140 ESL educators and a 2010 analysis of collected student reflections (Ansarin, Farrokhi, & Rahmani, 2015; Campoy, 2010). The presence of pre-reflective educators is also readily apparent in the authors' ongoing research. As such, being aware of the

pre-reflection stage is necessary for beginning practitioners, and this knowledge is perhaps most useful for designers of professional development programs.

The first true level of reflection is *surface reflection*. At this level, educators are concerned about achieving a specific goal, such as high scores on standardized tests. However, these goals are only approached through conforming to departmental norms, evidence from their own experiences, or otherwise well-established practices (Larrivee, 2008a). In other words, educators at this level question whether the specific pedagogical practices will achieve their goals, but they do not consider any new or non-traditional pedagogical practices or question the current education policies (Campoy, 2010). Educators' reflections are grounded in personal assumptions and influenced by individuals' unexamined beliefs and unconscious biases.

At the *pedagogical* level, educators "reflect on educational goals, the theories underlying approaches, and the connections between theoretical principles and practice." (Larrivee, 2008a, p. 343). At this level, educators also consider their own belief system and its relationship to their practice and explore the problem from different perspectives. A representative scenario at this level includes: a teacher contemplating their various teaching methods and considering their observed outcomes in student comprehension, alternative viewpoints, and also the current evidence-based research in education. Subsequently, they alter (or maintain) their previous teaching practices to benefit the students. In doing so, more consideration is given to possible factors than is seen with surface level reflection. This category is quite broad due to the various definitions present in the literature (Larrivee, 2008a). However, there is a common emphasis on the theory behind teaching practices, ensuring that practice matches theory, and the student-outcomes of enacted teaching practices (Larrivee, 2008a).

The last level of reflection categorized by Larrivee is *critical reflection*, wherein educators consider the ethical, moral, and political ramifications of who they are and what they are teaching to their students (Larrivee, 2008a). An approachable way of thinking about critical reflection is that the practitioner is challenging their assumptions about what is taught and how students learn. In doing so, educators evaluate their own views, assertions, and assumptions about teaching, with attention paid to how such beliefs impact students both as learners and as individuals (Larrivee, 2005, 2008b). Through practicing critical reflection, societal issues that affect teaching can be uncovered, personal views become evidence-based rather than grounded in assumptions, and educators are better able to help a diverse student population.

Larrivee used this classification to create a tool for measuring the reflectivity of teachers (see section 4.1 of the Supplementary Materials), which was leveraged for this dissertation.

Content of reflections

The third type of reflection is one in which *what* is being reflected on is the defining feature. One such example is Valli's five types of reflection (Valli, 1997): *technical reflection*, *reflection-in and on-action*, *deliberative reflection*, *personalistic reflection*, and *critical reflection*. Note that Valli's conceptions of the two types of reflection - *reflection-in and on-action*, and *critical reflection* - are congruent with the descriptions provided in the Time-Dependent and Depth of reflections earlier sections, respectively, and will thus not be detailed in this section.

In a *technical reflection,* an educator evaluates their instructional practices in light of the findings from the research on teaching and learning (Valli, 1997). The quality of this type of reflection is based on the educator's knowledge of this body of work and the extent to which their teaching practices adhere to it. For example, an educator would consider whether they are providing enough opportunities for their students to explain their reasoning to each other during

class. This type of reflection does not focus on broader topics such as the structure and content of the curriculum or issues of equity.

Deliberative reflection encompasses "a whole range of teaching concerns, including students, the curriculum, instructional strategies, the rules and organization of the classroom" (Valli, 1997, p. 75). In this case, deliberative comes from the practitioner having to debate various external viewpoints and perspectives, or research which maybe be in opposition to each other. As such, they have an internal deliberation when deciding on the best actions for their specific teaching situation. The quality of the reflection is based on the educator's ability to evaluate the various perspectives and provide sound reasoning for their decisions.

Personalistic reflection involves an educator's personal growth as well as the individual relationships they have with their students. Educators engaged in this type of reflection thoughtfully explore the relationships between their personal and professional goals and consider the various facets of students' lives with the overarching aim of providing the best experience. The quality of the reflection is based on the educator's ability to empathize.

In order to manage the limitations of each type of reflection, Valli recommended that reflective practitioners do not focus solely on a specific type of reflection but rather engage with multiple as they each address different questions. It is important to note that some type of reflection may be prerequisite to others and that some may be more important than others; for example, Valli stated that critical reflections are more valuable than technical reflections as they address the important issues of justice.

Scaffoldings for Reflection

Scaffoldings have been created to aid novice practitioners when first beginning the cyclical process of reflection. Bain et al. (2002) created the 5R framework in order to support the

development of pre-service teachers into reflective practitioners. The framework includes the following five steps (Bain et al., 2002):

1. **Reporting** involves considering a particular experience and the contextual factors that surround it.
2. **Responding** is when the individual practitioner verbalizes their feelings, thoughts, and other reactions that they had in response to the situation.
3. **Relating** is defined as the teacher making connections between what occurred recently and their previously obtained knowledge and skill base.
4. **Reasoning** then encourages the practitioner to consider the foundational concepts and theories, as well as other factors that they believe to be significant, in an effort to understand why a certain outcome was achieved or observed.
5. Finally, **reconstructing** is when the teacher takes their explanation and uses it to guide their future teaching methods, either to encourage a similar result or to foster a different outcome.

This framework facilitates an understanding of what is meant by and required for reflective practices.

Another popular scaffolding for promoting reflective practices is the reflective learning cycle described by Gibbs (1988). This cycle for reflection has been extensively applied in teacher preparation programs and training of health professionals (Ardian, Hariyati, & Afifah, 2019; Husebø, O'Regan, & Nestel, 2015; Markkanen et al., 2020). The cycle consists of six stages:

1. **Description**: The practitioner first describes the situation to be reflected on in details.
2. **Feelings**: The practitioner then explore their feelings and thoughts processes during the situation.
3. **Evaluation**: The practitioner identifies what went well and what went wrong.

4. **Analysis**: The practitioner makes sense of the situation by exploring why certain things went well while others did not.
5. **Conclusions**: The practitioner summarizes what they learned from their analysis of the situation.
6. **Personal action plans**: The practitioner develops a plan for what they would do in a similar situation in the future and what other steps they need to take based on what they learn (e.g., gain some new skills or knowledge).

These two models are complimentary to one another, and we have formulated a proposed scaffolding for reflection by combining the two models. In Table 2, we provide a short description of each step and examples of reflective statements. The full scaffolding is provided in Appendix A.1 in the distributed online survey.

Reflective practices are widely advocated for in academic circles. Reflective practices are a *process*, and are a time- and energy-intensive, but extremely valuable tool for educators when implemented with fidelity. Therefore, reflection is vital for efficacy as an educator and a requirement for instructors to advance their life-long journey as educators.

Aim of Part 1

While the effectiveness of reflection in teaching is widely acknowledged, there remains a limited understanding of both the fundamental nature of instructors' reflections and of the needs that instructors identify in when engaging in reflective practices. This gap in the literature is significant, as effective reflection necessitates deep thinking, a consideration of different types of content, and an explicit effort by instructors to consider societal, cultural, and contextual influences in their classrooms. However, such reflections may not be produced, thus, negating the potential effects when promoting reflection. Through the work described herein, insight is provided into the

starting point of instructors' reflections as they begin reflective practice. This can then enable targeted professional development programs that encourage effective adoption of this widespread approach to instructional reform.

Chapter 1. Exploring the Nature of STEM Instructors' Written Reflections

The following is available via open-access publication with copyright retained by the authors: Machost, H., Kable, E. A., Mitchell-Jones, J. K., Yik, B. J., & Stains, M. (2024). Characterization of physics and astronomy assistant professors' reflections on their teaching: can they promote engagement in instructional change? *Disciplinary and Interdisciplinary Science Education Research*, *6*(1), 14.

Introduction

In response to extensive evidence for the inequitable and poor learning outcomes experienced by students enrolled in science, technology, engineering and mathematics (STEM) courses (e.g., Hatfield, Brown, & Topaz, 2022; Koester, Grom, & McKay, 2016; Matz et al., 2017), calls for enhancing these learning environments have been broadcasted for decades by government bodies (Olson & Riordan, 2012), higher education organizations (Boyer Commission on Educating Undergraduates in the Research University, 1998; Miller & Fairweather, 2015), and STEM faculty themselves (Bradforth et al., 2015). Discipline-based education researchers and higher education researchers have been answering these calls by empirically investigating how students learn in STEM (e.g., Pond & Chini, 2017; Wu & Rau, 2019), cognitive and affective challenges they experience in these courses (e.g., Marshman et al., 2018; Rice, Lopez, & Richardson, 2013; Sorby, Veurink, & Streiner, 2018), and leveraging findings from these studies to develop and test the efficacy of innovative instructional practices (e.g., Chasteen et al., 2016; Henderson, Beach, & Finkelstein, 2011; Madsen, McKagan, & Sayre, 2017; Mooring, Mitchell, & Burrows, 2016). As these evidence-based instructional practices emerged, different communities have strived to propagate them to STEM instructors. A review of studies on strategies to promote instructional change demonstrates the complexity of this endeavor (Henderson, Beach, & Finkelstein, 2011) and recent studies suggest that the uptake of these practices has been slow across STEM fields (e.g., Beane, McNeal, & Macdonald, 2019; Stains et al., 2018; Yik et al., 2022a). The Henderson

et al. (2011) review concluded that one important first step to change instructional practices is for instructors to understand their practices, beliefs, and values around teaching and to help them problematize their teaching. While this alone is not sufficient and long-term support and cultural change around teaching at the department and institution levels are also required, this step is essential as the dissatisfaction experienced once a problem is identified can be a powerful initiator for change (Andrews & Lemons, 2015).

Engaging STEM instructors in reflective teaching practices is a promising strategy to help them problematize their teaching. Indeed, reflections provide opportunities for instructors to critically analyze their teaching practices and learn from these analyses to enhance instructional effectiveness and ultimately students' experiences (McAlpine & Weston, 2002). The positive impacts of instructors' reflections have been reported extensively, especially in the K-12 literature (Ansarin, Farrokhi, & Rahmani, 2015; Belvis et al., 2012; Fox, Campbell, & Hargrove, 2011; Markkanen et al., 2020; Tajeddin & Aghababazadeh, 2018). In higher education, many of the calls for reforms on the evaluation of teaching and evaluation of teaching frameworks have also recognized the importance of reflections (Accelerating Systemic Change Network, 2023; Bradforth et al., 2015; Dennin et al., 2017; Simonson, Earl, & Frary, 2022; The University of Kansas Center for Teaching Excellence, 2024; Weaver et al., 2020). For example, practicing reflective teaching is one of the four criterion described in the Framework for Assessing Teaching Effectiveness (FATE; Simonson, Earl, & Frary, 2022) and an essential component of the Benchmarks for Teaching Effectiveness developed by the Center for Teaching Excellence at the University of Kansas (2024).

These teaching evaluation frameworks and guidelines are built on the premise that engaging in reflections will lead instructors to engage in instructional growth and the adoption of

learner-centered practices. However, the literature on reflective practice has demonstrated that reflections can range in quality and therefore may not lead to expected outcomes (Dyment & O'connell, 2010; O'Connell & Dyment, 2011; Ryan, 2013; Spalding & Wilson, 2002). The teaching evaluation frameworks describe reflections in broad terms and provide limited scaffolding. For example, FATE describes an exemplary reflection as one that "demonstrates a high level of self-reflection around teaching broadly, objectively describing their strengths and weaknesses, consistent with evidence of teaching practices" (Simonson, Earl, & Frary, 2022, p. 170). Similarly, the Benchmarks for Teaching Effectiveness describes someone with an expert level of reflection as an individual who "regularly adjusts teaching based on reflection on student learning, within or across semesters and examines student performance following adjustments" (The University of Kansas Center for Teaching Excellence, 2024). The literature on reflective practice has demonstrated that certain scaffoldings and methods are more effective at prompting high level reflections (i.e., reflections in which the instructor considers their roles, beliefs system and knowledge about teaching and the place these play in the education of their students) and, therefore, at problematizing teaching. Unfortunately, few studies have explored the nature and quality of STEM instructors' reflections, whether as part of the teaching evaluation frameworks previously discussed or when instructors are provided with a specific, empirically-derived scaffold. It is necessary to first determine whether instructors are functioning as reflective practitioners on the level required to result in instructional change in order to design effective trainings and interventions involving reflective practice. Consequently, the goal of this study is to expand our understanding of the nature of STEM instructors' reflections by analyzing responses from physics and astronomy assistant professors to a specifically-designed reflective scaffold. The following research questions drive this study:

1. What is the nature of a difficult or challenging teaching experience (i.e., critical incident) new postsecondary physics and astronomy instructors choose to reflect on?
2. What is the content of new postsecondary physics and astronomy instructors' reflective writings when prompted to consider a critical incident?
3. What depth of reflection do new postsecondary physics and astronomy instructors spontaneously reach?
4. What types of plans are new postsecondary physics and astronomy instructors proposing to address their critical incident?
5. To what extent are the nature of the critical incident, content of reflections, and plans outlined associated to the depth of these new postsecondary physics and astronomy instructors' reflection?

Reflective practice

Reflective practice has a history grounded in philosophy and the concept of reflective thinking, particularly in the work of John Dewey (1933). The transition of reflective thinking to reflective practice- wherein the process of reflection is formalized and often recorded in some manner- lies in the realm of professional training, a shift which was catalyzed by the combined works of Schön (1987, 1991). Subsequently, Schön's concept of reflective practice has become extrememly influential in the training of educators and healthcare professionals (Munby & Russell, 1989). Reflective practice is a process by which one considers past, present, or hypothetical experiences in light of personal belief system, assumptions, and knowledge base related to these experiences in order to gain insight concerning the factors at play as well as to plan for future, similar situations (Machost & Stains, 2023).

Reflective practices can be implemented through a variety of written, recorded, and oral methods (Machost & Stains, 2023). No matter the modality, the effectiveness of reflective practices stems from enabling instructors to deeply contemplate both their experiences and the knowledge they gained through those experiences (Machost & Stains, 2023; Osterman & Kottkamp, 2004). Indeed, by practicing continual and cyclical reflective practices, instructors can become more aware of their current pedagogical content knowledge and how they continually develop knowledge (Loughran, 2002). For this reason, reflective practices have been adopted as important components of the professional development of educators (Marshall, 2019; McAlpine et al., 2004).

Reflection promotes greater effectiveness through encouraging planning for future experiences (Bain et al., 2002; Mohamed, Rashid, & Alqaryouti, 2022; Zahid & Khanam, 2019), focusing on one's strengths (Brookfield, 2017; Mohamed, Rashid, & Alqaryouti, 2022), and considering weaknesses and potential areas of improvement (Bain et al., 2002; Huda & Teh, 2018; Mohamed, Rashid, & Alqaryouti, 2022). In this way, reflection can problematize one's action and inspire the adoption of new approaches. Indeed, reflective practices are proposed to act as a "gyroscope" when navigating various external influences on the classroom, such as new departmental initiatives (Brookfield, 2017). Furthermore, it has been posited that "without routinely engaging in reflective practice, it is unlikely that practitioners in higher education will comprehend the effects of their inspirations, motivations, expectations and experiences upon their practice" (Lubbe & Botha, 2020, p. 290). For instance, through thoughtful reflection, instructors may realize how their own beliefs about the difficulty of a subject affect their explanations in class, or how their feelings of self-doubt affect their actions during office hours. Essentially, reflective practice acts as a magnifying glass, where instructors are able to analyze their actions and thoughts in relation to their experiences.

Analytical frameworks for reflections

Different frameworks have been presented in the literature to describe the nature and quality of reflections. Some frameworks focus on the variety of reflection types presented in one whole reflection (i.e., content), while others aim to evaluate hierarchically the depth of the reflection as a whole. The most popular frameworks leveraged in the literature that address these two aspects are presented below.

Content of reflections. One predominant method of analyzing reflections is based on the content discussed within the reflection itself. This method originated with the work of Valli (1997). Within this model, there are five distinct types of reflection (Table 1): reflection-in and on-action, deliberative, technical, personalistic, and critical reflections. *Reflections-in and on-action* were derived from the work of Schön (1983) and relate to when the instructor is engaging in reflection, either while teaching (in-action) or after the act of teaching (on-action). *Deliberative reflections* are concerned with weighing different perspectives, opposing research findings, or varying personal viewpoints to determine the best course of action. *Technical reflections* are specifically concerned with following the guidelines put forth by a professional organization outside of the instructor; additionally, these guidelines must be based on pedagogical research to be considered technical-type reflection. *Personalistic reflections* involve "an educator's personal growth as well as the individual relationships they have with their students" (Machost & Stains, 2023, p. 5). Finally, *critical reflections* center on an instructors' own values, assertions, and assumptions about topics such as gender, accessibility accommodations, and cultural differences. Notably, the different types of reflection can occur simultaneously within the same piece of reflective writing.

Table 1. Content of reflections based on Valli (1997)

Content type	Definition
In- and On-Action	Content focused on an instructor's own past experiences, either retrospectively or in the moment.
Deliberative	Content centered on debating viewpoints, perspectives, or research, which are in opposition to each other. This content is associated with an instructor deciding on which pedagogical practices to change, alter, or implement.
Technical	Content centered on an instructors' pedagogical practices and the ways in which they control the classroom or teach their students. These considerations are performed in relation to guidelines created by an entity outside of the instructor, and the guidelines must be based on education literature.
Personalistic	Content centered on the relationships present in a learning environment. This includes the relationship between an instructor with their students, an instructor with their peers, and an instructor with themselves.
Critical	Content centered on how societal and cultural phenomena (such as gender, accommodations, and cultural differences) affect the learning environment.

Depth of reflections. Reflective writings have been evaluated for depth through several different categorizations (Day, 1993; Farrell, 2003; Handal & Lauvas, 1987; Jay & Johnson, 2002; Larrivee, 2008a; van Manen, 1977; Zeichner & Liston, 1987). Larrivee (2008a) conducted an extensive review of this work in order to develop a four-level hierarchical model that represents the commonalities across these different categorizations (Table 2). Larrivee's model begins with *pre-reflection* where there is an absence of reflection. At the next level, we have *surface-level reflection* where an instructor is concerned about achieving a specific goal and also acknowledges a link between their actions and the observed outcomes; however, the desired outcomes are only approached through considering pedagogical norms, their own anecdotal experiences, or other practices established within the status-quo (Campoy, 2010; Larrivee, 2008a). In a *pedagogical-level reflection,* an instructor reflects on their educational goals and theories in light of observed

outcomes in student comprehension, recent education research and literature, and alternative viewpoints (Larrivee, 2008a). Finally, *critical-level reflections* consider the ethical, moral, and political ramifications of what is being taught in an educational environment; furthermore, educators are evaluating "their own views, assertions, and assumptions about teaching, with attention paid to how such beliefs impact students" (Larrivee, 2005, 2008a; Machost & Stains, 2023, p. 4). The clear connection between critical-type reflection (re: content; Valli, 1997) and critical-level reflection (re: depth; Larrivee, 2008) should be noted. However, unlike content-based analyses of reflection, depth-based analyses are mutually exclusive. A piece of reflective writing is judged holistically and can only have one associated depth. Thus, while critical-type content is required for the critical-level to be reached, the presence of critical-type content does not automatically indicate a critical-level reflection. Additionally, a piece of reflective writing is associated with a depth, and the individual doing the reflecting is not bound to a particular level of depth; i.e., multiple reflections from an individual may have different associated contents and depths. It is important to note that instances of both pedagogical and critical reflections are considered by the authors to be high-level reflection

Table 2. Depth of reflection based on Larrivee's (2008a) model

Level	Definition
Pre-reflection	Instructors base their teaching practices on preconceived notions, and do not comment about pedagogical goals they attempt to accomplish. There is a lack of connection between an instructor's actions and the observed outcomes.
Surface reflection	Instructors are concerned about achieving a specific goal, such as a specific passing rate for their class. However, these goals are only approached through conforming to departmental norms or their own anecdotal evidence. Thus, they are grounded in personal assumptions and influenced by unexamined beliefs and unconscious biases.
Pedagogical reflection	Instructors are willing to challenge the status quo and alter their pedagogical practices in light of evidence in observed student outcomes, relevant education literature, and alternative viewpoints. In this way, instructors also consider their own pedagogical belief system and its relationship to their practice.
Critical reflection	Instructors consider how societal and cultural phenomena affect the learning environment. In doing so, instructors evaluate their own views, assertions, and assumptions about teaching, with attention paid to how such beliefs impact their students holistically.

Methods

This study, including participant recruitment, was approved by the Institutional Review Board for the Social and Behavioral Sciences at the University of Virginia (Protocol #: 5248).

Reflection scaffold

When conducting a review of the literature and creating a primer on reflective practice for instructors (Machost & Stains, 2023), authors HM and MS created a scaffold for written reflection based on the works of Gibbs (1988), Larrivee (2000, 2008a, 2008b) and Bain et al. (2002); this scaffold was additionally inspired by other reflection scaffolds developed by the University of Edinburgh that were also based on some of this literature (The University of Edinburgh, 2021). This scaffold, which was used to guide participants' reflective writings, begins by prompting participants to self-identify a past challenging teaching situation, i.e. a critical incident. Following

this, participants are asked to describe the facts of the situation before being prompted to describe their feelings and the potential feelings of others involved. Afterwards, they evaluate the past experience for cause-effect relationships and positive/negative aspects and finally draw conclusions from the past experience and plan for future, similar situations. For each step of the process, an example of an answer to the scaffolding question was provided.

Participants

Participants were recruited from two iterations of a national workshop for new physics and astronomy instructors (Physics and Astronomy Faculty Teaching Institute, 2023). Participants represented instructors at a variety of degree granting institutions (i.e., AA/AS, BA/BS, MA/MS, PhD) across all regions of the continental United States.

The first cohort of participants completed a Qualtrics survey containing the previously described scaffold during the workshop held in July 2022. The second cohort of participants completed the survey as a pre-workshop activity in June 2023. This change was implemented as the workshop itself was redesigned to heavily focus on reflection; thus, the pre-workshop survey serves as a baseline for participants' engagement in reflective practices prior to receiving instruction on reflective practice.

Participants were included in the study if they met the following criteria: 1) the reflection submitted had to be about a time or situation when the participants were acting in the role of an instructor; and 2) the description provided by the participants had to be clear and detailed so as to i) be easily understood by the research team and ii) not require interpretation by the research team. Of the 62 instructors who attended the July 2022 workshop, 52 submitted a reflection, and 46 met the inclusion criteria. Of the 106 instructors who attended the June 2023 workshop, 57 submitted a reflection, and 52 met the inclusion criteria. A total of 98 reflections were included for analysis.

Scaffold analysis

A combination of in vivo and a priori coding was used in the creation of the codebook. The codebook itself is comprised of four sub-codebooks containing codes generated to describe the following categories: topics discussed in the reflections, content of the reflections, level of the reflections, and plans created in the reflections. Two code categories, topics and plans, were created solely from *in vivo* coding. It should be noted that none of these *in vivo* codes were mutually exclusive within each code category or across the different code categories. Two code categories, content and level, were created *a priori* from Valli's (1997) and Larrivee's (2008a) descriptions of content and depth, respectively.

Topic codes were used to capture the nature of the critical incident described in the reflections. In all, 18 topic codes were used by authors HM, JMJ, and BJY during coding and assessment of inter-rater reliability; after inter-rater reliability analyses, these 18 topic codes were condensed into 9 parent-categories following analysis by authors HM and MS (Table 3).

The plan codes were used to capture the actions participants either have taken or plan to take to prepare themselves for future, similar situations. In all, 25 plan codes were utilized by authors HM, JMJ, and BJY. Post inter-rater analyses, only the plan codes utilized in at least 5% of the written reflections were retained for further analysis. Authors HM and MS organized these remaining 15 plan codes into three categories based on the intent behind each individual plan.

The portion of the codebook used to describe the content of reflections, as depicted by Valli (1997), was created using a mixture of *a priori* and *in vivo* coding. As other analyses of reflection have done (Minott, 2008), Valli's five categories were utilized in *a priori* coding. However, these five categories were each expanded upon with subcodes derived from *in vivo* coding to give a better understanding of the content described (see Results and Discussion). As with the topics and

plans codes, the content-based codes were not mutually exclusive either across or within code categories.

Finally, the portion of the codebook depicting depth of reflection used *a priori* coding taken from Larrivee's (2008a) description of the different levels of reflection. Larrivee's four-level categorization has previously been used in the analysis of reflections (Ansarin, Farrokhi, & Rahmani, 2015; Campoy, 2010), and a modified version of Larrivee's categorization has also been used (Winchester & Winchester, 2011). However, it should be noted that other analyses use a different depth-model of reflection (Betrabet Gulwadi, 2009; Dyment & O'connell, 2010; Jensen & Joy, 2005; Lee & Abdul Rabu, 2022; O'Connell & Dyment, 2004; Plack et al., 2005; Richardson & Maltby, 1995; Sumsion & Fleet, 1996; Thorpe, 2004; Wong et al., 1995). The categorization used herein is described in the introduction and aligns with Campoy's (2010) and Larrivee's (2008a) works. As the depth of reflection is a holistic analysis, these codes were mutually exclusive. It is important to note that for a reflection to be classified at the critical level, the higher-level concerns (e.g., equity, accessibility, representation, etc.) must have been considered consistently throughout the entirety of the reflection.

Table 3. Topic categories, topic codes, and definitions

Topic category	Topic code	Definition
Poor student(s) behavior	Student in-class disruption	The instructor describes a situation where students are being disruptive (e.g., talking during lecture, arguing before class, etc.)
	Student cheating	The instructor describes a situation where a student cheated on an exam/assignment
	Student procrastination	The instructor describes a situation where a student is procrastinating
External to professor or class	Equipment failure	The instructor describes a situation where the necessary equipment fails during class or lab
	COVID transition	The instructor describes a situation dealing with the COVID transition to online classes or coming back to in person instruction
Class management	Made assessment too difficult	The instructor comments on an assessment or assignment they designed that was too difficult for their students (either due to time constraints or just the complexity of the material)
	Poor class time management	The instructor describes a situation during which they moved too quickly through a class, had too much material expected to be covered in a class, etc.
	Recommendation letter	The instructor reflects on a situation that arose while writing a recommendation letter
Student(s) negative feedback	Student direct negative feedback	The instructor describes a situation where students complain to the instructor directly or via a feedback survey
	Student indirect negative feedback	The instructor describes a situation where students' complained about the instructor/course to others (e.g., colleagues) or via end-of-course evaluations
Critical topics	Sexually inappropriate behavior	The instructor describes a situation where sexual harassment or actions contributing to sexual harassment were taking place in the classroom
	Cultural differences	The instructor describes a situation during which cultural differences contributed to difficulties experienced by either the student or the instructor
	Gender	The instructor describes a situation where gender norms, roles, or expectations play a part in the learning environment. The role of gender may be explicit in the description or assumed by either the student or instructor
Students' weak academic profile	Students lack fundamentals	The instructor describes a situation where students have a weak understanding of fundamental concepts and skills
	Student poor performance	The instructor describes a situation where a student is not performing well academically in class/lab
Struggling student(s)		The instructor describes a situation where a student is not doing well holistically
Student-instructor specific interactions		The instructor describes a difficult student interaction, including combative interaction on the part of the student, correcting students' behaviors in class, or the instructor being abrasive
Instructor's incorrect answer or explanation		The instructor describes a situation where they gave an incorrect answer or explanation or were not able to give any answer or explanation

Trustworthiness

The trustworthiness of this research is of primary concern. As such, steps were taken throughout this analysis to ensure credibility, transferability, and dependability.

Credibility. As outlined by Shenton (2004), there are numerous avenues to demonstrate credibility. First, we approached the analysis by adopting "research methods well established" in the literature (Shenton, 2004, p. 64); *a priori* coding taken from the well-established works of Valli (1997) and Larrivee (2008a) aided in ensuring that the analysis aligns with prior work when determining the content and depth of the reflections. Furthermore, we address the previous findings in the literature while discussing the findings from this study.

Throughout the analysis of the data, frequent debriefing sessions occurred within the entire research team. Finally, we aim to establish transparency of both the data and the data analysis through the information provided in the appendices.

Transferability. We promoted transferability by providing a thick description of the context of the study, its participants, the data collection, and analysis processes.

Dependability. The stability of our findings is primarily addressed via the two different samples, collected one year apart. Similar distributions of content and depth were seen at the two different time points, and the codebook developed after the first data collection readily applied to the second set of data. The initial codebook was created through an iterative code-recode strategy by author HM informed by whole-group discussions with the research team. Additionally, the final codebook demonstrates inter-rater reliability with percent agreements greater than 80% in all code categories; 16 of the 46 reflections from the initial data collection were fully cross coded between HM, JMJ, and BJY to demonstrate reliability (Table 4). Due to the non-mutually exclusive nature of the codebooks, Cohen's kappa values were not calculated.

This cross-coding was performed through stepwise replication across five rounds. Furthermore, throughout the initial sense-making and the intensive inter-rater reliability analyses, a detailed audit trail was kept about the iterative modifications of the code books. Changes made during the first three rounds of inter-rater reliability analyses include: altering the names of codes (e.g., changing 'student indirect complaints/evaluations' topic code to 'student indirect feedback'), adding onto the definitions of codes (e.g. definition of 're-explain course material' planning code was expanded to explicitly include utilizing a different method or approach), and verbal clarifications (e.g. that not all sections of the codebook needed to be utilized in each reflection). The final two rounds of inter-rater reliability analyses resulted in no further changes to the codebook.

Table 4. Inter-rater reliability metrics.
All codebooks were fully cross-coded by HM, JMJ, and BJY.

Codebook	Number of interviews	Percent agreement
Topics	16/46	83%
Content	16/46	89%
Level	16/46	89%
Plans	13/46	87%

Results and Discussion

The findings discussed herein provide insight into the written reflections of physics and astronomy assistant professors who are untrained in reflective practices. The presentation of the results is aligned with the research questions.

Nature of critical incidents

Participants in this study focused their critical incident on nine different topics (Table 3). The top three topics most discussed were *Student(s) weak academic profile*, *Student-instructor*

specific interactions, and *Student(s) negative feedback* (Table 5). The following three excerpts provide examples for each of these three topics, respectively:

> "I was teaching a grad student class. Before the mid-term, nearly the entire class was shaking their hand in agreement when I tried to gauge the clarity of my lectures. I did ask questions and encouraged different people to participate, but the first midterm performance was extremely poor and revealed a knowledge gap that I didn't expect to see." –Instructor 129

> "When I asked a question from a student to increase her engagement in class, she didn't answer. I helped her to get to the answer, but she didn't show any interest either. I provided the answer and asked her to make sure that she understood the process of getting to the answer and she said, "I will just say 'yes'". It was clear her 'yes' was only to make me to leave her alone." –Instructor 114

> "I got very poor course evaluations and students made complaints to the department on grading. However, I asked students to talk to me at the very beginning of the semester if they have questions on their grades and no one talk to me during the semester." –Instructor 138

Overall, the data indicate that most instructors' reflections were focused on negative events with students. Indeed, 79% of the critical incidents contained at least one topic code about negative experiences with students. At the time of the writing of this manuscript, we could not find studies that had explored the focus of teaching reflections written by higher education instructors in STEM and other disciplines. This study thus provides a first insight into what STEM instructors consider challenging situations within their teaching.

Table 5. Distributions of topics discussed in critical incidents

Topic	Proportion of reflections
Student(s) weak academic profile	39%
Student-instructor specific interactions	28%
Student(s) negative feedback	26%
Instructor's incorrect answer or explanation	17%
Poor student(s) behavior	14%
Class management	8%
Critical topics	7%
Struggling student(s)	6%
External to professor or class	6%

Content of reflections

We leveraged Valli's (1997) framework to analyze the content of the reflections, which includes five types (Table 1): in- and on-action, deliberative, technical, personalistic, and critical. Since the scaffold used to guide the written reflections requires the participants to reflect on a past teaching experience, the in- and on- action content type was not relevant to code.

Neither technical nor deliberative content were present in any of the participants' reflections. The lack of technical content aligns with a prior study investigating pre-service teachers enrolled in a course that required students to maintain a reflective journal throughout the term (Minott, 2008). In this study, reflections were collected from 20 pre-service teachers where participants submitted five entries from their reflective journals for assessment. In these submissions, there were no instances of technical content, mirroring the findings from the present study. Importantly, Minott's participants had months to record reflections in a journal and chose which of their reflections to submit. Our study collected spontaneous reflections from participants who were without previous training in reflective practices. Thus, the lack of technical reflection in either participant pool may indicate that technical reflection needs to be deliberately prompted. Unlike what is observed in our study, Minott (2008) noted instances of deliberative content in 10%

of the study sample. This difference may be due to the scaffold used in our study (see Appendix A.1), which does not directly probe instructors to consider opposing perspectives or viewpoints.

Personalistic content was the most common content present in the reflections with 57% (*n* = 56) of instructors addressing it. The prevalence of personalistic content aligns with Minott's (2008) prior study, as personalistic content was the second-most prevalent content type among Minott's participants, only surpassed by in- and on-action. We identified six subcodes that fit within personalistic content (Table 6). Our participants reflected mostly on themselves and their flaws or on negative perceptions that they thought others had about them. Few considered their students' holistic improvement or empathized with them, two key criteria for personalistic content (Machost & Stains, 2023; Minott, 2008; Valli, 1997).

Table 6. Personalistic content subcodes, definitions, exemplary quotes, and distribution of subcodes within the reflections that contained personalistic content

Personalistic content subcode	Definition	Exemplary quote	Proportion of personalistic content reflections
Failure to facilitate learning	Instructor reflects on their inability to facilitate their students' learning	**"**I felt I failed the students on properly introducing them to a key concept in the course and felt like I was not a good teacher." –Instructor 257	54%
Perceived negative opinions of instructor by students	Instructor reflects on how they perceive their students to view them or the course	**"**I immediately felt a sense of dread and panic - thinking that my students would think I was a fraud." –Instructor 209	36%
Negative personal traits	Instructor reflects on their own negative personal traits (short temper, insecurities, etc.)	**"**I think it also reflected my own insecurities. I always had a bit of imposter syndrome, especially in grad school, so any sort of criticism of my teaching made me very defensive." –Instructor 134	16%
Failure as advocate	Instructor reflects on their inability to advocate for their students or their failure while advocating for them	"No one had prepared me for what I should do when a student starts having a breakdown/crisis in the middle of class. After the student left, I was mostly concerned that the student would be able to get help. I hope the student felt supported." –Instructor 249	12%
Peer interpretations or opinions of instructor	Instructor reflects on how they perceive their peers to view them or the course	**"**My colleague observing me definitely pitied me and tried to offer helpful suggestions." –Instructor 233	9%
Student personal struggles	Student is having difficulties due to situations/context outside of class/lab	**"**After the student described their situation, and how much they were working, on top of taking so many classes, I felt surprised and empathetic." –Instructor 247	2%

Critical content was observed in significantly fewer reflections (12%, n = 12) and felt into one of four subcodes: 1) Accommodations, 2) Gender, 3) Cultural differences, and 4) Grouping (Table 7). The presence of critical content in a minority of participants again aligns with Minott's (2008) findings, who noted critical content in only 3% of the reflections in their study. The most common critical content written about by our participants related to the need to accommodate students.

Table 7. Critical content subcodes, definitions, exemplary quotes, and distribution of subcodes within the reflections that contained critical content

Critical content subcode	Definition	Exemplary quotes	Proportion of critical content reflections
Accommodations	Accommodations for students with disabilities and/or difficult situations. May relate either to the implementation of the accommodations themselves or to how accommodations affect non-accommodated students.	**"**Some students, especially those with a family, might be too busy to do two homework assignments per week and to commit to outside class activities. That gave me a new understanding to accommodate everyone in class." –Instructor 204	58%
Gender	Gender adding a more complex layer to situations. Most often, it is the result of a female existing in a male-dominated field or class. The effects of this gender discrepancy can be expressed as either experienced by the student or be expressed as a concern by the professors navigating a situation	**"**A student neglected to write their pronouns in my get-to-know-you survey and asked me not to bring up their gender again. I was happy to respect that, but then they asked me to write them a letter of rec. I needed to write their pronouns in the letter, and I didn't know what to use." – Instructor 101	17%
Cultural differences	Cultural differences contributing to difficulties experienced by either the student or the instructor	**"**He was previously educated in another country, where the students were not able to ask questions (as that generally was viewed as meaning they weren't able to do things themselves). So I realized that when I told the class that I expected them to come talk to me about things they didn't understand, he still didn't think it was really an option." –Instructor 147	17%
Grouping	Grouping students together without reason (note: NOT due to race, gender, ethnicity, sexuality)	**"**I think things went poorly because I painted half of the class with a generalization. I realized afterwards that it would be better to address students' resistance to participation when they were in smaller groups, or perhaps individually." –Instructor 112	17%

Depth of reflections

The depth of the reflections collected were analyzed using the four-level hierarchical categorization of reflections developed by Larrivee (Table 2; 2008a). Over 80% of the reflections written by our participants felt to the low-level of reflection with 23 reflections classified at the pre-reflection level and 59 at the surface-level (Fig. 4). A hallmark of pre-reflection was a lack of

connection between an instructor's actions and words and the observed outcome. This is exemplified with Instructor 108:

> "There was a girl in the class who did very well in almost all the homework. She never came to the office hour. But since she did well in homework, I thought she understood the materials well. But she didn't do well in mid-term. I discussed the midterm with her, and she told me that she had schedule conflict with the office hour. I then offered very flexible time to her, but she then never came. She continues to do okay on her homework until she did very poorly on the final ... I feel confused about her performance on homework and exam. It seemed that she was cheating on her homework." –Instructor 108

Instructor 108 saw themselves as a bystander; they failed to see any reason for the conflicting performance of their student other than cheating. Additionally, there is no connection between a minimal action on the instructor's part and the student's continued mixed performance. This contrasts to surface-level reflection where instructors do make a connection between themselves and the outcomes; however, the plans to achieve different outcomes in surface-level reflections are based on anecdotal experiences or the status quo as Instructor 102 illustrates:

> "I learned that I need to be more prepared for my lectures, although this is an ongoing challenge for me. I do need to learn to handle my own mistakes with more grace. I'm OK with admitting that I'm wrong or don't know something, but I do that too much in my lectures." –Instructor 102

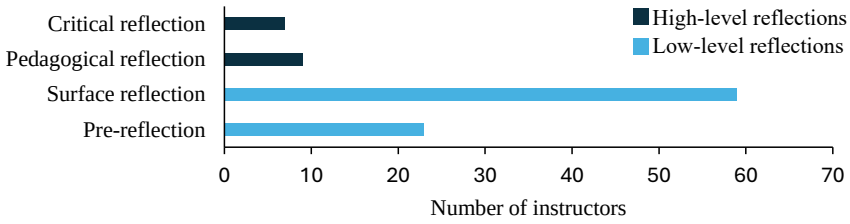

Figure 4. Distribution of physics and astronomy instructors' reflections across the different levels of depth of reflection based on Larrivee's (2008a) model.

A minority of participating instructors (16%) completed high-level reflection (Fig. 4): pedagogical-level reflection (n = 9); critical-level reflection (n = 7). As seen with Instructor 128, instructors who reached the pedagogical level focused on how they can improve their teaching based on observed outcomes in student comprehension, alternative viewpoints, and/or current educational research and literature:

> "I learned that my style of sort of more casual research instruction ... does not always help my student. I think I should learn more about teaching scientific programming to undergraduates, and what are some successful strategies or techniques I can impart to them. Hopefully in the future I'll be better prepared because I will have developed structured mini-lessons on best coding practices, and my student and I will have done those together, and I will have also developed mechanisms for soliciting specific feedback from my students on what I can do to help them better learn." –Instructor 128

Instructor 128 took from their experience that they need to change the status quo of how they taught coding to researchers. In doing so, they exhibit a high-level of reflection regarding their pedagogical practices. An added layer of complexity is present in those instructors who reach critical-level reflection as they examine the role that larger societal issues, trends, and differences play in learning environments. Instructor 147 details this relationship, as they had a student who

had the potential to perform better in their course but did not do so because of cultural differences where the student was not comfortable asking questions. Furthermore, rather than problematizing the student, Instructor 147 acknowledges that it is their role as the instructor to make the classroom norms easily understood by students.

> "I was working under the assumption that when I told students that not only could they ask questions and/or come to me for help, [that] they accepted it when I made the offer. This situation made me understand that some students (especially from certain backgrounds) had preconceived notions about what they should do as students, and that I needed to do more to encourage them." –Instructor 147

These findings may seem in contrast to a prior study investigating Iranian English as a Foreign Language teachers which found the predominant depth of reflection among these instructors to be at the pedagogical level. Importantly, the researchers found a positive correlation between an instructor's years of teaching experience and the depth of their reflection (Ansarin, Farrokhi, & Rahmani, 2015). In their study, instructors had a broad range of teaching experience with an average of 8.39 ± 4.59 years. In contrast, our study sample had less teaching experience; based on the demographic data that were collected from the 2023 cohort (no such data was collected from the 2022 cohort), the 2023 cohort had an average of 3.2 ± 4.8 years of teaching experience. Therefore, our cohort is more similar to the group of instructors in the Ansarin et al. (2015) study who were classified in the low level of teaching experience. That group wrote a significantly larger proportion of pre-reflection and significantly less pedagogical and critical reflections. In light of these results, it may be that our sample provided fewer high-level reflections because they did not have enough teaching experiences.

Plans to address similar situations in the future

Instructors were asked to describe, based on what they learned from the experience, their plans for preparing themselves better when faced with a similar challenging situation in the future. Through in vivo coding, three majors plan codes emerged (Table 8): *Self-preserving*, *Self-reliant*, and *Seeking knowledge* outside of self. *Self-preserving* plans entail emotional regulation regarding either oneself or others (i.e., personal grace), or standard practices of instructors (i.e., pre-planning). *Self-reliant* plans go beyond the explicit duties of instructors and are based solely on an instructor's own experiences, speculations, and abilities to address the topics at hand (e.g., establishing clear expectations in the classroom, correcting mistakes made by oneself, meeting students where they are academically, discussing issues privately or in small groups). *Seeking knowledge* plans rely on an instructor going outside of their current knowledge or personal past experiences, and include soliciting student feedback, implementing successful strategies (either in the literature or as used by peers), communicating with peers, and participating in professional development. As Table 8 indicates, instructors' plans relied mostly on the instructors themselves and their knowledge and experiences. Only about a quarter thought to reach out and leverage other resources (e.g., peers, books, peer-reviewed journal articles) to better equip themselves to handle future challenging situations. The nature of the plans presented in these reflections indicate that the engagement in the reflection is unlikely to lead to pedagogical growth among the participants.

Table 8. Types of plans described in the reflections for managing future similarly challenging situations.

Only plan subcodes present in at least 5% of the reflections were analyzed and are presented in this table. For a full list of planning subcodes and definitions, see Appendix B.

Types of plan	Example subcode	Exemplary quote(s)	Proportion of reflections
Self-reliant	Communicate with students – establish clear expectations	"I learn from this situation that it is super important to set up classroom culture and be more attentive to class dynamics. I think in the future, I could integrate the activity of building community agreement at the beginning of the course and continue to revisit it throughout the semester to remind students of ways of working that they are expected to do and agreed to do" –Instructor 125	58%
	Re-explain course material	"I would step back and engage with some basic cross-product concepts that students know from mathematics, and then, once they get familiar, they move on to the idea they need to know." – Instructor 237	
Self-preserving	Pre-planning	"I've learned that preparation is important to avoid stressful and regretful situations (for both the instructor and the students)." – Instructor 232	45%
	Personal grace	"But I also realized I needed to give myself grace for not taking more frequent/detailed notes the first time I was teaching as it was ridiculously busy semester." – Instructor 236	
Seeking knowledge	Participate in professional development	"I have had to develop leadership skills to address interpersonal issues more directly, including attending workshops and taking courses on equity, inclusion, and social justice." – Instructor 121	27%
	Communicate with peers	"I am always looking for advice from teachers with experience running extra large classrooms." – Instructor 250	
No plan	Did not write a plan	"Students sometimes seem to assume (based on their predicted grades in the teaching evaluations) that I will curve more than I do. I worry this affects the amount of work they put into the class." – Instructor 216	9%
Other types of plan	Plans that were present in 5% or less of the reflections	"I don't know what to do under this situation." – Instructor 108 "I am unlikely to address questions that aren't strictly about content at my university ever again, which I think is a loss for both the students and for me." – Instructor 127	5%

Relationship between the nature of the critical incident and depth of reflections

We analyzed the relationship between the nature of the critical incident (i.e., topics; Table 3) and the depth of the reflection to explore whether certain situations are more prone to engage instructors in higher-level reflections. Table 9 displays the distribution of the topics explored in the critical incidents across the four levels of depth of reflection described by Larrivee (2008a; Table 2).

The topic of *Student(s) weak academic profile*, which was the most common topic discussed by our participants (Table 3) is equally represented across all levels of depth. Therefore, reflecting on students' academic difficulties can but does not necessarily lead to high-level reflections.

The topics that most distinctively separate reflections at the critical level from other levels were *Student-instructor specific interactions* and *Struggling student(s)*. The *Student-instructor specific interactions* were over twice as prevalent in the critical reflections than in the other levels of reflection. However, no notable qualitative differences were found between the descriptions of *Student-instructor specific interactions* at the critical level and lower levels. Therefore, similarly to the *Student(s) weak academic profile* topic, the focus on student-instructor interactions does not seem to drive the depth of the reflection. The *Struggling student(s)* topic, which is when instructors are considering their students who appear to be struggling holistically rather than solely as students or academically, was only present in 7 of the 98 reflections, but half of these reflection were at the critical level. The presence of this topic is in alignment with the definition of critical level by Larrivee (2008a). However, it is worth noticing that few of the lower-level reflections covered this topic as well. While these instructors had described students struggling holistically, they did not make it the focus of their reflections and were thus not classified in the higher-level of reflections.

This points to a missed opportunity for instructors to engage in more transformative reflections but also indicates that instructors need to be guided towards unpacking more this type of topics.

Overall, the data in Table 9 do not provide a clear trend (except for *Struggling student(s)*) between the topic being discussed in the critical incident and the depth of the reflection. This finding indicates that it may not be necessary to coach faculty to think about particular types of situations in order for them to engage in high-level reflections. Other aspects, such as the content of the reflection, might play a bigger role and will be explored in the next section.

Table 9. Distribution of topics discussed across the four levels of depth.

Cell percentages represent the proportion of reflections at a specific level (i.e., depth) of reflection that included each topic category. Topic categories are not mutually exclusive; thus, the sum within a level is greater than 100%.

Topic (from most to least reported)	**Pre-reflection (*n* = 23)**	**Surface (*n* = 59)**	**Pedagogical (*n* = 9)**	**Critical (*n* = 7)**
Student(s) weak academic profile	43%	36%	44%	43%
Student-instructor specific interactions	30%	25%	---	71%
Student(s) negative feedback	22%	30%	33%	---
Instructor's incorrect answer or explanation	9%	24%	11%	---
Poor student(s) behavior	26%	12%	11%	---
Class management	---	10%	22%	---
Critical topics	9%	7%	11%	---
Struggling student(s)	9%	2%	---	43%
External to professor or class	4%	5%	11%	14%

Relationship between the content and depth of reflections

While the connection between content and depth of reflection may appear to be intuitive, few studies simultaneously analyze reflections for both content and depth (e.g., Lee, 2005). This is an important gap in the literature as understanding the content that appears in high-level reflections can aid in the development of reflective practitioners.

Fig. 5 depicts the relationship between the content and depth of reflections. As the level of reflection increases so does the presence of personalistic content. At the pre-reflection level, most reflections contain neither personalistic nor critical content, while all critical reflections contain both personalistic and critical content. Interestingly, critical content is a distinctive feature of reflection at the critical level since it is mostly absent in the pre-reflection, surface, and pedagogical reflections. Therefore, it is essential to guide instructors towards exploring critical content (e.g., gender, accommodations, and cultural differences) when they engage in reflection. However, as the presence of critical content in the low-level reflections indicates, it might not be sufficient. Similar to our previous recommendations about guiding instructors to further unpack the topic of *Struggling student(s)*, instructors also need to be guided in exploring critical content for them to reach higher level reflections.

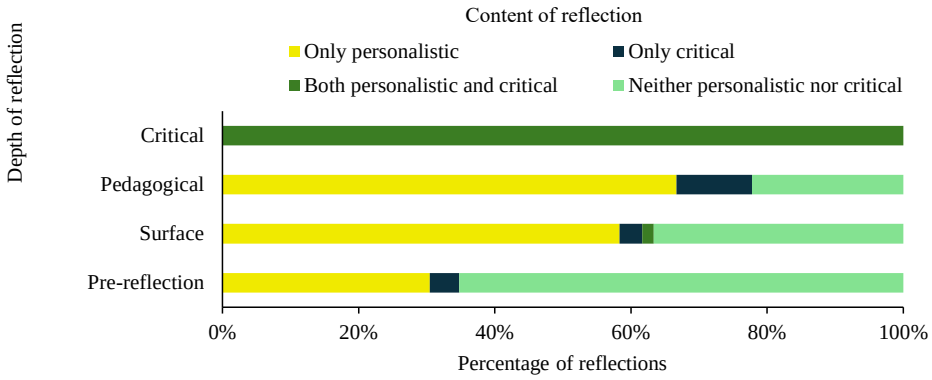

Figure 5. Overlay of content and depth of instructors' reflective writings.
Percentages are normalized for each level of reflection.

Relationship between the plans outlined and depth of reflections

Each type of plan (i.e., *Self-reliant*, *Self-preserving*, and *Seeking knowledge*) was observed across all depth levels (Table 10), but each level of reflection had a different combination of plans (Fig. 6).

Low-level reflections contained more diverse plans and were more likely to have a combination of plan types when compared to high-level reflections. However, low-level reflections were also the only reflections for which the *No plan* code was used, albeit at a small rate (Fig. 6). The most common type of plans in each of the low-level reflections was *Self-reliant* (Table 10). Reflections in both the pre-reflection and surface reflection levels also had roughly a quarter of the plans focused on *Seeking knowledge*. What clearly differentiated the two low-levels of reflection was the proportion of *Self-preserving* plans, which was higher in the surface reflections when compared to the pre-reflections.

High-level reflections had limited types of plans and were dominated by *Self-reliant* plans (Table 10). Nearly half of the reflections at the high reflection levels also included *Self-preserving* plans. A key distinction between the pedagogical and critical levels was the much higher proportion of *Seeking knowledge* plan in the pedagogical reflections (67% versus 14%, respectively). Indeed, the critical level had the smallest proportion of reflections with *Seeking knowledge* plans (14%); this could be due to the difficult subject matters broached in the critical-level reflections which instructors may be hesitant to discuss with outside sources.

Overall, the data show that regardless of the level of reflections, instructors rely on themselves to prepare for the next time they face a similar critical incident. Therefore, instructors' engagement in these reflections are not likely to result in pedagogical growth. Our data indicates that we need to normalize seeking help from others when facing challenging teaching situations.

A recent study that qualitatively explored the teaching social network of STEM faculty had probed help-seeking behaviors of STEM instructors when faced with issues with their teaching (Lane et al., 2022). They found that many of the 19 interviewees would only reach out to their discussion partner if they knew that this instructor had the expertise and experience that was directly related to the problem they were encountering. This current study and the Lane et al. study (2022) demonstrate the need to promote communications among instructors so that they can learn about the breadth of expertise of their peers, and thus have resources that they can feel comfortable reaching out to when facing a challenging situation.

Table 10. Distribution of plans described by instructors across the four levels.

Cell percentages represent the proportion of reflections at a specific level of reflection (i.e., depth) that included each type of plan. As instructors could describe multiple types of plan within the same reflection, the sum within a level is greater than 100%.

Type of plan (from most to least reported)	Pre-reflection (n=23)	Surface (n=59)	Pedagogical (n=9)	Critical (n=7)
Self-reliant	43%	56%	89%	86%
Self-preserving	17%	56%	44%	43%
Seeking knowledge	30%	24%	67%	14%
No plan	17%	8%	---	---
Other types of plan	13%	3%	---	---

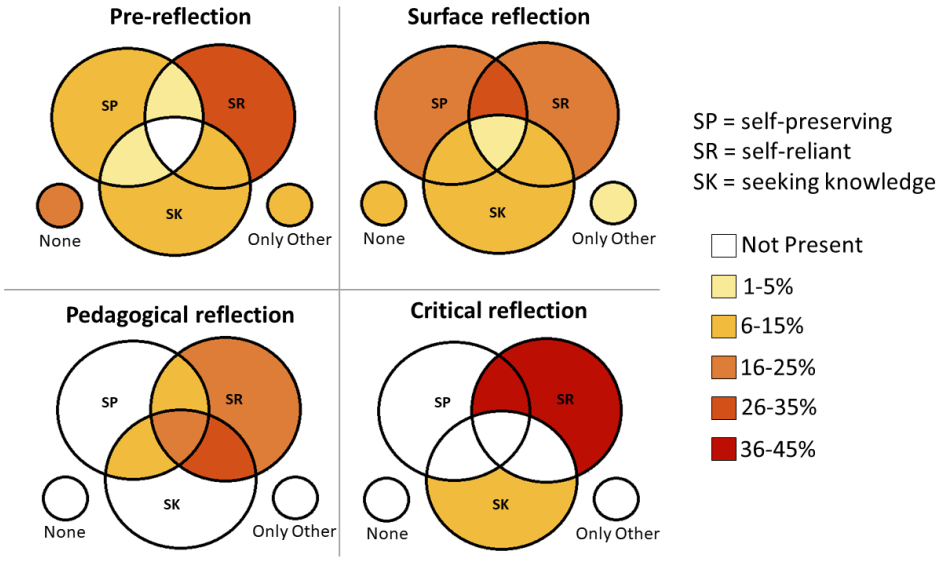

Figure 6. Combinations of plans among the four levels of reflection

Implications

Findings from this study lead to several implications regarding the promise of reflective practice in promoting pedagogical growth and the research agenda around reflective practice.

The required inclusion of reflections on teaching evaluations is likely not enough to promote pedagogical growth: instructors need to be trained on reflective practice.

This study showed that instructors with limited teaching experience wrote low-level reflections. Low-level reflections mean that instructors are not considering their beliefs and values about teaching, nor educational literature when reflecting on a critical incident. Our data also show that instructors are primarily looking inward when elaborating plans to address future similar situations. As Henderson et al. (2011) remarked in their review of the literature on instructional change, it is essential for instructors to face their beliefs/values around teaching in order to better problematize their teaching. Moreover, their self-reliance is unlikely to lead these instructors

towards learning new instructional approaches or ways of supporting their students. Consequently, the instructors in this study are unlikely to experience pedagogical growth as a result of their writing of these reflections.

As indicated in the introduction, reflections are becoming a center-piece of new teaching evaluations and are seen as a mean to help instructor improve their teaching practices (Simonson, Earl, & Frary, 2022; The University of Kansas Center for Teaching Excellence, 2024). Our data suggest that this requirement alone is insufficient to achieve this goal and that training instructors is necessary. This is also in-line with prior research on reflective practice (Belvis et al., 2013; Dinham et al., 2021; Zahid & Khanam, 2019). Our data points to the need to train instructors in recognizing and unpacking critical topics and in considering students more holistically. Trainings should also provide instructors with educational resources and trusted networks of pedagogically-trained colleagues that they can leverage to gain insight about their particular situation and identify strategy to mediate similar future situations.

A more extensive research agenda around reflective practice in STEM instructors is needed to design effective training.

This study is one of the first studies to characterize the nature of STEM instructors' reflections on teaching. Consequently, more studies ought to be conducted to characterize the generalizability of these results across STEM fields (we only have physics and astronomy instructors in this study) as well as a range of teaching experiences and contexts (e.g., type of course, class size, type of institution). Extending this research agenda is essential to assist institutions and teaching and learning centers in the development of training programs that cater to the need of the different types of populations of instructors.

Limitations

The exploratory nature of this study limits the generalizability of the results. Indeed, the sample size is small and only represents a particular slice of the STEM teaching professorate (i.e., physics and astronomy assistant professors). Thus, extrapolation to other STEM and non-STEM disciplines is not supported. Moreover, the participants in this study voluntarily chose to attend this pedagogical-focused workshop. Consequently, they may not represent typical new instructors in physics and astronomy. Finally, as reflective practice is inherently personal, it is possible that participants were not inclined to write about critical scenarios or to include controversial topics despite the confidential nature of this study.

Conclusion

This study is one of the first to provide an insight into the nature of STEM instructors' reflections on their teaching. The results show that physics and astronomy instructors with limited teaching experience are mostly unable to write reflections at a level that would promote pedagogical growth. This study thus points to the need to support and train STEM instructors on their reflective practices, especially if the intent of the inclusion of reflections in teaching evaluation processes is to promote instructional transformation.

Chapter 2. Uncovering the complexity of emotions experienced by physics faculty when reflecting on a past teaching experience: Shining a light on an often-overlooked aspect of post-secondary instruction.

This chapter is adapted from a soon-to-be-submitted manuscript.

Introduction

Preparing the next generation of mathematicians, scientists, and engineers is of undeniable importance. In 2012, the President's Council of Advisors on Science and Technology (PCAST) highlighted a critical need for a million more STEM professionals (Olson & Riordan, 2012). However, meeting this need is challenging as only 40-50% of students who begin a STEM degree program go on to receive a STEM degree (Seymour et al., 2019). Students have often cited poor instructional practices provided in the first and second years of STEM courses as reasons they left their STEM majors; while this issue has been known for over twenty years, it still persists (Seymour & Hewitt, 1997; Seymour et al., 2019). Research has previously been conducted to address student retention in STEM including the development, testing, and dissemination of EBIPs (Freeman et al., 2014; Lorenzo, Crouch, & Mazur, 2006; Ruiz-Primo et al., 2011; Theobald et al., 2020). In addition to promoting EBIPS, discipline-based education researchers have explored contextual factors influencing their adoption. While contextual factors – such as course size, classroom layout, and course level – are well studied (Henderson & Dancy, 2007; Hora & Anderson, 2012; Lund & Stains, 2015; Michael, 2007; Prosser & Trigwell, 1997; Shadle, Marker, & Earl, 2017; Sturtevant & Wheeler, 2019; Yik et al., 2022a, 2022b), the individual factors which can affect instructor adoption should not be ignored (Kraft et al., 2024). These individual factors are known to include instructors' familiarity with EBIPs (Kraft et al., 2024; Lund & Stains, 2015; Oleson & Hora, 2014; Yik et al., 2022a, 2022b), instructors' pedagogical beliefs (Popova & Jones,

2021; Popova et al., 2020), and the emotions that instructors experience while teaching (Cubukcu, 2013; Frenzel, Daniels, & Burić, 2021; Geng & Yu, 2024; Mattanah et al., 2024; Mendzheritskaya & Hansen, 2019; Trigwell, 2012). Notably, emotions are well-documented to affect individuals' decision-making (Kordts-Freudinger, 2017; Mattanah et al., 2024; Postareff & Lindblom-Ylänne, 2011; Trigwell, 2012). Furthermore, differences in emotive cultural norms are known to alter how emotions correlate to pedagogical decisions in different countries (Kordts-Freudinger, 2017; Mattanah et al., 2024; Mendzheritskaya & Hansen, 2013; Trigwell, 2012). Thus, understanding the emotions involved in teaching can prove invaluable to the promotion of EBIPs and can enable targeted interventions focusing on emotional health of instructors. However, within STEM, there is currently a lack of understanding of what emotions instructors experience while teaching in the classroom. Given the relationship between emotion and choices of pedagogical approaches, there is a need to characterize the emotions STEM instructors' report feeling while teaching.

What are emotions?

Emotions have been defined in many ways in the literature. Scherer (2005) has argued that emotions are episodes of synchronized changes in most, if not all, of a person's systems (i.e., cognitive, neurophysiological, motivational, motor expression, and subjective feeling components) as they evaluate a stimulus. Emotions are typically short, high intensity responses that change quickly as one re-evaluates what they are experiencing; correspondingly, emotions can impact actions (Scherer, 2005). This contrasts with definition of other emotion terms, such as moods, which tend to be lower-intensity and can be experienced for longer durations (Scherer, 2005). Thus, people typically recall an event that triggered an emotion more reliably as opposed to what triggered a mood (Schutz et al., 2006). Frenzel et al. (2021) argue that instructors' emotions align with the Sherer's definition as instructors are constantly evaluating different social

interactions (e.g., engaging with students) and their response to those interactions (Frenzel, Daniels, & Burić, 2021).

Due to the multiple systems involved when experiencing emotions, research which identifies emotions often uses mechanisms for tracking facial expressions and for measuring physiological responses, such as heart rate (Ekman, 1992; Scherer, 2005). However, with the growing usage of social media and other online platforms to communicate emotions via text, there is an emergent need to characterize emotions from written words alone.

When characterizing emotions from written text, emotion analysis has centered on characterizing the sentiment of reflections, meaning the overall affect is characterized as positive, negative, and in some instances, neutral (Mohammad & Turney, 2013; Nandwani & Verma, 2021). Sentiment analysis has been used previously to understand the sentiments expressed by customers in product reviews (Hu & Liu, 2004), sentiments from tweets concerning politics (Onyenwe & et al., 2020), and the sentiments of movie reviews (Crossley, Kyle, & McNamara, 2017). More recently, researchers have been trying to detect specific emotions in written text to characterize a person's specific emotions (i.e., joy or anger), a process called emotion detection (Nandwani & Verma, 2021). There are many different models and tools that are used for emotion detection, but the number of specific emotions and emotion categories can vary (Nandwani & Verma, 2021). However, despite this variance, the determination of emotions expressed in text can provide a more nuanced analysis.

Why is it important to characterize instructors' emotions?

Instructors in the classroom can experience a wide variety of emotions (Cubukcu, 2013). These emotions can be influenced by their attitudes and predispositions, but also their interactions with their students (Frenzel, Daniels, & Burić, 2021; Trigwell, 2012). Characterizing instructors'

emotions is essential, because instructors serve a pivotal role in helping students' engagement with learning and can influence a student's decision to remain in their chosen discipline (Mattanah et al., 2024; Seymour & Hewitt, 1997; Seymour et al., 2019). The link between emotions and instructional decision-making stems from how an individual evaluates a stimulus event (e.g., students being disruptive during class), and how they subsequently act in response to that event (e.g., instructing the students to focus) (Ellsworth & Scherer, 2002; Scherer, 2005). The association between instructors' emotions and their teaching practice has been explored more in depth in the K-12 literature than in post-secondary education. The K-12 literature highlights that positive teaching emotions allow instructors to generate more varied instructional strategies, which can improve learning outcomes (Sutton & Wheatley, 2003). This linkage is proposed as positive emotions are known to promote cognitive flexibility, enabling instructors to adopt different and new teaching strategies (Jiang, 2021; Mattanah et al., 2024).

A similar association was found within higher education, although this topic is overall understudied (Trigwell, 2012). Trigwell (2012) explored the emotions that five hundred instructors at an Australian university experienced and their approaches to teaching (Trigwell, 2012). The results showed that there was a relationship between instructors' emotions and their instructional decisions, where instructors who had more positive emotions (e.g., happiness and pride) used more student-centered teaching approaches (Trigwell, 2012). A similar finding and relationship was observed in a study conducted in Germany among their university instructors; experiencing positive emotions was associated with the adoption of student-centered instructional strategies (Kordts-Freudinger, 2017). Within the United States, Mattanah *et al.* (2024) conducted a similar study and explored one hundred and forty-one instructors from two institutions. Their work revealed an association between instructors' emotional experiences and whether their instructional

approaches were primarily instructor-centered or student-centered (Mattanah et al., 2024). The results again showed that instructors who reported feeling more positive emotions in the classroom (i.e., satisfaction) also reported using more student-centered instructional practices (Mattanah et al., 2024).

The effect of instructors experiencing negative emotions is more nuanced due to a cultural or environmental dependence on the relationship between negative emotions in the classroom and the adoption of student-centered strategies (Mattanah et al., 2024). Studies conducted in Australia and the United States found that instructors who reported experiencing negative teaching emotions (e.g., anger and disappointment) also adopted more teacher-centered strategies (Mattanah et al., 2024; Trigwell, 2012). However, other studies conducted in Germany and Russia found no relationship between instructors negative emotions and their adoption of teacher-centered strategies (Kordts-Freudinger, 2017; Mendzheritskaya & Hansen, 2013). This difference is thought to be caused by cultural norms around emotional regulation. Emotional regulation is defined as an individual controlling the emotions they experience as well as when and how they experience these emotions (Gross, 1998). Instructors in some countries, such as Australia and the United States, are encouraged to regulate their emotions in the classroom and to display only positive teaching emotions (Hagenauer, Gläser-Zikuda, & Volet, 2016; Mattanah et al., 2024; Shapiro, 2010). Instructors in other countries, such as Germany and Russia, are less likely to mask negative emotions, and instead share their feelings with their students (Hagenauer, Gläser-Zikuda, & Volet, 2016; Mendzheritskaya & Hansen, 2013). This assertion is supported by a qualitative study which conducted interviews among university instructors in Germany and Australia to explore how they display positive and negative emotions while teaching (Hagenauer, Gläser-Zikuda, & Volet, 2016). The results indicated that instructors from both countries viewed being open about positive

emotions as important to teaching; however, German instructors also reported being more open in displaying negative emotions as well (Hagenauer, Gläser-Zikuda, & Volet, 2016).

Instructors' emotions being associated with instructional decision-making highlights a need to understand what specific emotions instructors are feeling while teaching. Additionally, it is necessary to understand rationale instructors link to their emotional experiences. Within higher education, Cubukcu explored 10 language instructors' emotions that they showed in their class and their rationale for why they felt these emotions (Cubukcu, 2013). The results highlighted that instructors typically felt anger or frustration as well as joy in relation to students' actions (Cubukcu, 2013). Younger university instructors were shown to describe feeling more guilty due to not feeling able to answer students' questions (Cubukcu, 2013). More recently, a study in higher education investigated instructors' sentiment while teaching a writing class from non-native speakers (Geng & Yu, 2024). The results showed that an instructor in the classroom experienced more negative than positive sentiments, and the rationales for these sentiments were due to students' actions, such as students' low participation in class (Geng & Yu, 2024). The positive sentiments that were described by the instructors were also linked to students' actions, such as students' commitment to writing and revising their writing (Geng & Yu, 2024). Within STEM higher education, few studies have focused on instructors' emotions while teaching. What work has been done regarding STEM instructors' emotions focuses on their emotions in relation to their professional identities (Jiang et al., 2021), engineering instructors' emotions while teaching during the COVID-19 pandemic (Rehmat, Diefes-Dux, & Panther, 2021), and physics instructors' sentiment regarding transitioning to online learning during the COVID-19 pandemic (Green et al., 2024). Recently, there have been calls to focus on students' and instructors' emotions in STEM education environments (Lönngren et al., 2024; Tea, 2024).

Cumulatively, prior studies on emotions in the education settings suggest that there is a need to support instructors and destigmatize feeling and expressing a range of emotions in the classroom in environments where instructors implement more emotional regulation and display only positive emotions. Therefore, instructors in the United States are likely in need of this targeted support in order to facilitate their pedagogical growth. However, there is a current lack of understanding of the types of emotions American STEM instructors are feeling in the classroom, and why these instructors are feeling negative or positive emotions in the classroom. An understanding of both aspects is necessary to provide appropriate support to instructors.

What tools have been used to characterize emotions?

There are many different approaches that researchers have taken to characterize emotions, each with their strengths and weaknesses.

A smaller emotion categorization model is Ekman's universal six basic emotions which is a theory-driven categorization containing six emotions (anger, disgust, fear, happiness, sadness, and surprise) which are posited to be universal across the human population (Ekman, 1971, 1992). Ekman's six basic emotions model is one of the most widely used in emotion recognition research (Maruf et al., 2024). With there only being six emotion categories with this model, the classification scheme is simple and allows for higher agreement between researchers when identifying emotions in text (Williams et al., 2019). However, some limitations to having only six emotion categories are an oversimplified representation of emotions and misrepresentations of emotions in order to accommodate the six categories (Williams et al., 2019). Additionally, out of these six emotion categories, happiness is the only positive emotion; thus, all positive emotions would be forced into this category (Williams et al., 2019).

Another commonly used emotion model is the wheel of emotion (Plutchik, 1980). The wheel of emotion depicts at the center eight basic emotions (i.e., anticipation, joy, trust, fear, surprise, sadness, disgust, and anger) (Plutchik, 1980). Additionally, the wheel has both a lower and higher intensity word for each of the basic emotions (Plutchik, 1980). For example, for the basic emotion joy, the lower intensity is serenity, and the higher intensity is ecstasy (Plutchik, 1980). In total, the wheel of emotion allows for 24 emotion categories. With having more emotion categories, there are more options; thus, less oversimplification occurs (Williams et al., 2019). Additionally, the wheel of emotions has shown to have good agreement between researchers categorizing emotions (Williams et al., 2019). However, an evaluation of Plutchik's wheel of emotions found that the basic emotions are not readily distinguishable from the additional categries (Smith & Schneider, 2009).

To address the limitations of these two broadly applied models, researchers created an additional emotion model that categorized emotions from participants' free-response data (i.e., participants' written description of their emotions) (Scherer, 2005). While people may not know how to explicitly describe the emotions they recall feeling, the free-response format allows for individuals to describe their emotions more accurately (Scherer, 2005). To create a systematic tool to categorize emotions from free-response data, the Geneva Affect Label Coder (GALC) was developed (Scherer, 2005). GALC has 36 semantic emotion categories which were built from both empirical data and lexicons (Scherer, 2005). GALC functions by searching the written-free response for indexed emotion terms that are present; it then calculates the frequency of each emotion category. An added benefit of this tool is that it does not require manual coding of emotions, thus limiting human bias in the interpretation of others' subjective emotional experiences.

More recently, natural language processing tools (NLP) have become increasingly more common to the analysis of emotions. These are automated tools to characterize emotions and perform sentiment analysis from written text (Nandwani & Verma, 2021). Examples of these NLPs are the Linguistic Inquiry and Word Count (LIWC) and Sentiment Analysis and Social Cognition Engine (SEANCE) (Crossley, Kyle, & McNamara, 2017; Pennebaker et al., 2007). NLPs function through inputting written text, and emotions and sentiment are extracted typically through a bag-of-words approach, which counts the frequency of words from a pre-defined lexicons (Maruf et al., 2024; Nandwani & Verma, 2021). While NLPs do contribute to limiting human bias when interpreting an individual's emotion, the output of these tools is still a representation of the overall sentiment and a list of emotions that were identified from the written text. However, the outputs of NLPs require an individual to make sense of what these values and identified emotions mean. Therefore, there remains a need to have a human interpretation or coding to provide nuance when reporting the outputs from various NLPs.

Research Aim

An effective method for examining teachers' emotions is through reflective writing (Shapiro, 2010). Reflections have previously been defined as the process of thinking about experiences that are currently occurring or have previously occurred, learning from the outcomes of that experience, and planning for similar, future events (Machost & Stains, 2023). Instructors' reflections are an important starting point to understand the link between teachers emotions and student outcomes (Frenzel, Daniels, & Burić, 2021). Thus, this work examines the reflective writings of higher education instructors to characterize the emotions and overall sentiment expressed when instructors reflect on a remembered teaching experience. The two main research questions driving this investigation are as follows:

1) What emotions are present when instructors reflect on a remembered teaching experience?
2) What reasons do instructors connect to the emotions they recall feeling during their teaching experience?

Methods

This study was conducted under Protocol #5248, "Collection and Evaluation of Reflective Writings of Stem Instructors," as approved by the University of Virginia Institutional Review Board for the Social and Behavioral Sciences.

Participants

Participants were recruited from three iterations of a national workshop whose aim is to empower and improve the pedagogical practices of new physics and astronomy instructors (Physics and Astronomy Faculty Teaching Institute, 2023).

The participants in this study were asked to complete a Qualtrics survey that included a written scaffold (Appendix A) asking instructors to reflect on a previous teaching situation. The first cohort of participants completed the survey in July 2022; the survey was distributed during the workshop. The subsequent two cohorts completed the survey as a pre-workshop activity in June 2023 and November 2023, respectively, after solicitation via email. Across the three cohorts, a total of 125 instructors participated in this study (Table 11)

Participants were included in the study based on two criteria: (1) their written response needed to describe a time or situation when they were acting as an instructor, and (2) their description of the situation had to be clear and detailed enough to be easily understood without requiring interpretation by the research team. The participants of this study are associated with a variety of degree granting institutions (i.e., AA/AS, BA/BS, MA/MS, PhD) across all regions of

the continental United States and Ontario, Canada. Thus, all participants are employed by North American academic institutions despite their individual contextual differences.

Table 11. University demographics of participants

Highest degree awarded	n
Associates	3
Bachelors	18
Masters	27
Doctoral	71
Other	6
Total	125

Demographic information is provided for the 79 participants from the second and third iteration of the workshop (Table 12); such information was not collected in the first iteration.

Table 12. Participant self-identified demographics

Race or Ethnicity	Asian or Asian American	18
	Black or African American	3
	Latino/a/x	4
	White	48
	I prefer not to respond	0
	An identity not listed	6
Gender	Men	47
	Woman	28
	I prefer not to respond	1
	An identity not listed	3
	Total	**79**

Survey Design

This study is emergent from a multifaceted project centered on understanding and developing reflective practices among novice practitioners. In their literature review on reflective practice (Machost & Stains, 2023), authors HM and MS developed a scaffold for written reflection drawing on the works of Gibbs (1988), Larrivee (Larrivee, 2000, 2008a, 2008b), and Bain et al. (2002). The survey analyzed herein is composed of open-ended prompts for instructors to respond to this scaffold. The scaffold begins by prompting participants to identify a past teaching situation which met the provided definition of a critical incident. Participants then describe the facts of the situation. This is followed by a consideration of their feelings and their interpretations of the feelings of others involved. Next, they evaluate the incident for cause-effect relationships and for areas of improvement and aspects that went well. Finally, the instructors are asked to draw conclusions and plan for future similar situations. An example answer was provided for each step of the process.

Survey Analysis

A multifaceted approach – utilizing human coding and an automatic text analysis tool – was used in the analysis of the emotions and sentiments expressed by instructors in their written reflections. While the open qualitative coding provides the basis for much of the findings described herein, an automated text analysis tool was also used in order to aid in monitoring and flagging any biases which may be subconsciously carried by the authors. The measure is deemed necessary due to the connotations each person can have about their own experiences or with particular emotions. To ensure rigor, both qualitative coding and text analysis needed to be performed from the same body of text. Thus, prior to the two analyses, the written responses were cleaned in order to isolate the emotions that the instructors indicated experiencing during situation being described.

Authors EAK and HM met to outline the parts of the reflection survey responses which should be removed or cleaned. Subsequently, both HM and EAK independently cleaned each response before meeting to come to complete consensus on the text to be used for analysis. Specifically, we removed sections that were not concerned with in-the-moment experiences, emotions that the instructors themselves expressed as a delayed or retrospective response, and/or emotions that the instructors associated with others involved (i.e., students' emotions). Furthermore, due to the nature of the automatic analysis tool, words common to the classroom which have negative connotations (e.g. solving a 'problem' during a 'lecture') were replaced with their part of speech (e.g. solving a 'noun' during a 'noun'). An example of this process is depicted in Figure 4.1.

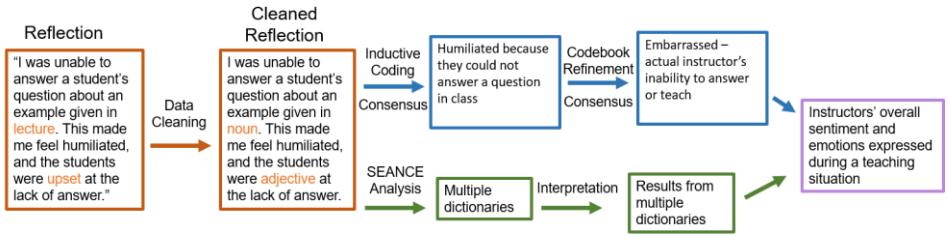

Figure 7. Outline of response cleaning and analytic processes

Qualitative Coding

HM prepared qualitative memos on the survey responses collected during the first two iterations of the workshop which captured the emotions expressed by instructors and the reasonings instructors linked to their emotions. After data from the three cohorts was collected and cleaned, HM and EAK independently analyzed each survey response using inductive in vivo coding. Following this, complete consensus was reached using inductive in vivo coding to capture

the emotions expressed by instructors and the reasonings they tied to their emotions. Once consensus was established using the in vivo codes, EAK and HM returned to the original qualitative memos to guide the process of grouping similar reasonings together. The original memos prepared by HM served as a starting point for secondary coding of the instructors' self-identified reasons for their emotions, enabling the inductive in vivo codes to be grouped through thematic analysis. Next, HM and EAK grouped similar inductive in vivo codes describing the emotions felt based upon ability to clearly articulate differences and consultation with Webster's dictionary. Once this codebook was established, authors HM and EAK reached complete consensus on all survey responses.

SEANCE

Once the responses were cleaned and consensus reached, the responses were converted into .txt files and analyzed using the SEntiment Analysis aNd social Cognition Engine (SEANCE). The SEANCE scores are primarily utilized as a metric to determine the overall emotions in and ANEW metrics for participants' responses. However, there is an additional purpose; SEANCE scores act as further evidence for the findings of the qualitative coding via a method that minimizes potential human bias when interpreting instructors' emotion. SEANCE is an analytical tool which produces simultaneous outputs from various natural language processing (NLP) programs and has shown utility across multiple domain types (e.g., movie reviews and social media posts) – a somewhat unique characteristic for NLPs which tend to only be accurate for the domain they were designed for (Crossley, Kyle, & McNamara, 2017). SEANCE outputs can include up to 3,000 potential metrics depending upon the combination of components the NLP is directed to analyze. This work only analyzes metrics from the Affective Norms for English Words (ANEW) and the Geneva Affect Label Coder (GALC) (Table 13).

Table 13. SEANCE metrics utilized in analysis

SEANCE metric	Type of Categories	Categories	Outputs	Interpretation	Pros	Cons
Hu-Liu polarity	Sentiment Score	Strongly Negative to Strongly Positive	Ratio between negative:positive proportion*	Strongly Negative: >1.50 Negative: 1.26-1.50 Neutral: 1.25-0.75 Positive: 0.66-0.76 Strongly Positive: <0.66	Ease of interpretation. Single score for each document.	No clear precedent for interpretation in the literature. Limited in detail.
VADER	Sentiment Score	Strongly Negative to Strongly Positive	Compound score ranging from -1 to 1	Strongly Negative: < -0.6 Negative: -0.6 to -0.06 Neutral: -0.05 to +0.05† Positive: +0.06 to +0.6 Strongly Positive: > +0.6	Ease of interpretation. Single score for each document.	Limited in detail.
ANEW	Sentiment Dimensions	Valence (Happiness), Dominance (Control), Arousal (Excitation)	Score between 1-9 for each dimension	Low: 1 to 2.99 Low-mid: 3.00 to 4.49 Neutral: 4.5 to 5.5 High-mid: 5.51 to 7.00 High: 7.01 to 9	Greater detail when examining sentiment	More complex interpretation
EmoLex	Emotion Categories	8 Major Emotions‡	Decimal per emotion category	Larger score indicates greater frequency. 0 Indicates absence	Ease of Interpretation.	Limited categories with large lexicons can inflate the scores. Few positive emotion categories.
GALC	Emotion Categories	36 Emotions∞	Decimal per emotion category	Larger score indicates greater frequency. 0 Indicates absence	Ease of Interpretation. Greater emotion delineation.	Limited lexicons result in small scores and frequent scores of 0.

Negation that occurs within three words of the classified word was accounted for using the SEANE program for all metrics where this is not done automatically.
*Original output from SEANCE is positive:negative. We transformed this to negative:positive
†Neutral range was taken from literature precedent
‡EmoLex Categories: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, positive, and negative
∞GALC Categories: Admiration, Amusement, Anger, Anxiety, Being Touched, Boredom, Compassion, Contempt, Contentment, Desperation, Disappointment, Disgust, Dissatisfaction, Envy, Fear, Feeling Loved, Gratitude Guilt, Happiness, Hatred, Hope, Humility, Interest/Enthusiasm, Irritation, Jealousy, Joy, Longing, Lust, Pleasure, Pride, Relaxation, Relief, Sadness, Shame, Surprise, Tension

Trustworthiness

As this is an exploratory qualitative study, the trustworthiness of the data is of upmost importance. Thus, steps were taken throughout this analysis to ensure credibility, transferability, dependability, and confirmability.

Credibility. Shenton (2004) highlights numerous methods to demonstrate credibility, and we have taken the steps applicable to this study. Frequent debriefing sessions occurred within the research team from conception of the study to the review of the final manuscript. Furthermore, qualitative coding is complimented by SEANCE scores, providing evidence from "research methods well established" in the literature (Shenton, 2004, p. 64). This includes analyses by GALC and ANEW. Importantly, authors took steps to ensure the transparency of data analysis. We used SEANCE due to the established performance of the program in sentiment analysis and its freely available software, enabling the analysis methods used herein to be available to other researchers.

Transferability. We promoted transferability by providing a detailed description of the context of the study, the study participants, the data collection procedures, and analytic processes. However, the transferability of this exploratory study is inherently limited due to the specific participant pool (instructors of physics or astronomy attending a new faculty workshop).

Dependability. The stability of our findings is primarily addressed in this exploratory study through reaching complete consensus between authors HM and EAK. Additionally, the stability is furthered by the application of analytic methods to three separate cohorts of participants. The codebook readily applied across all responses which were collected in three groups across two years. Finally, the NLP analytical tool used for sentiment analysis (SEANCE) is well established in the literature as a reliable tool (Crossley, Kyle, & McNamara, 2017; Rebora, 2023).

Confirmability. We address confirmability through two primary methods: coding to complete consensus and utilization of an NLP tool. Qualitative coding to complete consensus ensures that no one researcher's biases, point of view, or background knowledge unduly impacts the results. Additionally, an audit trail was kept throughout the process of both cleaning and coding the responses. The use of SEANCE furthers the confirmability of this study as the results from the NLP will be the same for anyone using the same texts. Thus, when the results from SEANCE corroborate the findings from qualitative coding, we can be reasonably certain that human biases are not unduly impacting the results.

Results and Discussion

The results first presented originate from SEANCE, specifically the ANEW metrics. Following this, the results of human coding for emotions are presented and validated through comparison to the GALC results. Finally, the qualitative coding for the reasons instructors associate with their emotions are presented.

ANEW

One response received ANEW scores of zero; as the scale for the three domains used in ANEW ranges from 1 to 9, this indicates that the words utilized in this response were not part of the ANEW lexicon lists. Thus, the results described in the following paragraphs are limited to the 124 responses which did receive ANEW outputs for valence, dominance, and arousal.

The valence scores for the survey responses were the most widespread, meaning our participants had a significant range in the domain which indicates the happiness associated with the written text. However, more responses were scored as some degree of 'happy' (High-mid or High) as compared to a degree of 'unhappy' (Low-mid or Low). Both the arousal and dominance measures of our sample were primarily neutral, meaning that the instructors indicated feeling

neither very animated nor overly lethargic as the situation unfolded, and neither in-control nor out-of-control of the described situations. This is furthered by the lack of responses which were categorized as low-dominance, high-dominance, and low-arousal. Additionally, only one response was classified as having high-arousal. However, there is enough variance from Low-mid to High-mid in the dominance and arousal domains to enable a preliminary check for correlations between scores across the three domains (Figure 8).

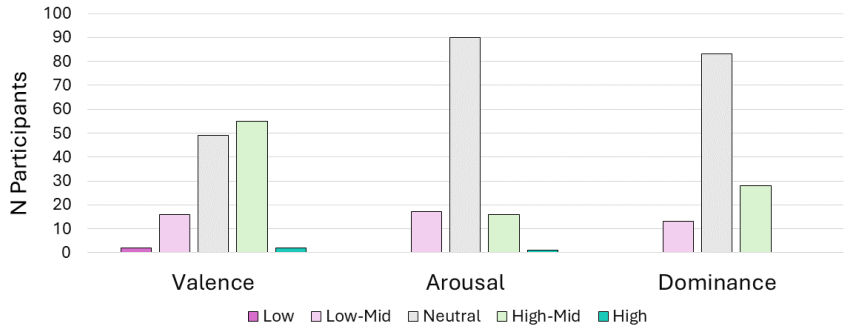

Figure 8. Abundances of ANEW scores

Correlation analyses were completed for all combinations of the three domains; however, the only trend observed was between Valence and Dominance (Figure 9). As the responses move from Low Valence (unhappy) to High Valence (Happy), the Dominance scores also shift from feeling out-of-control (Low-mid Dominance) to in-control (High-mid Dominance). Appendix C3, Figure C3.1 shows the lack of correlation between Valence scores and Arousal scores. These results together indicate that while an instructor's intensity of feelings is not associated with either their perception of control in a situation or their happiness during it, instructors did feel happier when they felt more in-control of their environment.

Feelings of control are strongly linked to an individual's behavior in the context of teaching strategies. Among instructors who either did not use or had discontinued certain strategies, fear of losing control over the classroom environment emerged as a key reason for avoiding these practices (Brooks et al., 2024). This fear can be connected to the concept of perceived behavioral control, a core component of the Theory of Planned Behavior (TPB), which asserts that an individual's belief in their ability to perform a behavior influences their intention to engage in it, and ultimately their actual behavior (Ajzen, 1991). In educational settings, as seen in the work of Reinholz and Andrews (2020), TPB suggests that instructors' attitudes towards EBIPs, their perceptions the status quo relating to a practice, and their perceived control over executing these strategies significantly impact their likelihood of adopting such practices.

The strong linkage evidenced herein between feelings of control and happiness may aid in explaining prior work which showed a connection between positive feelings and adoption of student-centered strategies (Kordts-Freudinger, 2017; Mattanah et al., 2024; Trigwell, 2012). While this exploratory study does not provide enough data to perform a causal analysis, it is likely that the promotion of training in regards to student-centered strategies can assist instructors feelings of control over their classroom when implementing such practices. This combination potentially would result in instructors being more likely to adopt EBIPs and other pedagogical innovations they receive training for.

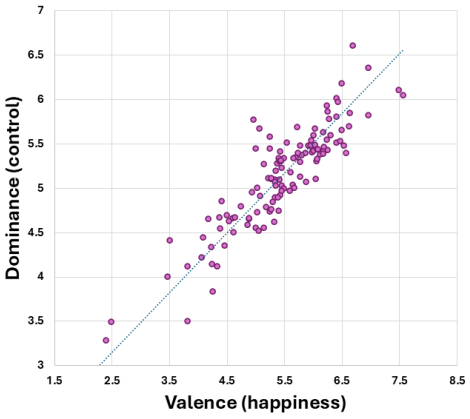

Figure 9. Correlation graphs between ANEW valence and dominance metrics

Emotions present

The emotions present in our participants' responses were primarily determined through qualitative coding, with authors HM and EAK readily reaching complete consensus on all responses. The results of the qualitative coding for the emotions expressed by participants are supported by the results of SEANCE's GALC analysis.

Qualitative Coding

Qualitative coding showed 23 different emotions expressed by our participants. However, only emotions present in at least 5% (n≥7) of the participants are further explored herein. Thus, while the following emotions were present, they were not explored in Table 14: *relieved, exhausted, betrayed, hopeful, overwhelmed, confident, conflicted, helpless, fearful*, and *defensive*.

Table 14. Common emotions as detected through qualitative analysis

Emotion Qualitative Coding	% responses
Guilty	49
Angry	41
Anxious	23

Confused	21
Surprised	20
Nervous	18
Happy	18
Embarrassed	13
Uncomfortable	10
Empathetic	8
Sad	6
Doubtful	6
Disappointed	6

Qualitative analysis revealed 14 emotions that were present in at least 5% of participants' responses when reflecting on a teaching-related experience (Table 14). The analysis of our instructors' reflections reveals high frequencies of both guilt and anger being experienced. These two emotions are on the negative end of the spectrum, associated with a low valence. Thus, it is likely that instructors who experienced guilt and anger also did not feel in-control of the situations they recalled in the reflection. A smaller but still significant portion of our instructors reported feeling anxious, confused, and surprised which can also be indicative of feeling not in-control of the situation. The sixth most common emotion, that of happiness, is the most common emotion associated with high-valence scores despite still occurring less frequently than feelings of guilt, anger, confusion, or anxiety.

GALC comparison for review of bias

The 36 emotion categories in GALC are provided in the supplemental information (Appendix C2, Table C2.3). When comparing the GALC outputs to the qualitative coding, authors HM and EAK classified the qualitative codes into GALC categories based on the GALC lexicon lists. For the analysis herein, we focus on the GALC emotion categories which were present among our sample, and we then provide the comparison to the qualitative coding as determined through the GALC lexicon analysis. Of the 36 GALC emotion categories, ten are wholly absent from our

participants' responses ('Boredom,' 'Contempt,' 'Envy,' 'Feeling Loved,' 'Hatred,' 'Humility,' 'Jealousy,' 'Joy,' 'Lust,' and 'Pleasure'). A further 13 were present in fewer than 5% ($n \leq 6$) of the responses: 'Compassion,' 'Contentment,' 'Guilt,' 'Amusement,' 'Relaxation,' 'Relief,' 'Sadness,' 'Admiration,' 'Desperation,' 'Disgust,' 'Dissatisfaction,' 'Gratitude,' and 'Pride.' Of these, 'Guilt,' 'Joy,' and 'Sadness' are far more prevalent in the inductive qualitative coding. However, as described below (Table 15), there are additional GALC emotion classifications which likely correspond to the common qualitative codes of Guilty, Happy, and Sad. Notably, qualitative codes of *Overwhelmed, Conflicted, Betrayed, Exhausted, Defensive, Doubtful, and Confused* cannot be confidently aligned with a GALC category; this is likely due to the brevity of the lexicon lists for each GALC category. Overall, despite the somewhat mixed alignment due to differences in emotion categories, GALC provides evidence in support of the emotions classified during the qualitative analysis. Indeed, there is a predominance of negative emotions such as anger and disappointment, with a smaller proportion of the reflections exhibiting emotions such as happiness.

Table 15. Alignment between GALC and qualitative coding

GALC Emotion Category	% responses with GALC emotion	Alignment with emotions from qualitative coding*
Anger	34	Angry
Disappointment	20	Angry, Disappointed
Happiness	14	Happy
Surprise	13	Surprised
Longing	12	Guilty
Shame	11	Guilty, Embarrassed
Anxiety	11	Anxious, Nervous
Tension/Stress	11	Nervous, Uncomfortable
Fear	9	Surprised, Fearful
Being Touched	8	Empathetic
Hope	7	Hopeful
Interest/Enthusiasm	6	Happy
Irritation	6	Angry
Contentment	4	Happy
Compassion	4	Empathetic
Guilt	3	Guilty
Admiration/Awe	2	Hopeful, Happy
Amusement	2	*No Match*
Relaxation/Serenity	2	Happy
Relief	2	Relieved
Sadness	2	Sad
Desperation	1	Helpless
Disgust	1	Angry
Dissatisfaction	1	Sad
Gratitude	1	*No Match*
Pride	1	Confident

*the alignment was determined by authors HM and EAK after analyzing the qualitative codes for their presence in the GALC lexicons

Reasons for emotions

In the remainder of this section, we will be focusing on the emotion code present in at least 5% of the sample and their associated reasonings. Notably, the reasoning codes are not mutually exclusive, as an instructor may feel more than one emotion for the same reason, such as feeling both disappointed and empathetic in response to a student's actions.

Table 16. Most common reasons associated with emotions

Reasoning Code	Number of Occurrences
Experienced student's actions	69
Experienced instructor's inability to answer or teach	58
Experienced instructor's actions	50
Experienced student academic challenges	19
Experienced instructor's ability	18
Experienced feedback	16
Experienced unknown/unclear solution or path	16
Experienced instructor effect on students	12
Experienced lack of student prior knowledge	12
Experienced instructor unanticipated adjustment	9
Anticipated student's actions	8
Experienced student's circumstances or context	8
Anticipated instructor effect on students	7

The predominant reasoning tied to the emotions instructors described experiencing was Experienced student's actions. However, this was closely followed by Experienced instructor's inability to answer or teach and Experienced instructor's actions (Table 16). Of the top reasons

that instructors connect to their feelings, roughly half are directly related to an instructor's actions, inactions, or response to the situation. The remaining half are split between students' actions and students' circumstances, such as their academic challenges and personal situation.

Further analysis enables mapping of the most common emotion codes to the most common reasoning codes. As expected, there is a significant overlap between the top three in each category. Below, the five most common emotions are reported with the reasons which, together, account for at least 70% of that emotion's occurrences.

Guilty is primarily associated with *Experienced instructor's ability to answer or teach* and with *Experienced Instructor's actions,* as seen with Instructors 66 and 87, respectively.

"I felt horrible; these students are here to be taught correct physics and I did not deliver." –66

"I also felt regret that I did not slow down my brain while I was reading the question to think through it again. Instead, I just read off what I selected last year as the right answer."–87

However, the third most common reason for feeling *Guilty* is the instructor's view of their own abilities or inabilities (*Experienced instructor's ability*). This is clearly articulated by Instructor 102, who felt guilt when they provided advice and guidance to their student, but the student still did not succeed in their course.

"I was feeling regretful after this. I wish I could help my student more." –102

Guilt, the emotion most commonly experienced by instructors, primarily stems from their perceived inadequacy in teaching or answering questions. Additionally, guilt is linked to instructor's perceived responsibility for their students' outcomes. This finding resonates with a previous study which found that younger university instructors felt more guilty due to similar

reasons, including not feeling equipped to handle students' questions (Cubukcu, 2013). These findings indicate that these instructors are very self-critical of their own ability to teach or be a resource to their students. To alleviate the prevalent feelings of guilt among new instructors while teaching, there is a need to help instructors navigate these feelings of self-criticism. Additionally, there is a need to understand how instructors' self-criticism could impact their actions while teaching or interacting with students.

Angry is overwhelmingly associated with *Experienced student's actions*, with that lone reasoning code accounting for over 50% of all occurrences. Instructor 95 exemplifies this link.

"I typically feel frustrated at multiple times: during the semester when the student fails to use the resources I provide or respond to my suggestions as to how to get back on track. At the end of the semester when they finally realize that they have not completed enough to pass the class, when multiple reminders do not get the student to submit additional work." –95

Previous research had a similar finding (Geng & Yu, 2024). In this case study of one instructor, the authors observed that the instructor experienced a negative sentiment due to students' actions (Geng & Yu, 2024). A similar result was also found where university instructors' feelings of anger or frustration occurred based on students' actions (Cubukcu, 2013). *Angry* is further depicted by our instructors as being caused by *Experienced instructor's actions*, indicating anger being directed inwards. Instructor 110 described being *Angry* because "*1) [they] had made a mistake, and 2) had tried to mindlessly argue it was correct without thinking about what the student had said." Experienced feedback* is as prevalent a reason for causing *Angry* as *Experienced instructor's actions.* However, *Experienced feedback* was only directed outwards, as seen with Instructors 82 who experienced frustration with their TA.

"I was trying my best to adjust, but my TA was giving me a hard time - stating over and over again that students found the homework adjective even though I adjusted." –82

The second most common emotion, anger, was primarily triggered by the actions of others, particularly students, as well as feedback received from others. To a lesser extent, instructors also felt angry with their own actions, especially when they were unable to effectively teach.

Half of the instructors describe feeling ***Anxious*** due to their own inability to answer a question completely and correctly (*Experienced instructor's inability to answer or teach*), their own actions taken during a situation (*Experienced instructor's actions*), such as with Instructor 60 who had an emotional reaction after not being able to answer a question asked in class.

"When the question was asked, I immediately felt a sense of dread and panic" –60

Some also felt anxious in response to the actions taken by the students during a situation (*Experienced student's actions*).

Confused is associated most with instructors not knowing how to best address or approach a situation (*Experienced unknown/unclear solution or path*).

"I felt frustrated and conflicted on how to address the situation so the students listening could get the last of the material before the midterm the following class."–120

As shown above, Instructor 120 did not know how to proceed with a review session after a student caused a disruption. Still significant yet accounting for less than half the occurrences as *Experienced unknown/unclear solution or path*, is the reasoning code *Experienced instructor's inability to answer or teach*, as seen with Instructor 2 who did not know how to proceed with solving a type of in-class problem.

"Instead, I got lost, then flustered, and forgot how to use the chain rule for derivatives."–2

Overall, instructors feel confusion during situations in which they felt stuck or unable to move forward; some instructors were also confused by' actions such as Instructor 101 who had not anticipated student resistance to an active learning activity.

"…This was not the first time we had done this, so I was confused by the greater than normal student reluctance." –101

Surprised is most linked to *Experienced lack of student prior knowledge* and to *Experienced student's actions.* Less common yet still contributing to the overall prevalence of surprise among the instructors, is *Experienced student academic challenges*. Thus, as shown with Instructors 50 and 130, the instructors primarily experience surprise when students do not meet performance or behavioral expectations. Whereas Instructor 50 was surprised when one of her students made an inappropriate model during class, Instructor 130 was surprised by a student not having the appropriate math background.

"One of the students asked me about how to calculate the sine of theta. The problem was related to motion along the inclined plane. Since I was expecting it to be a high school level calculation, at least the sine of theta. and assume that student should have already known about it. But to my surprise, student did not know it" –130

Thus, instructors experience *surprise* due to their students, particularly in reference to their students' behaviors or academic journeys.

Implications

The purpose of this study was to characterize physics and astronomy instructors' emotions expressed when reflecting on a previous teaching experience, and the reasonings that instructors

connected to these emotions. The results highlight a complex emotional environment, with varied sentiments found throughout the sample and many instructors expressing more than one emotion. Instructors are feeling a range of emotions and sentiments in the classroom, and, therefore, there is a need to provide ways to support them and destigmatize the feelings instructors experience while teaching.

Implications for professional development

This study provides an insight into the emotions that instructors experience while teaching in their classroom. Instructors feeling negative emotions regarding their students' actions, such as students asking for additional instructional resources, highlights a need for instructors to prepare for the challenges associated with teaching. One avenue that should be explored when supporting instructors in this effort is the advocation for reflective practices. When instructors engage in reflective practices, they can further understand why they are experiencing feelings of anger or guilt. Reflective practices also encourage instructors to reflect on their role as a teacher and the purpose of teaching. This can lead to them better understanding their students' experiences and their own rationales that lead to their decided actions. Furthermore, reflective practices can combat the apparent feelings of guilt which was prevalent among the instructors. Through reflection, instructors recognize their successes while teaching (Brookfield, 2017; Machost & Stains, 2023; Mohamed, Rashid, & Alqaryouti, 2022). This was observed in the present study with a common emotion that instructors expressed, though far less prevalent than anger or guilt, were feelings of happiness. Reflective practices are thus a tool for instructors to consider potential areas for improvement while constructively exploring their emotional responses (Mohamed, Rashid, & Alqaryouti, 2022).

Instructional training should also consider instructors' emotions while teaching and implement activities aimed at helping them navigate the variety of positive and negative emotions they experience. Training instructors, especially novice instructors, through role-playing instructional experiences could allow instructors to practice instructing and interacting with students, to reflexively learn and develop their practice in a lower-stakes environment than in their own classroom (Schön, 1987). Such role-playing based training can also provide one avenue for destigmatizing instructor's emotional responses while teaching. Through designed scenarios, novice instructors can be guided through emotionally charged situations common in education, and their own feelings can be validated. In addition to this potentially having a positive impact on students, as seen with the association between negative emotion suppression and teacher-entered teaching, such practices can also benefit instructors through normalizing common experiences which could otherwise diminish their sense of self-efficacy. The practice scenarios can be further used to train instructors regarding how to redirect their negative emotions of guilt and anger toward productive actions, such as seeking additional training or learning about departmental resources.

Implications for education researchers

When supporting instructors, especially new faculty who are relatively inexperienced in teaching, it is important to consider instructors' emotions. Teaching, itself, is said to be an emotional practice (Cubukcu, 2013). The results from the present study show that instructors do experience a wide variety of emotions while teaching, but there is a tendency for them to focus on negative experiences when reflecting on past teaching experiences. Future research, as well as instructional institutions, need to provide support for instructors' feelings to help destigmatize emotional responses in the classroom which could potentially impact their instructional decision making. When continuing to explore the lack of adoption of student-centered instructional

strategies in STEM classrooms, individual factors such as instructors' emotions is an important aspect to consider. Trigwell suggested that there is a relationship between the way instructors emotionally experience teaching and their approach to teaching (Trigwell, 2012). Future research within STEM higher education needs to continue exploring how instructors' emotions intersect with their feelings of control and connect to their instructional decision. Additionally, research should further explore instructors' emotions in the classroom to support instructor's development of their emotional identity. For instructors to understand their emotional identity, which is how positive and negative teaching emotions impact their professional self, it must become more practiced in teaching (Shapiro, 2010). SEANCE provides a useful tool for characterizing instructors emotions from written text (Crossley, Kyle, & McNamara, 2017). Future research could use this tool as a method to understand STEM instructors' emotions in the classroom to continue understanding how to support their development of their emotional identity.

Limitations

The exploratory design of this study limits the generalizability of its findings. The small sample size, which consists solely of assistant professors in physics and astronomy, restricts the applicability of the results to other STEM or non-STEM fields. Additionally, since participants voluntarily chose to attend this pedagogical workshop, they may not represent the broader population of new instructors in physics and astronomy. Also, given the personal nature of the incidents discussed in this study, participants might have been hesitant to fully explore their emotional states or responses despite the survey's confidentiality. Additionally, due to the examples provided in the scaffold for instructors to reflect, instructors potentially could have leaned towards reflecting on a negative topic. However, with the large number of instructors who reflected with a positive sentiment, this was not a major concern. Lastly, due to the nature of the study with

instructors describing a prior teaching experience, instructors may experience an immediacy bias and the intensity of the emotions they recall feeling could differ from what they felt in the moment.

Conclusion

Instructors' emotions in the higher education classrooms have previously been shown to be associated with their adoption of student-centered teaching strategies. However, within STEM, there is a lack of understanding the emotions instructors experience while teaching in the classroom. This study addresses this gap by characterizing instructors' emotions they recall feeling on past teaching experiences. An analysis of the three affective dimensions from the Affective Norms for English Words (ANEW) framework revealed a distinct correlation between happiness and feeling in control of the situation, with no similar correlation observed among other dimensions. When examining the emotions expressed, guilt and anger were predominant. Guilt was the most frequently expressed emotion and was typically linked to instructors' own actions or their struggles with effectively addressing questions or teaching concepts. While anger was not uni-directional, and could be caused by students' actions, instructors' actions, or additional factors, over half of all instances of anger were linked to instructors' experiences with students. The prevalence of these negative emotions underscores the need to better normalize and address emotional experiences within the classroom. This is especially pertinent in Western academic contexts, such as those in the United States and Australia, where such emotional experiences are tied to teacher-centered pedagogical approaches. It is thus crucial for educational institutions to prepare instructors for the emotional challenges they may encounter while teaching and to foster environments where a wide range of emotional experiences are normalized.

Part 1 Conclusions and Future Directions

Conclusions

The studies presented thus far highlight the importance of training and professional development among novice STEM instructors. Training in reflective practice must inform instructors of different aspects to consider and work to normalize emotional responses in educational environments. Many new faculty do not reach high levels of reflection, indicating that their reflections may not result in pedagogical change. However, the presence of high-level and complex content in certain reflections indicates that the developed scaffold can foster effective reflection. The need for professional development is further seen with the negative emotions instructors report experiencing in the classroom. These emotions, including anger and guilt, are associated with teacher-centered practices in American classrooms. As such, providing novice instructors with tools to navigate emotions in educational environments is vital. Together, chapters three and four showcase the necessity of continually studying faculty in order to better support their pedagogical practices. Specifically, a greater understanding of instructors' reflective practices and of their emotional expression is vital for effective professional development of STEM faculty as is the development of a system of sustained support especially during the early formative years of being an instructor.

Future Directions

The first future direction will focus on the analysis of novice STEM instructors' written reflections to determine the support structures they utilized when confronted with self-identified critical incidents in their teaching. Initial coding for this project has already been completed, and further analysis will be carried out by our research team to identify key patterns and themes. The results of this investigation will be invaluable for informing educational change agents, department

heads, and faculty development programs about the support structures that are used by STEM instructors. Understanding which support systems utilized by novice faculty, particularly when navigating challenging teaching moments, will guide the development of targeted interventions and resources to foster instructional growth and resilience.

Building upon the analyses presented in chapters 1 and 2, a second future investigation will cross-reference the nature of instructors' reflections, their plans for future improvements, and the emotions they report experiencing in those critical teaching moments. By examining these three interrelated aspects, this study will deepen our understanding of the role emotions play in shaping instructors' reflective processes and their subsequent instructional decisions. In particular, we will explore whether certain emotional responses – such as anger or guilt – correlate with specific plans for improvement and with the quality of reflections themselves. A follow-up study will further investigate the support structures used by instructors when faced with teaching challenges and examine the potential connections between these supports and the emotional experiences of instructors. By focusing on the emotional dimensions of reflection, this line of inquiry will help to illuminate how emotions may impact decision-making and pedagogical approaches within American STEM programs, ultimately contributing to the growing body of research on the emotional complexities of teaching.

A third future investigation will explore the longitudinal effects of reflective practices on novice instructors' development, particularly how these practices evolve over time. Data collection is currently ongoing for 6-month and 12-month follow-up reflections, using the same reflective writing prompts as in previous studies. Once sufficient data has been gathered, the analysis from chapter 2 will be repeated, supplemented by semi-structured interviews, to assess how reflective practices change as instructors gain more teaching experience. This longitudinal study aims to

uncover how novice instructors' reflections evolve in complexity and depth over time. The concurrent interviews will provide rich insights into the contextual factors and professional development aspects that foster effective reflection among novice instructors. Understanding the long-term trajectory of reflective practices will help inform faculty development programs, providing evidence-based strategies for supporting instructors in their ongoing growth as reflective practitioners.

These three future investigations will advance our understanding of the role of instructor reflection in STEM education, focusing on the support structures that facilitate novice instructors' growth, the emotional dimensions of teaching, and the evolving nature of reflective practices over time. Together, these studies will inform the design of more effective professional development initiatives, with an emphasis on providing instructors with the tools and support they need to navigate the emotional and pedagogical challenges they face in the classroom. By shedding light on how reflective practices can be optimized and sustained, these investigations will contribute to the broader goal of enhancing teaching quality and fostering a more supportive and dynamic teaching environment in STEM education.

Part 2 Introduction: A brief overview of specifications grading

In recent years, many innovative pedagogical practices have been proposed in post-secondary education to address students' needs in the ever-evolving learning environment (Beach, Henderson, & Finkelstein, 2012; Henderson, Beach, & Finkelstein, 2011). One class of innovation that is gaining attention from both education researchers and practitioners is the methods instructors use to assign grades to their students. Specifically, there is an ongoing shift among some instructors away from traditional grading and towards alternative grading systems (Clark & Talbert, 2023; Hackerson et al., 2024). It is argued that traditional grading, which assigns evaluative grades on individual assignments, is not effective at focusing students on the learning process, nor in promoting an equitable learning environment. Alternative grading schemes aim to address these issues. One of the most represented alternative grading schemes in chemistry education is specifications grading (Hackerson et al., 2024).

The path to alternative grading

Receiving a points- or percentage-based grade is a ubiquitous experience in western schools, from primary through post-secondary education. However, the practice of assigning such grades has not always been the standard. Performance at academic institutions has always been assessed; this assessment was originally a determination of whether a student had "mastered [their studies] to a level comparable to and determined by other masters" during a leaving exam (Williams, 2022, p. 8). Indeed, such examinations are still in use today as seen in doctoral defenses. A shift away from this classical measure of learning occurred in the 18th and 19th centuries. As academia grew and a desire to compare students' performances emerged, universities began to implement grades or simple marks (e.g. Senior Optimes, Junior Optimes), and slowly

incorporated more frequent assessment (Clark, 2019; Schneider & Hutt, 2014). By the mid 1900's, this scheme had evolved into the more granulated and formalized A-F system, which is still in use today (Clark, 2019; Schneider & Hutt, 2014). It is at this time that such grades also "generally aligned with numerical values —an A reflecting work between 90 and 100, for instance, and a B reflecting work between 80 and 89" (Schneider & Hutt, 2014, p. 215). This emergent grading scheme, called traditional grading, thus encapsulates the "assigning of points to one-time assessments and aggregating those points into a letter grade for the course" (Clark & Talbert, 2023, p. 11).

While such grades undoubtedly have use as means of inter-institutional communications, such as in the case of students applying to universities from many different high schools, there are numerous characteristics of traditional grading that have been critiqued. Indeed, the traditional grading scheme is argued to have several limitations. Firstly, the utility of traditional grades is limited. These grades are evaluative, meaning they are used to indicate students' performance on individual assignments and in the course overall; however, they do not provide actionable feedback or instruction regarding how students can improve (Cain et al., 2022). This is compounded by the fact that traditional grades can reflect a student's access to mental health resources, high-school preparation, and/or financial stability as opposed to their learning gains (Feldman, 2019a, 2019b; Link & Guskey, 2019; Matz et al., 2017; McKay, 2019). Furthermore, the grades students are assigned can differ greatly depending on one instructor's standards as compared to (Cain et al., 2022; Donaldson & Gray, 2012; Herridge & Talanquer, 2020; Herridge, Tashiro, & Talanquer, 2021). Beyond the unreliable measurements of learning, students in traditional grading were found to have increased levels of anxiety and decreased self-motivation to learn (Chamberlin, Yasué, & Chiang, 2018; Lewis, 2020; Pulfrey, Buchs, & Butera, 2011; Schinske & Tanner, 2014).

Resultant from these issues is the movement towards alternative grading systems (Clark & Talbert, 2023; Danielewicz & Elbow, 2009; Kohn, 2011; Nilson, 2015). The primary goal of alternative grading systems are summarized by Clark and Talbert (2023) and include having clearly defined standards, incorporating helpful feedback, ensuring marks are representative of learning, and the ability of students to reattempt without penalty. *Clearly Defined Standards* are a set of criteria provided to students, such as specified learning outcomes, complimented by outlined steps that students can take to demonstrate their understanding and abilities. *Helpful Feedback* is provided to students; said feedback must be actionable and provided in relation to the clearly defined standards and steps students can take to improve. *Representative Marks* are meant to replace point- or percentage-based grades. These marks indicate whether students have completed assignments to the desired level (e.g., "satisfactory," "needs revision"). Finally, students are able to *Reattempt* assignments *Without Penalty.* These opportunities for revision enable students to use the helpful feedback provided in order to improve their performance to meet the clearly defined standards while learning from their mistakes. This approach fosters a learning environment where improvement and mastery are prioritized (Clark & Talbert, 2023). Additionally, alternative grading schemes which implement each of these four pillars are anticipated to ensure that students have equal access to opportunities and to aid in the removal of structural barriers to success inherent in the traditional grading system (Clark & Talbert, 2023).

Specifications Grading

One alternative grading system that has been gaining increasing attention from both researchers and educators is specifications grading. First formalized by Nilson (2015), specifications grading evaluates students' assignments based on a simple 2-level system, where each assignment is assessed according to a set of predefined standards that are made clear to the

students beforehand. These standards, which are explicitly outlined, determine whether the students' work meets the criteria for receiving credit. If the student meets the standards for an assignment, they earn credit for that particular assignment. However, if the student does not meet the standards, they will not receive credit for the assignment immediately. Instead, they receive detailed, constructive feedback from the instructor, which they can use to improve their work. The students are then given the opportunity to revise their assignment and resubmit it. It is important to note that the instructor can set limitations on how many revisions are allowed, depending on their preferences or any constraints related to the course context, such as available grading time or the course's length.

A key feature of specifications grading is that the standards for each assignment are intentionally set to represent, at a minimum, the quality of work that would be considered at least "B-level" under a traditional grading system. This ensures that students cannot pass the course without demonstrating a basic level of proficiency and quality in their work.

Additionally, specifications grading organizes the grading into "bundles" of assignments, each of which is linked to a specific letter grade. These bundles group together related assignments, and students' final grades are determined by how many assignments meet the preset standards within each bundle (Nilson, 2015). The bundling of assignments can be done in many ways; three such methods are depicted in Figure 10. One model of bundling involves a different number of assessments that meet specifications. For example, there are five different assessment types (represented by the five color blocks) and meeting specifications on a different number of assignments for each assessment type yields a different letter grade; the highest-level bundle where all specifications are met determines the course letter grade. A second model of bundling involves different assessment types. The distinction between bundles, and thus between the final grades, is

determined by the number of specifications, the assignment types that met specifications, and/or whether the specifications met align with foundational or additional learning objectives.

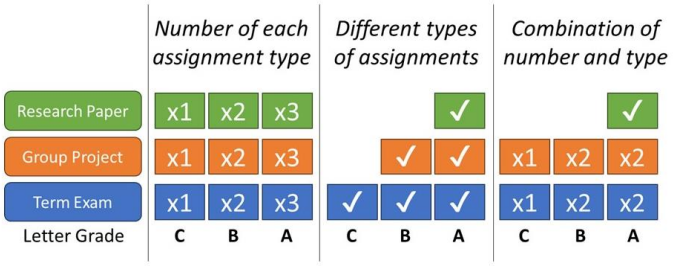

Figure 10. Depiction of various bundling schemes in specifications grading

In addition to the bundling system, specifications grading is unique in the alternative grading world due to the use of tokens. Tokens act as a form of currency for students within specifications graded courses; these tokens can be provided to students or earned by students through completing surveys, assignments, or other course-relevant tasks. Once students have either received or earned tokens, they can be exchanged for various purposes, including receiving the ability to reattempt assignments, excusing absences, receiving deadline extensions, etc. Thus, specifications graded courses often use token systems to address the fourth pillar of alternative grading (reattempts without penalty); however, tokens are not a requirement for the fourth pillar to be implemented in specification grading schemes. While not all specifications grading systems use tokens, the token system itself is designed to provide flexibility for students outside of what must be asked for of the instructor. Additionally, token systems can increase student autonomy and choice as the students will be the ones deciding when and how to use their tokens.

Specifications grading thus has six main distinguishing features: (1) individual assignments are graded on a pass/fail basis, (2) students are provided clear specifications regarding what assignment expectations, (3) specifications reflect the standards of B-level or better work, (4)

students are provided opportunities to revise and resubmit work (5) final grades are determined through a bundling system as opposed to a weighted average, and (6) bundles are aligned with the course learning outcomes (Nilson, 2015). When implemented completely and correctly, Nilson claims that specifications grading achieves 15 outcomes (Table 17); many of these outcomes correspond with Talbert and Clark's (2023) four pillars of alternative grading. It is important to note that these 15 outcomes are only hypothesized by Nilson and are not yet empirically backed.

Table 17. Proposed outcomes of specifications grading

Outcome	Description	Outcome	Description
1	Uphold high academic standards	8	Minimize conflict between faculty and students
2	Reflect student learning outcomes	9	Save faculty time
3	Motivate students to learn	10	Give students feedback they will use
4	Motivate students to excel	11	Make expectations clear
5	Discourage cheating	12	Foster higher-order cognitive development and creativity
6	Reduce student stress	13	Assess authentically
7	Make students feel responsible for their grades	14	Have high interrater agreement
		15	Be simple

Diffusion of specifications grading

The popularity of specifications grading was gained quickly, with specifications grading being featured in the American Chemical Society's C&EN magazine just six years after its inception (Arnaud, 2021). Implementations in chemistry that have resulted in publication include courses in general chemistry (Bunnell et al., 2023; Howitz et al., 2023; Martin, 2019; Noell et al., 2023; Saluga et al., 2023), organic chemistry (Ahlberg, 2021; Houseknecht & Bates, 2020; Howitz, McKnelly, & Link, 2021; McKnelly et al., 2023; Ring, 2017), writing for chemists (McKnelly, Morris, & Mang, 2021), analytical chemistry (Hunter, Pompano, & Tuchler, 2022), physical chemistry (Closser, Hawker, & Muchalski, 2024), and biochemistry and chemical biology (Donato

& Marsh, 2023; Kelz et al., 2023). Further evidence of this near-unprecedented propagation is the increasing numbers of symposia and presentations at the Biennial Conference on Chemical Education (BCCE) (Figure 11). Despite being first presented in 2015 (Nilson), in the four-year span of 2017-2020, there were 7 peer-reviewed manuscripts, 2 BCCE symposia, and 8 BCCE presentations on specifications grading in chemistry. These numbers drastically increased in the following four years (2021-2024), growing to 16 manuscripts, 10 BCCE symposia, and 58 BCCE presentations. The increase in BCCE presentations is perhaps most indicative of the accelerated spread of specifications grading; BCCE is a conference for educators as well as education researchers. Thus, the increase in presentations at BCCE is partially due to increased adoption of specifications in real courses which are then presented on by the educators themselves.

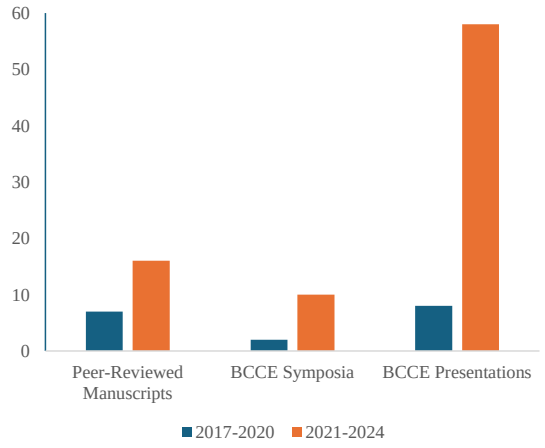

Figure 11. Growth of specifications grading

Previous studies on specifications grading

Despite the growing popularity of specifications grading, there remains little evidence for its effectiveness in the literature. A major focus of the prior studies that do exist is the method of

implementation, with many publications investigating its use in one course or by one instructor (Howitz et al., 2023; Howitz, McKnelly, & Link, 2021; Katzman et al., 2021; Kelz et al., 2023; Martin, 2019; Noell et al., 2023; Saluga et al., 2023). The work which explored specifications grading beyond its implementation mainly investigated the associations between specifications grading and student academic outcomes. For example, students are reported to earn higher letter grades (Bunnell et al., 2023; Katzman et al., 2021; McKnelly et al., 2023) and have positive student attitudes toward their course (Bunnell et al., 2023; Katzman et al., 2021). Additionally, previous literature has indicated an increased quality of interactions between instructors and students, with an emphasis on how students can better learn as opposed to receive better scores (Ahlberg, 2021; Bunnell et al., 2023; McKnelly et al., 2023). However, more work is vital to understanding this innovative practice that is rapidly gaining popularity in chemistry higher education in order to ensure its effective and appropriate implementation and propagation.

Aim of Part 2

While specification grading is gaining popularity among chemistry instructors, its associated benefits remain largely unsupported by evidence-based literature. It is, therefore, essential to investigate the propagation, implementation, and associated effects of specifications grading to better inform instructors and professional development coordinators about this grading scheme. As insights into specifications grading are developed through the work described herein, instructors who choose to move away from traditional grading will be able to make more informed decisions about the grading strategies they plan to implement into their classroom. Additionally, should specifications grading prove to be beneficial, the following chapters will provide valuable information regarding how to best implement specifications grading depending on an instructor's context.

Chapter 3. Benefits and challenges of specifications grading: Perceptions from chemistry instructors who have adopted this grading scheme

This chapter is adapted from a soon-to-be-submitted manuscript.

Introduction

The adoption of instructional strategies that have empirical evidence supporting their effectiveness (referred to as evidence-based instructional practices; EBIPs) is still lagging in chemistry undergraduate courses. A national survey study of introductory chemistry instructors (n=1,232) reported that only 51% consistently use EBIPs in their courses (Wang et al., 2024). Another national survey of chemistry instructors (Raker et al., 2021) explored the level of use of three specific EBIPs, i.e., Peer-Led Team Learning (PLTL) (Cracolice & Deming, 2001), Problem-Based Learning (PBL) (Servant-Miklos, 2019; Wood, 2003), and Process Oriented Guided Inquiry Learning (POGIL) (Moog, 2019). The statistical analysis of the data collected (n=829 faculty) produced an estimated 11%-17% of users of these strategies. Extensive research efforts have been devoted to characterizing factors related to the adoption of EBIPs by post-secondary science and chemistry instructors. For example, it has been repeatedly determined that contextual factors (e.g., departmental climate towards pedagogical innovation and physical classroom layouts) and instructor's personal factors (e.g., mindsets and personal experiences with specific pedagogical practices) are associated with instructors' decision to adopt innovative pedagogical practices (e.g., (Andrews & Lemons, 2015; Connor & Raker, 2023; Kraft et al., 2024; Lund & Stains, 2015; Sturtevant & Wheeler, 2019; Yik et al., 2022a, 2022b). Personal experiences, among these factors, have been reported to particularly concern instructors' adoption of these practices. For example, Andrew and Lemons (2015) found that instructors prioritized their personal experiences and

associated inclinations towards or away from pedagogical practices over literature evidence (Andrews & Lemons, 2015). Furthermore, instructors' personal experiences are also linked to the initial step of seeking out pedagogical innovations; it is instructors' dissatisfaction with the status quo, frequently their dissatisfaction with student learning outcomes and their own teaching practices, that drives instructors to consider different EBIPs (Andrews & Lemons, 2015; Feldman, 2000; Yik et al., 2022a).

All these studies focused on identifying factors related to the adoption or lack thereof of EBIPs. However, much can also be learned from exploring instructors' motivation to adopt pedagogical innovations that may not yet have broad empirical support for their effectiveness but are popular among instructors. Elucidating the decision process that leads instructors to choose this type of practice can provide valuable insight for the development of new instructional methods and the dissemination of already-existing EBIPs. An example of such a pedagogical innovation growing in popularity is specifications grading. Since it was first introduced in 2015 (Nilson, 2015), specifications grading is increasingly gaining interest in postsecondary Science, Technolgy, Engineering and Mathematics (STEM) education and has become the most represented alternative grading scheme in chemistry (Hackerson et al., 2024).

This relatively fast adoption of specifications grading is in contrast to the slow adoption among chemistry instructors of many EBIPs that have been disseminated for years and that have empirical evidence for their effectiveness. Given that few empirical studies exist on the impact of specifications grading on student learning, other factors or attributes of this innovation must compel chemistry instructors to its adoption. The goal of this study is thus to start elucidating the motivation of chemistry instructors to adopt specifications grading. In particular, we aim to explore the benefits of specifications grading chemistry instructors perceive to encourage them to use it

and the challenges they were aware of before their implementation by answering the following two research questions:

1. What are the benefits of specifications grading perceived by chemistry instructors who adopt it?
2. What are the challenges that chemistry instructors anticipated before they implemented specifications grading?

Specifications Grading

Extensive research has demonstrated that the traditional grading scheme (0-100% and A-F) has several limitations. Instructors have thus begun the shift away from traditional grading methods and towards alternative grading schemes such as mastery grading, specifications grading, and standard-based grading. Talbert and Clark (2023) have outlined four distinctive features of alternative grading methods, which shift the focus from evaluating students to promoting their learning process. These features include: *clearly defined standards, helpful feedback, representative marks, and reattempts without penalty*. *Clearly defined standards* include specific learning outcomes and detailed rubrics that outline what students need to achieve. These standards provide transparent expectations of students, and the paired *helpful feedback* provides students with actionable guidance on how to demonstrate their understanding or improve their work. After students complete an assignment, they receive marks (or scores) that clearly reflect their performance, such as “needs revision” or “exceeds expectations.” Finally, students are allowed to revise and resubmit their work without penalty, using the feedback and marks they’ve received to meet the established standards. This approach fosters a learning environment where improvement and mastery are prioritized (Clark & Talbert, 2023).

One of the alternative grading schemes that has gained increasing attention from chemistry researchers and educators is specifications grading. First formalized by Nilson (2015), specifications grading aligns with several of Clark and Talbert's criteria. Specifications grading evaluates assignments on a 2-level basis as determined by preset standards outlined to students. If the standards for an assignment are met, students receive credit on the assignment. Alternatively, if the standards are not met, students can receive meaningful feedback from the instructor and revise their work, though the opportunities for revisions may be limited per the instructor's preference or contextual limitations. Importantly, specifications grading requires that the standards be at minimum representative of what would be classified as B-level work under a traditional grading scheme. Thus, students cannot pass a course without producing at least some high-quality work. Finally, specifications grading schemes use bundles of assignments to correspond to letter grades (Nilson, 2015). In the bundling systems, students' final letter grades can be associated with the number of assessments in each type which meet the set standards, the completion of different assessment types to the set standards, or a combination of the two.

When the different features of specifications grading are combined, there are 15 anticipated outcomes (Nilson, 2015). These outcomes may be part of what motivates chemistry instructors to adopt specifications grading in their courses. Indeed, in the literature surrounding implementations of specifications grading in chemistry courses, Nilson's outcomes are frequently cited. In particular, instructors reference the motivators of lowering student stress (Mary E Anzovino et al., 2023; Kelz et al., 2023; Noell et al., 2023), increased learning outcomes (Ahlberg, 2021; Hunter, Pompano, & Tuchler, 2022; Noell et al., 2023), greater flexibility for students (Kelz et al., 2023), and reduced instructor grading time (Mary E Anzovino et al., 2023). However, these outcomes have not been systematically investigated in the literature until recently. Yik *et al.* (2024)

developed the first psychometric instrument aimed at measuring the extent to which several of these outcomes are met. They found that in comparison to traditionally graded courses, students from two general chemistry laboratory courses (n=~ 1,300) perceived that they were less anxious and were clearer on expectations in their specifications-graded course compared to their traditionally graded courses. However, no difference was observed in students' perceptions in terms of the three other learning outcomes (i.e., reflects student learning outcomes, useful feedback, and promotes motivation to learn). Interestingly, while another study had also found that students felt less anxious in a specifications-graded upper-level analytical chemistry course (Hunter, Pompano, & Tuchler, 2022), a third study focusing on the implementation of specifications grading in a general chemistry course found that students were more anxious (Noell et al., 2023). There are thus mixed results regarding the extent to which the hypothesized outcomes laid out in the book describing specifications grading and its implementation (Nilson, 2015) are realized.

Other motivating factors for its use may include prior publications on the impact of specifications grading on student learning. Few studies have reported on this potential outcome however and most were published within the last two-three years. For example, studies have reported that students receive higher letter grades (Bunnell et al., 2023; Katzman et al., 2021; McKnelly et al., 2023) and have positive attitudes toward their courses (Bunnell et al., 2023; Katzman et al., 2021). Additionally, previous literature has indicated an increased quality of interactions between instructors and students, with an emphasis on how students can better learn as opposed to receive better scores (Ahlberg, 2021; Bunnell et al., 2023; McKnelly et al., 2023).

Finally, instructors may also opt to try specifications grading because of their dissatisfaction with the traditional grading system. Indeed, several of the authors mentioned

various types of dissatisfactions in their articles describing their implementation of specifications grading. For example, they highlighted that the grades students achieve with the traditional grading scheme do not reflect their understanding of the course material (Mary E Anzovino et al., 2023; Howitz, McKnelly, & Link, 2021), students experience stress with the traditional grading scheme (Kelz et al., 2023; Noell et al., 2023), and there were issues with grading consistency across graders (Howitz, McKnelly, & Link, 2021) or due to partial credit (Martin, 2019).

Theoretical Framework

Within Rogers' Diffusion of Innovation (DOI) theory (2003), there are five key stages that describe an instructor's decision to adopt, or not adopt, a particular practice. The first stage involves instructors acquiring knowledge about a pedagogical innovation. Notably, this innovation is only required to be new to the instructor rather than 'new' in terms of its original inception. Next, during the persuasion stage, instructors evaluate various aspects of the innovation in light of some dissatisfaction they have with a current practice. They also take into account their personal context; these different factors come together to inform an instructor's opinion. Following this, instructors decide on whether or not to adopt the innovative practice. If they choose to adopt, instructors proceed to implement it. After implementation, they assess the outcomes and any associated costs, ultimately deciding whether to continue with the current approach, modify their implementation, or abandon the practice altogether. While these stages present a logical progression, they are highly interconnected; thus, the progression is not strictly linear. Moreover, as indicated in the confirmation stage, the process can be cyclical. Indeed, Rogers emphasized that the first three stages, in particular, are not a rigid pathway that individuals must follow (Rogers, 2003). Furthermore, Rogers (2003) identifies four key factors that influence the rate of innovation adoption: (1) the prior conditions and context of the individual before reaching the knowledge

stage, (2) the personal characteristics of the individual involved, (3) the attributes they perceive in the innovation, and (4) the communication channels that affect all stages of the cyclical, non-linear process.

As our interest lies in the features of specifications grading that lead instructors to its adoption, this study focuses on the perceived attributes of the innovation. Rogers identified five attributes which account for the majority of the variation in rate of adoption (Rogers, 2003): *relative advantage*, *compatibility*, *complexity*, *trialability*, and *observability*. *Relative advantage* describes "the degree to which an innovation is perceived as being better than the idea it supersedes" (Rogers, 2003, p. 212). *Compatibility* refers to how well the innovation is perceived to align with an individual's personal values, past experiences, and situational needs. *Complexity* describes how "difficult to understand and use" an innovation is (Rogers, 2003, p. 242). Finally, *trialability* describes "the degree to which an innovation may be experimented with on a limited basis" (Rogers, 2003, p. 243), and *observability* "is the degree to which the results of an innovation are visible to others" (Rogers, 2003, p. 244).

An additional part of Rogers' DOI model is the progression describing the adopters of innovations. Those who first create or attempt to implement an innovation are referred to as the innovators. Following this, early adopters will implement the practice. Once the pedagogical innovation has some evidence or anecdotal experience in a field, the early majority will adopt an innovation; this is then followed by the late majority. Finally, the laggards, who still comprise 16% of a population, will be the last to adopt an innovation. The ability to describe the targeted practitioners based on the stage of innovation adoption is vital to understand the practitioners' motivations, as there are distinct differences among the stages.

Methods

This study was conducted under Protocol #5936, which was approved by the University of Virginia Institutional Review Board for the Social and Behavioral Sciences.

Participants and data collection

To enable an in-depth analysis of potentially context-dependent motivations, only instructors utilizing specifications grading in their chemistry courses were recruited as participants. Furthermore, to increase the sample size of chemistry instructors, two pilot interviews were conducted with a biology and mathematics instructors who utilized specifications grading in their courses. These two pilot interviews were solely used to inform the design of the interview protocol and are not considered henceforth in the participant identification, participant pool, or data analysis sections. A combination of methods was used to identify potential chemistry instructors as participants in this study. Methods included available conference abstracts, journal article publications, snowball sampling, and social media posts. Participants were also identified through online searches, personal communications, and published book chapters.

Once identified, participants were contacted via email. These recruitment emails contained an invitation to participate in the study as well as a link to an online survey (see Appendix D for the pre-interview survey, interview protocol, and post-interview survey, respectively). In the online survey, participants were first asked to provide their consent to participate in the study. Subsequently, participants were asked to provide details about their current academic position, current and past teaching experiences, usage of specifications grading, and to upload relevant course artifacts (e.g., syllabi, assignment rubrics, and any other supplemental information provided to their students about the course structure and grading).

In all, 85 instructors were identified as potential participants in this study and were sent recruitment emails. In addition to the two instructors who did not teach chemistry and were used to pilot the semi-structured interview, 32 instructors agreed to participate and were interviewed by either BJY or HM. Of these 32 instructors, 3 were excluded from the sample for the following reasons: one instructor, who was a part of the biochemistry department at their institution, taught a molecular biology course, i.e., not a chemistry course. Another instructor, while belonging to a chemistry department and teaching chemistry courses, belonged to an institution which does not utilize the final A-F grading scheme. They were thusly removed from the sample due to their unique institutional context preventing the generalizability of their motivations. One additional instructor belonged to a chemistry department and taught chemistry courses, but they were removed from the sample due to confidentiality concerns and the sensitive nature of personal anecdotes, which were freely shared by the instructor during the interview. Thus, our final sample consists of 29 chemistry instructors teaching chemistry courses which use specifications grading at an institution that assigns students A-F grades after completion of a course. At the time the study was conducted, these 29 participants held appointments at 24 different institutions, which vary in type (Table 18). Each interview was audio recorded and transcribed verbatim using Temi, an automated transcription software.

Table 18. Instructors' academic ranks, teaching experiences, terms using specifications grading and demographics.

Category		Number of instructors
Academic Rank	Professor	4
	Associate Professor	7
	Assistant Professor	6
	Professor of Teaching/Practice	5
	Associate Professor of Teaching/Practice	4
	Assistant Professor of Teaching/Practice	1
	Lecturer or Instructor	2
Gender	Man	17
	Woman	11
	Agender	1
Race or Ethnicity	Non-Hispanic White or Euro-American	27
	Asian or Asian American	1
	Black, Afro-Caribbean, or African American	1
Years of Teaching Experience	2-4	5
	5-9	4
	10-14	13
	15+	8
Number of Terms Using Specifications Grading	1-3	7
	4-6	9
	7-9	6
	10+	7
Number of Unique Courses Taught with Specifications Grading	1	6
	2	10
	3	5
	4	5
	5	3
Total		**29**

Interview Protocol

As mentioned above, the perceived relative advantage of an innovation is a key attribute correlated with the likelihood of adoption according to the DOI framework (Rogers, 2003). In alignment with this framework, we investigated instructors' motivations for adopting specifications grading by exploring what they perceive as advantages and disadvantages of specifications grading through semi-structured interviews. In particular, we aim to identify these

instructors' perceived advantages of specifications grading that lead to their adoption. First, we asked instructors about their perceived benefits of adopting specifications grading, posing questions such as "*Why did you decide to use specifications grading?*" and *"What goals did you have when deciding to use specifications grading in this course?"* In their responses, instructors often spontaneously compared the characteristics of specifications grading with those of traditional grading schemes. For those who transitioned from traditional grading to specifications grading, additional questions were asked regarding the factors that led them to move away from traditional grading. This provided additional insight into their perceptions of the two grading schemes. Secondly, we asked questions about their anticipated challenges with the implementation of specifications grading, such as *"Before implementing specifications grading, what challenges or worries did you have?"* in order to understand the perceived disadvantages of specifications grading.

Data analysis

The transcripts were checked for accuracy and uploaded into NVIVO for initial review and memo creation as well as coding. Qualitative analytic memos were created by authors HM and BJY; these memos informed an initial codebook. Authors HM and YW subsequently reviewed a subset of the transcripts while using the initial codebook to refine codes, clarify definitions, and combine similar codes. HM and YW leveraged the DOI framework when creating the parent codes used in the codebook. Combined, these efforts resulted in a refined codebook. HM and YW then independently coded the transcripts at the paragraph level in three rounds. In the first round, 10 interviews were coded; a meeting between HM and YW achieved complete consensus and minor alterations to the codebook were made as detailed in Appendix E. The latter two rounds of independent analysis and attainment of complete consensus were performed with 10 and 9

interviews, respectively. After full consensus was reached between YW and HM, the two authors revisited all instances of coding across the 29 interview transcripts which utilized a code that had been altered between the creation of the refined codebook and the final codebook. The final codebook, complete with example quotes, is available in Appendix E.

Trustworthiness

Steps were taken throughout the data collection, data analysis, and sense-making processes to ensure credibility, transferability, and dependability (Shenton, 2004).

Credibility: The first steps to ensure credibility were taken during data collection as participants were asked to be truthful and assured of the study's confidentiality. Furthermore, data analysis by HM and YW occurred with intermittent debriefing sessions with authors MS and BJY. Sensemaking by authors HM and YW was enhanced by frequent debriefing sessions with their entire chemistry education research group at their institution.

Transferability: To maintain the transferability of the findings, care was taken in sample selection to ensure only instructors of chemistry who were teaching chemistry courses in higher education institutions which provide students with final letter grades were included. Furthermore, we aimed to provide more transparency regarding the characteristics of our sample, which includes reporting information regarding participants' demographics, institutions, and chemistry courses (Appendix E). Finally, thorough descriptions of the data collection and analysis processes are included herein and supplemented by Appendix E.

Dependability: The dependability of this study's findings is primarily established through utilizing iterative qualitative coding to reach a complete consensus regarding all transcripts. Furthermore, a detailed audit trail was kept beginning with the initial memoing by authors HM and BJY and continued through the completion of this manuscript. Thus, all alterations to the

collected memos were recorded. All iterations of the codebook are archived, and the alterations made during the inter-rater reliability process (e.g. verbal clarifications between coders, the combination of codes, deletion of unused codes, and expansions of definitions) are reported in Appendix E.

Results

This study provides the first empirical report of factors influencing chemistry instructors' decision to adopt specifications grading in their courses. Leveraging the Diffusion of Innovation theory (Rogers, 2003), we will first present instructors' perceived benefits that lead to integrating specifications grading over traditional grading in their practice, followed by the challenges they anticipated.

RQ 1: What are the benefits of specifications grading perceived by chemistry instructors who adopt it?

The instructors participating in this study cited an array of benefits, which were grouped into fifteen distinct codes. However, only the responses identifying benefits perceived by at least 10% of the sample are presented herein (Table 19).

Table 19: Chemistry instructors' perceived benefits of specifications grading

Code	Definition: Instructors want to implement specifications grading…	Number of Participants
Increased flexibility	in order to increase the flexibility available to their students (e.g. ability to miss assignments due to illness or sports, ability to revise or retake assignments)	20
Increased student learning gains	in order to increase student proficiency with and/or retention of the course material and related skills or to increase the rigor of the course or to increase students' focus on the learning process	13
Transparent expectations	because expectations will be clearer to students (e.g. what assignments are necessary, how to complete assignments)	8
Accommodating different groups of students	in order to increase equitable opportunities for all students as specifics grading can enable students with different cultural backgrounds or contexts to succeed in the course (e.g. differing education backgrounds among students)	8
Grades reflect students' proficiency	in order to have final grades that are representative of students' proficiency with course material and to enable instructors to accurately track their students' understanding of the material	8
Increased opportunities for feedback	in order to increase the frequency at which their students receive feedback and/or to ensure that the feedback students receive is meaningful or actionable	8
Alignment between course learning objectives, and assessments	because it enables better alignment of the course learning objectives and assessments	7
Increased student agency over learning	in order to increase student self-regulation (e.g., autonomy, metacognition, motivation)	6
Reduced student stress	in order to decrease their students' anxiety or stress	6
No partial credit	because it does not use partial credit on individual questions or assignments which are not wholly correct or not completed to standard	5
Reduced grading burden	in order to reduce the mental effort of grading on either themselves and/or the TA's	5
Reduced tension between instructor and student	in order to reduce instructor's perceived tension in the student-instructor relationship (e.g., move away from the instructor being a gatekeeper of the students' desired grade)	3

Increased flexibility. Over two-thirds (69%) of the interviewed instructors perceived specifications grading as offering more flexibility for students than the traditional grading system.

Indeed, 17% of instructors cited the lack of flexibility with traditional grading as a contributing factor in their switch to specifications grading. Instructors thought that specifications grading provides students with more opportunities to revise and retake their assignments or allows students to miss assignments or extend deadlines due to various reasons (e.g., illness or sports) without penalty. This is exemplified by instructor #117, who teaches a chemical and synthetic biology lecture course:

> "*I like that it [the token system] gives the students some built-in flexibility in the course so, then they can request a(n) assignment due-date extension, or, you know, they really just couldn't submit something. Maybe you let them trade in two tokens at the end to make up that assignment, especially if it's a complete/incomplete assignment.*[…] *And then you look at your evaluation system [traditional system] and it's like you're penalizing them for getting things in not on time. You're penalizing them for learning things later instead of earlier and, and you're penalizing them again and again*."

Increased students' learning gains. Approximately half of the interviewed instructors (45%) believe that specifications grading can increase students' learning gains in their courses. Notably, the increased learning gains are in reference to students' knowledge and skill base, as opposed to the numerical or symbolic grades students achieve. For example, instructor #184, who teaches an analytical chemistry lecture course mentioned that "*we were motivated to try specifications-based grading because of how much it incentivized mastery or at least proficiency*." Many instructors in this group also mentioned that specifications grading provides opportunities for students to focus on learning as well as develop knowledge and related skills due to its flexible mechanism, highlighting their perception of how one benefit can lead to another. As an example, instructor #118, who teaches an organic synthesis lab, cited specifications grading as "*a mechanism*

that says 'you're gonna try things. If you don't meet the mark, you get second chances, and that's gonna help reinforce these set of knowledge and skills you need so that you're ready for these more higher stake assessments towards the end'." Indeed, a third of the interviewed instructors (32%) think that the point-based nature of traditional grading can impede student learning. For instance, some instructors believe that a subpopulation of students will focus on their grades instead of learning the material. As a result, students can leave the class without a concrete understanding of the course. This is one of the reasons driving instructors away from traditional grading, as exemplified by instructor #120, who teaches an organic chemistry lab course:

> "*And at the end of the day, like, should there be enough points where the student literally doesn't have to pass any tests and they still get a B, a D in the class? Or a C in the class? That's probably problematic because you don't know if the students have actually learned anything. Um, so yeah. Yeah, I guess I do have a lot of issues with points-based.*"

Transparent expectations. Just under a third of the interviewed instructors (28%) emphasize that specifications grading offers greater transparency in term of learning expectations. They believe that specifications grading clarifies what students must learn and do to achieve their goals, including the required level of content knowledge, desired depth of understanding, and necessary assessments. As instructor #125, who teaches an analytical chemistry course, said:

> "*I was hoping that it could clarify the expectations for students in terms of what they needed to do to, to earn whatever grade they wanted; and also my expectations of them in terms of what learning outcomes I expected them to meet on each kind of assignment in the class.*"

Echoing this idea, a concern instructors (n=3) have with traditional grading is that it often leaves students uncertain about their progression, and thus their final grade, throughout the course

of the semester. For example, instructor #149, who teaches a general chemistry course, mentioned that, with traditional grading, they are unable to answer students' questions regarding their final course grades in the middle of the semester because students haven't completed all the exams and other assignments. Instructor #149 elaborates on this concern: "*if I have to guess, you could imagine student[s] have no idea what's going on with their grades and lots of speculation in it.*"

Accommodating different groups of students. Nearly a third of the interviewed instructors (28%) indicated their perception that specifications grading provides equitable opportunities for all students, enabling students with different backgrounds or contexts to succeed in the course. For example, instructor #152, who teaches a graduate-level organic chemistry course, explained how specifications grading allows them to adapt their instruction to students with different preparation in organic chemistry:

> "*Students that had a strong organic background could just go through and do everything … the first try and then the students that needed more time, I would be able to give them more retries and feedback. And so, it would help me to differentiate the class a little bit more … and tailor to the different populations we have in there.*"

Correspondingly, 7% of the instructors perceived that traditional grading only benefits some of the students, as exemplified by instructor #189, who teaches an organic chemistry lecture: "*I also feel like a lot of the times our traditional grading is rewarding people that can test well.*"

Grades reflect student proficiency. Roughly one out of every three interviewed instructors (28%) indicated that they wanted to adopt specifications grading so that final grades would be representative of students' proficiency with course material, thus enabling instructors to

accurately measure students' understanding. This is exemplified by instructor #171, who teaches an organic chemistry lecture:

> "*I just started thinking more about what a grade means and something more tangible to say, 'okay, a student who left my class having earned a C, they can do these core things.' And so I could say, yeah, that basically anyone who passed a class C or higher can do these core things and, you know, B students can do some number more, A students can do pretty much everything*."

Related sentiments were captured in instructors' comments about traditional grading. Specifically, instructors (n=15) did not believe that grades in traditional grading schemes reflect the student's knowledge and skills. As instructor #184 indicated,

> "*We realized that several students were not actually reaching mastery or proficiency on really important concepts that we knew would be important in future chemistry courses. And our letter grades at the end of the semester were indicating that that was okay, when we knew that it actually wasn't.*"

Notably, this perceived benefit of specifications grading and disadvantage of traditional grading focuses on the accurate measurement of student learning, as opposed to the prior benefit of increased learning gains which explores the relative amount of student learning itself.

Increased opportunities for feedback. Slightly less than a third of the interviewed instructors (28%) believed that specifications grading allows them to provide more frequent and actionable feedback to students. Some of these instructors also mentioned that the feedback can help them gauge their students' learning progress. For example, instructor #186, who teaches a

general chemistry course, indicated that both students and instructors benefit from the feedback provided in specifications grading on assessments that closely align with the learning objectives. They explain that under a traditional system, instructors invest additional time and effort into understanding how the feedback given corresponds to the course aims:

> "*If you give a large exam, right, 'yeah, they didn't do well,' but now you've really gotta drill down, 'okay, they didn't do well on questions one through three. That means it's this, you know, this topic.'"*

This contrasts with their perception of specifications grading, which is more streamlined for both instructors and students:

> *"Whereas, you know, specs grading, you sort of get that immediate feedback on this objective or this series of objectives, this one concept. And so, you know, if I'm getting that data out of that exam … you know, out of that, then the students also can get it*."

Alignment between course learning objectives, and assessments. Some of our interviewed instructors (24%) believe that specifications grading enables the alignment between their learning objectives and assessments. For example, instructor #163, who teaches a general chemistry course, talked about the opportunity that specifications grading provides for them to reflect on how their instructional practices, including assessment, are aligned with the big ideas of the course:

> "*It made me take another look at what I was teaching. And then it was like, 'Okay, well here's my big ideas, and this big idea. I only have one module and one assessment. And for other big ideas, I might have more modules and more assessments.' So it also allowed me*

to reflect on that and adjust, so that if I considered something important enough to be a big idea that I was assessing it and covering it in as much detail as I would … something else. So that allowed me to go back and like look at what I was covering and reflect on how much time and emphasis I was putting on certain things."

Additionally, instructor #186 also commented that students don't always make the connection between the exam questions and specific learning objectives; however, specifications grading helps make this connection more transparent due to "*the one-to-one alignment between the assessment and the learning objectives.*"

Increased student agency over learning. 21% of the interviewed instructors thought that specifications grading can increase student agency over learning, such as having more self-regulation skills, feeling the autonomy or motivation to learn, and developing more metacognition skills. For example, instructor #102, who teaches an organic chemistry course, believed that specifications grading can be "*improving their (students') agency over their own learning*", which is a "*critical component of student success.*" They went on to explain "*I think that was what caused me to get started on it. That's why I'm still doing it.*" Additionally, instructor #154, who teaches an instrumental method course, valued the idea of metacognition and commented that the alignment between assessment questions and learning objectives allows students to be aware of what they are learning, therefore "*specifications grading was an idea where it was a way of getting students to do metacognition without trying to explain to them what metacognition is*."

Reduced student stress. Two out of every five (21%) interviewed instructors mentioned that they believe specifications grading can decrease students' stress or anxiety. This may be due to the adoption of more low-stake assessments, which focus students more on the learning process

instead of grades, as well as students being offered opportunities to retake assignments, which reduces the worry associated with making mistakes. For example, instructor #118 believed that specifications grading can help cultivate students' growth mindset, which decreases their anxiety levels, because

> "*they start to realize 'I can make mistakes and it's not in (an) immediate penalty.' And so instead of students viewing each assignment as an opportunity for their grade to go down, it's now viewed as every assignment's an opportunity for me to show growth and improvement. And so that mentality shift I think is then what couples with them feeling more positively about the course and about what they're learning*."

Indeed, several instructors (21%) cited "traditional grading increasing student stress" as one of the reasons for them to switch to specifications grading. These instructors mainly linked students' anxiety to the high-stake(s) nature of the exams in traditional grading. This is exemplified by instructor #120:

> "*And our method of assessment is, 'okay, here's a test' and then the test should be, 'okay, I'm gonna measure and see if you actually meet these learning outcomes.' … but then that makes the class rather high stakes, right? Everything in the class, their whole grade is based on the test. Which is another issue too, right? Because then you just have all these high-stake(s) things and it's stressful.*"

No partial credit. Just under a fifth (17%) of the interviewed instructors underscored the benefit of having no partial credit on individual assignments or questions under specifications grading. For example, instructor #173, who teaches a biochemistry course, stated that

"*one of the things that was appealing to me with specifications grading is getting rid of that whole idea of partial credit and having to decide, okay, you know, did this person get 9 points out of 10 and things like that*."

Several instructors (17%) also mentioned their dissatisfaction with traditional grading in terms of having to assign partial credit to students, which often leads to students' arguing for points. This is exemplified by the same instructor (#173):

"*a lot of students who are taking chemistry, I think are highly anxious students. They're very grade driven, even though I'm trying to, trying to get rid of that focus. But that still doesn't seem to always happen. So this very much grade-focus point-grabbing, partial credit, you know, sorts of things, was becoming very tiresome*."

Reduced grading burden. 17% of the interviewed instructors reported that they expect specifications grading to reduce the mental effort of grading on themselves or the teaching assistants (TAs). For instance, instructor #163 mentioned that they would assign more project-based assignments in specifications grading, instead of multiple-choice questions. The specifications grading approach was perceived to help them streamline the grading process, especially as they often grade by hand. In addition, instructor #180, who taught a multi-section chemistry laboratory facilitated by multiple TAs, mentioned that "*students don't wanna spend too much time on labs, so, they were always not doing what they were supposed to. So, the TAs were spending all this time grading*." However, specifications grading can focus students and graders on what is important, thus, "*it allowed the TAs to spend less time grading cause everything was very standardized, amongst all sections and between all the labs."*

Reduced tension between instructors and students. One tenth of the interviewed instructors explained that specifications grading can help ease the tension in the student-instructor relationship. Notably, this tension was linked to students' intensive focus on grades and partial credit under the traditional grading scheme, which was discussed by 28% of the interviewed instructors. For instance, instructor #180, who teaches a quantitative analysis and methods lab course, stated that

> *"[specifications grading] also was a way to de-emphasize the grade focus for our pre-med students who are primarily taking this course. And it allowed them to say, 'Hey, you know, …, just focus on doing the things.' Like … we don't want you to be asking about sig figs, … we don't want you to be, you know, those are important, but that's, I don't want you to be arguing those points every single lab*."

Additionally, instructor #117 detailed the different roles they perceived themselves to have in the two grading schemes. Specifically, they saw themselves as "*the gatekeeper*" and "*the withholder of points*" when using the traditional grading scheme to get rid of points from students. However, using specifications grading, this instructor can "*feel much more like their coach or their ally."* Instructor #117 wanted to convey this message to students: "*I want to help, and I will give you information that helps you to make progress*." The instructor does not feel that they are pointing out students' weaknesses and comparing students to a standard in specifications grading. Instead, they perceive the relationship as more supportive and less adversarial.

RQ2: What are the challenges that chemistry instructors anticipated before they implemented specifications grading?

The instructors participating in the study generally cited far fewer challenges than benefits when elaborating on their decision to implement specifications grading. Indeed, only seven distinct

codes were identified to describe the anticipated challenges before implementing the grading scheme. However, only the responses identifying challenges present within at least 10% of the sample are presented herein (Table 20).

Table 20: Chemistry instructors' perceived challenges of specifications grading

Code	Definition: Instructors are concerned…	Number of Participants
Increased instructor workload	about having an increased workload in terms of time commitment or mental effort (e.g., spending more time creating assignments, more time grading due to retakes, translation into letter grades)	19
Student resistance	that students will actively not buy-in to specs grading and will choose to resist it	11
Unfamiliar system for students	that the system is unfamiliar to the students and that the students may not be able to understand the system	9
Maintaining rigor	about maintaining the rigor of their course when using specs grading	4
Lack of support	about the lack of support they may have from their peers/chair/department/institution if they start specs grading	3

Increased instructor workload. More than half of the interviewed instructors (66%) reported their worries about the increased workload in implementing specifications grading. The major concern was their time commitment or mental effort in figuring out the logistics of specifications grading. This includes making decisions on the criteria for student proficiency and translating students' level of proficiency to letter grades. For example, instructor #188, who teaches a general chemistry lab course, stated, "*One was logistical, just figuring out how to make … those buckets of … what's gonna make a bucket and how is that bucket gonna turn into a letter grade at the end of the semester*." Additionally, their concern also includes writing different versions of exams for students to retake. For example, instructor #174, who teaches a biochemistry course, mentioned that

"*So if I'm doing specifications grading, the way that I do it is with flexible deadlines. And so students are taking different quizzes at different times. You know, maintaining the confidential nature of the assessment pieces … was another concern*."

Instructors were also worried about the time burden associated with regrades, as instructor #119, who teaches an organic chemistry course, expressed:

"*Yeah, I was really worried. So like thinking back to that first exposure in 2018, it just seemed like so much work 'cause you're just allowing so many retries and you have to regrade. And I was just worried that it was gonna end up being more work for me."*

Student resistance. 38% of the interviewed instructors were concerned about how they could get students to buy into specifications grading. Several instructors anticipated a negative student attitude towards specifications grading because students may prefer the point-based grading scheme they are accustomed to. For example, instructor #120 stated that

"*The biggest drawback for me has been trying to establish student buy-in, just because students are habituated to other types of grading systems, most commonly points-based grading systems. And there seems to be a lot of student resistance to the new system, and they tend to blame them not doing well or them not understanding how the grading system works as well - just to the fact that it's a new system and it doesn't work very well*."

Other instructors attributed this concern to students' unfamiliarity with specifications grading, which will be unpacked in the following subsection.

Unfamiliar system for students. Around a third of the interviewed instructors (31%) mentioned that the specifications grading system is a novel system that students are not familiar

with, and they may not understand how the system works. For example, instructor #145, who teaches a general chemistry course, mentioned, "*I mean, it's complicated to explain to students who have never seen it before … and probably even students who have seen something similar before*."

Maintaining rigor. 14% of the interviewed instructors reported concern about maintaining the rigor of their course when using specifications grading. This is exemplified by instructor #186:

> "*I think my biggest concern then, and arguably, I think now even still, is okay, at what point are we really sure after, I don't know, five retakes, right? Did the student just memorize the concept to pass the objective? Can they truly apply it? Right? Do, have they actually learned it? Can they truly apply it down the road? Is a concern.*"

Lack of support. One-tenth of the interviewed instructors were apprehensive about potential push-back or a lack of support from their colleagues and institutions when implementing specifications grading. For instance, instructor #184 mentioned that

> "*we had a little bit of concern, but not too much about how the department head or administrators might like, if there might be push-back from that. But … we got buy-in from them relatively early in the process of planning so that it wasn't really too much of a concern*."

Discussion

Chemistry instructors adopted specifications grading due to the perceived relative advantages of specifications grading over traditional grading

Although instructors were not specifically asked to compare specifications grading with traditional grading in their interviews, many of them spontaneously discussed these comparisons as they explained their reasons for adopting specifications grading and, thus, moving away from traditional grading. At an aggregate level, the perceived advantages of specifications grading are often associated with corresponding dissatisfactions with traditional grading (Figure 12).

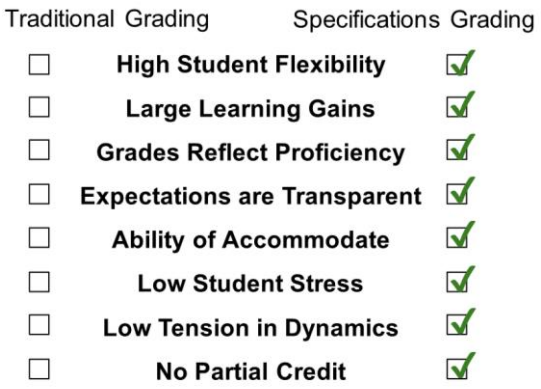

Figure 12. Instructors' perceived relative advantages of specifications grading

Instructors appear to seek out a new grading scheme due to their dissatisfaction with traditional grading and implement specifications grading because it addresses their concerns. This aligns with the claim in the DOI framework, which suggests that the innovators (i.e. instructors) do consider the attribute of "relative advantage" before adopting an innovation (Rogers, 2003). Additionally, it also resonates with results from a previous case study with biology instructors, where the sense of dissatisfaction was found to be a necessary prior condition that led instructors to change their teaching (Andrews & Lemons, 2015). In our study, instructors believed that specifications grading, as compared to traditional grading, increases flexibility, student learning gains, and transparency of expectations. It is also perceived to reduce student stress, instructor grading burden, and tension between instructors and students. Additionally, instructors noted that

the design of specifications grading, which does not allow for partial credit, results in grades that more accurately reflect student proficiency than in traditional grading. Importantly, instructors view specifications grading as allowing for more opportunities to accommodate students from diverse backgrounds and contexts, which they often see as lacking in traditional grading schemes. Previous literature has shown that traditional grading schemes may be representative of factors such as students' zip code and access to tutoring, as opposed to representing their knowledge or skills (Feldman, 2019a, 2019b; Link & Guskey, 2019; Matz et al., 2017; McKay, 2019). This is corroborated by work which highlights that grades do not always correspond to job performance (Cain et al., 2022). Furthermore, students' grades under traditional grading can be a reflection of the instructor rather than of the students' learning. An individual instructor's leniency and grading criteria can result in different grades for the same quality of work (Cain et al., 2022; Donaldson & Gray, 2012; Herridge & Talanquer, 2020; Herridge, Tashiro, & Talanquer, 2021), and instructors' implicit and unconscious biases affect the grades assigned to students (Feldman, 2019b). Our findings indicate that instructors are aware of such issues with traditional grading, which can contribute to their decision to implement specifications grading.

Flexibility and capability of increasing learning gains are perceived as major relative advantages of specifications grading

While instructors recognized the majority of Nilson's hypothesized benefits of specifications grading, flexibility emerged as the most frequently perceived benefit. Instructors often stated that flexibility can result in other benefits, such as reducing students' stress and providing accommodations for different groups of students. Flexibility is often provided in specifications grading through the student's ability to revise their work. The hallmark of allowing for revisions in specifications grading is associated with four of Nilson's hypothesized outcomes:

reducing student stress, discouraging cheating, minimizing conflict between students and instructors, and providing feedback to students that they will use. Nilson draws a clear connection between the ability to revise work and the reduction of student stress, as students have a "safety net" if they make mistakes. This safety net is also connected to discouraging students from cheating as they will experience less pressure when submitting work to be graded. Through this reduction in academic dishonesty, there is an associated reduction in conflict between students and the instructor, which is furthered by students choosing to revise work as opposed to arguing for points (Nilson, 2015). The instructors we interviewed further explained that the ability to miss or revise assignments enables students with different contexts to have a path towards academic success that fits their other responsibilities, such as caretaking or working, and their non-academic needs, such as attending doctor appointments. Thus, given the widely perceived relative advantage of flexibility offered by specifications grading and the other benefits that stem from this characteristic, flexibility can act as a key feature when promoting the adoption of specifications grading.

Increased student learning gains are seen as another major benefit of specifications grading. The anticipated increased learning gains are the result of a combination of factors. Primarily, specifications grading emphasizes the mastery of learning objectives, compared to traditional grading which emphasizes the earning of points. Furthermore, specifications grading aims for increased transparency which is deemed to enable students to understand exactly what is expected of them and to plan their studying appropriately. The previously mentioned flexibility also plays a role as opportunities like revisions are expected to promote a growth mindset while encouraging students to engage with the material. The iterative process students engage in under specifications grading is posited to also contribute to knowledge retention. Ultimately, the enhanced learning

gains are the result of students focusing on clearly defined aims and goals while being able to learn from their mistakes.

Chemistry instructors' perceived benefits of specifications grading align with hypothesized, yet untested, outcomes of specification grading.

The majority of instructors' perceived benefits of specifications grading are well aligned with the hypothesized outcomes of specifications grading proposed by Nilson. This alignment is perhaps to be expected. As Nilson proposed these outcomes when writing the first formalization of specifications grading, these hypothesized results are closely associated with specifications grading as a practice.

Recent work examined students' perceptions of specifications grading as it related to the student-centered hypothesized outcomes laid out in Nilson's book. While the students saw some benefits, namely reduced anxiety and clearer expectations, students did not perceive any difference in the alignment between grades and learning outcomes or in the feedback they received. Furthermore, students actually expressed that they felt a decreased motivation to learn in specifications grading as compared to traditional grading (Yik et al., 2024). The disparity in how specifications grading is perceived by instructors and by their students, paired with the disparity in instructor motivation to increase flexibility and accommodations and in the mixed empirical evidence, indicates a rich area for future investigations.

Typical deterrents to the adoption of EBIPs (time and student resistance) also concern adopters of specifications grading but not to the point of preventing adoption

In her book, Nilson hypothesized several instructor-centered outcomes, one of which is the benefit of saving faculty time (Nilson, 2015). Time is a typical factor mentioned by instructors

when ask about barriers to the implementation of EBIPs (Sturtevant & Wheeler, 2019). Nilson argued that the clear passing criteria and simplified framework for marking assessments would streamline the grading process, thus reducing the time instructors spend evaluating student work. However, the interviewed instructors worried that adopting specifications grading would increase their devoted time and mental effort. Specifically, instructors find it mentally challenging to decide the cut-offs for marks on assignments, as well as to take said marks and translate them into a final letter grade for the course. More importantly, the flexibility of "retake and resubmit" brings in instructors' concerns about maintaining a manageable workload while ensuring their assessments are not distributed in a way that would enable cheating. Instructors felt the need to write different versions of quizzes that cover the same learning objectives, which may eventually require more time commitment. This concern is not unique to our study, as skepticism about whether specifications grading saves faculty time has been in the published literature since the inception of specifications grading (Prescott, 2015). The process of aligning assessments directly with specific learning outcomes, while helpful for instructors (Walden, 2022), does add a burden to those first implementing specifications grading (Carlisle, 2020; Shields, Denlinger, & Webb, 2019; Williams, 2018). In addition to the potential barrier of additional effort being required during pre-term planning, instructors using specifications grading may also spend more time going through detailed feedback with students during the term (Lovell, 2018).

Additionally, there are conflicting accounts concerning the change in grading workload transitioning from traditional grading to specifications grading, with some reporting an overall decrease in time and effort spent grading (Elkins, 2016; Lovell, 2018; Mendez, 2018b; Mirsky, 2018; Williams, 2018) and others experiencing no change or an increased grading effort (Carlisle, 2020); (Hunter, Pompano, & Tuchler, 2022; Shields, Denlinger, & Webb, 2019; Spurlock, 2023).

Thus, future endeavors in disseminating and promoting specifications grading should address these valid concerns. Clear instructions on designing a specifications-grading course are needed to support instructors' adoption and continued implementation.

Furthermore, students' resistance or unfamiliarity with the system is perceived as another major challenge. Instructors are often worried that students may have a negative attitude toward specifications grading as they are used to the traditional grading scheme, which offers partial credit. This is not surprising as student resistance is a relatively common concern with pedagogical innovations and practices (Bentley, Kennedy, & Semsar, 2011; DeMonbrun et al., 2017; Genné-Bacon, Wilks, & Bascom-Slack, 2020; Lake, 2001; York & Orgill, 2023). Indeed, student resistance to specifications grading has been documented, with students resistant due to the nature of the grades they receive (Graves, 2023; McKnelly et al., 2023). Furthermore, students have been shown to need time to understand the specifications grading system (Howitz, McKnelly, & Link, 2021; Williams, 2018). Research on the adoption of pedagogical innovations has shown that the challenge of student resistance can be mitigated by clear explanations and active facilitation on the part of instructors (Tharayil et al., 2018). These strategies can be adapted for specifications grading and provided to instructors adopting the grading system if they encounter student resistance. For example, should students be opposed to accepting specifications grading, instructors could clearly articulate the purpose of using specifications grading in their course. Additionally, the instructors could encourage students to ask questions about the grading system, ensure a clearly communicated and consistent grading and re-grading routine, and continually encourage students to strive for success throughout the course.

Notably, these commonly cited concerns about, or barriers to, specifications grading have caused instructors not to adopt various EBIPs (Brownell & Tanner, 2012). The fact that these

barriers are not preventing our participants from adopting specifications grading may be due to the unique characteristics of our sample. Indeed, given the recent formalization of specifications grading, it is possible that the instructors currently using the grading system are innovators or early adopters. As highlighted by Rogers, it is possible that our sample is more inclined towards risk-taking, comfortable with navigating adversity, and/or confident in challenging the status quo (Rogers & Shoemaker, 1971). Additionally, our sample is strongly averse to traditional grading methods, and this intense dissatisfaction may encourage them to attempt an alternative grading system.

Implications

Professional development

To effectively motivate instructors to adopt innovative pedagogical practices, it is essential to target their underlying dissatisfaction with the status quo. Instructors have shown a willingness to accept challenges and potential resistance when the relative advantage is great enough. Indeed, the observed willingness to confront the challenges associated with specifications grading suggests that when a practice is believed to directly address their needs – such as a need to increase flexibility for their students – they are more likely to introduce the practice into their courses. This combination of dissatisfaction and direct, explicit alignment between innovation and an instructor's values or needs should be a primary component of advocacy for pedagogical innovations. This indicates a need for a strategic shift in how pedagogical practices should be presented; rather than focusing majorly or even solely on empirical evidence, we should instead highlight the advantages that clearly address instructors' real-world concerns. When promoting and providing training on specifications grading, it is vital to highlight the increased flexibility

given to students and the focus on learning as it clearly addresses chemistry instructors' dissatisfaction with traditional grading.

Research agenda

Continued efforts to effectively improve the implementation of pedagogical innovation must be rooted in a better understanding of the real-world needs of instructors. In the current study, the increased flexibility for students and increased student learning gains appeared to be aligned with instructors' need for grading schemes, thus becoming large motivators for instructors to overcome the challenges and attempt pedagogical change. Despite the valuable insights this work provided, our sample is probably composed of innovators and early adopters, meaning the major motivators for our sample to undergo pedagogical change may not translate to the early majority, late majority, or laggards. Thus, future research can focus on exploring the propagation of specifications grading within postsecondary chemistry courses to determine effective dissemination strategies across the different stages of adopters. Through studying the spread of specifications grading, it may be possible to determine the relative advantages of pedagogical innovations that resonate most with instructors at each stage. Such insight can enable effective and adaptive advocation for EBIPs.

Limitations

The sample size in this exploratory, qualitative study inherently limits the generalizability of our findings. However, we intentionally sought to recruit a diverse range of post-secondary chemistry instructors across the U.S. who utilize specifications grading, aiming to provide rich and transferable data. Although caution should be taken in drawing broader conclusions with this data, the data may provide informative information for other instructors to make decisions about implementing specifications grading in their chemistry courses. A further limitation is due to the

likely characteristics of our sample population. The instructors interviewed are likely innovators or early adopters, as described by Rogers. Thus, their motivations may differ from the early majority, late majority, and laggard adopters due to their relative comfort with risk-taking and handling adversity. Finally, the study only reflects the perceptions of the instructors who have already implemented specifications grading. The perceived challenges they reported may differ from those perceived by instructors who have not yet adopted this approach, as the latter group may face distinct barriers preventing them from doing so. Further research is needed to explore the perspectives of this group to gain a more comprehensive understanding of potential challenges related to the adoption of specifications grading.

Conclusion

The current study explored chemistry instructors' perceived relative advantages (i.e. benefits and challenges) of specifications grading that are linked to their decision to adopt this practice, drawing on Rogers' Diffusion of Innovations (DOI) theory. The results demonstrate that instructors adopt specifications grading due to their perceptions of its advantages over traditional grading, primarily increased flexibility and improved student learning gains, despite their concerns about potentially increased workload and student resistance and in spite of a lack of evidence for increased learning gains. This work provides valuable insights for future dissemination efforts aimed at chemistry instructors who are considering implementing specifications grading. Specifically, to encourage broader adoption, dissemination efforts should emphasize how perceived benefits, even if not yet empirically supported, align with instructors' dissatisfaction with the status quo and relate to their real-world needs and aspirations for their classroom.

Chapter 4. An investigation into the implementation of specifications grading in chemistry courses

This chapter is adapted from a soon-to-be-submitted manuscript.

Introduction

Receiving a points- or percentage-based grade is a ubiquitous experience in western schools, from primary through post-secondary education. However, the practice of assigning such grades has not always been the standard. Performance at academic institutions has always been assessed; however, this assessment was originally a determination of whether a student had "mastered [their studies] to a level comparable to and determined by other masters" during a leaving exam (Williams, 2022, p. 8). Indeed, such examinations are still in use today as seen in doctoral defenses. A shift away from this classical measure of learning occurred in the 18th and 19th centuries. As academia grew and a desire to compare students' performances emerged, universities began to implement grades or simple marks (e.g. Senior Optimes, Junior Optimes), and slowly incorporated more frequent assessment (Clark, 2019; Schneider & Hutt, 2014). By the mid 1900's, this scheme had evolved into the more granulated and formalized A-F system, which is still in use today (Clark, 2019; Schneider & Hutt, 2014). It is at this time that such grades also "generally aligned with numerical values —an A reflecting work between 90 and 100, for instance, and a B reflecting work between 80 and 89" (Schneider & Hutt, 2014, p. 215). This emergent grading scheme, called traditional grading, thus encapsulates the "assigning of points to one-time assessments and aggregating those points into a letter grade for the course" (Clark & Talbert, 2023, p. 11).

While such grades undoubtedly have use as means of interinstitutional communications, there are numerous characteristics of traditional grading that have been critiqued, and instructors

have begun to move toward alternative grading methods. Specifications grading is currently the dominant alternative grading scheme used in postsecondary chemistry courses. Specifications grading was first formalized by Nilson (2015) and combines aspects of mastery learning (Diegelman-Parente, 2011; Kulik, Kulik, & Bangert-Drowns, 1990; Winget & Persky, 2022), competency-based learning (Diegelman-Parente, 2011; Gervais, 2016; Ying et al., 2023), and contract grading (Danielewicz & Elbow, 2009; Harrington et al., 2024; Inoue, 2019; Offerdahl, Hodgson, & Krupke, 2016). In specifications grading, each assignment has set criteria, or *specifications*, that must be met in order to receive credit for the assignment (Nilson, 2015). Assignments are graded on a two-level, pass/fail basis. Assignment specifications align with course learning objectives and are meant to represent at least B-level work; these specifications are typically formalized in the form of rubrics where specifications correspond to rubric criteria. To achieve this higher-level of work, students are provided with opportunities to retake or resubmit assignments after receiving meaningful feedback. Assignments are bundled together to determine course-level final letter grades (Nilson, 2015).

The bundling of assignments can be done in many ways. One model of bundling involves a different number of assessments that meet specifications. For example, there are three different assessment types and meeting specifications on a different number of each assessment type yields a different letter grade; the highest-level bundle where all specifications are met determines the course letter grade. A second model of bundling involves different assessment types. For example, students must complete and meet specifications for a different assessment type to earn higher letter grades. A third mode of bundling involves a combination of the previous two models. Here, a hybrid number of assessment types and numbers determines the final letter grade. The distinction between bundles, and thus between the final grades, is determined by the number of specifications,

the assignment types that met specifications, and/or whether the specifications met align with foundational or additional learning objectives.

In addition to the bundling system, specifications grading is unique in the alternative grading world due to the use of tokens. Tokens act as a form of currency for students within specifications-graded courses; these tokens can be provided to students or earned by students through completing surveys, assignments, or other course-relevant tasks. Once students have either received or earned tokens, they can be exchanged for various purposes, including receiving the ability to reattempt assignments, excusing absences, receiving deadline extensions, etc. Thus, specifications-graded courses often use token systems to address the fourth pillar of alternative grading (reattempts without penalty); however, tokens are not a requirement for the fourth pillar to be implemented in specification grading schemes. While not all specifications grading systems use tokens, the token system itself is designed to provide additional inbuilt flexibility for students. Additionally, token systems can increase student autonomy and choice as the students will be the ones deciding when and how to use their tokens.

Specifications grading thus has six main distinguishing features (Table 21): (1) individual assignments are graded on a pass/fail basis, (2) students are provided clear specifications for each assignment, (3) specifications reflect the standards of B-level or better work, (4) students are provided opportunities to revise and resubmit work, (5) final grades are determined through a bundling system as opposed to a weighted average, and (6) bundles are aligned with the course learning outcomes (Nilson, 2015). When implemented completely and correctly, Nilson claims that specifications grading achieves 15 outcomes; many of these outcomes align with Talbert and Clark's (2023) four pillars of alternative grading. It is important to note that these 15 outcomes are

hypothesized by Nilson. For an empirical measure of some of these hypothesized outcomes, refer to chapter 5.

Table 21. Critical components of specifications grading

Component	Description
2-Level System	All grading is done on a pass/fail basis
Clear Expectations	Students are explicitly aware of the expectations for meeting specifications on assignments
B-Level Work	Meeting specifications on an assignment is indicative of B-level work or better
Revisions	Students have the ability to revise work in order to meet specifications and receive credit on assignments
Bundling Grade Determination	Final grades are determined through a bundling system rather than a weighted average
Bundles Based on Learning Outcomes	The bundling system is such that increasing final grades are representative of an increasing number of learning outcomes achieved

While Nilson's book provides an excellent guide on the implementation of specifications grading, the nature of the discipline, the particular student population, instructors' pedagogical beliefs, and instructors' departmental and institutional context can lead to variations in implementation (Nilson, 2015). Notably, Nilson recognized that hybrid models of specifications grading are inevitable with an array of different instructional contexts. These hybrid forms were hypothesized in her book to incorporate some elements of traditional grading, such assigning numerical scores to communicate students' performance, and to include some adjustments to the grading scale in order to comply with institutional policies. However, these hybrid forms were still described to align with the major characteristics of specifications grading, such as grading being based on meeting different learning outcomes and grading scales communicating pre-defined standards (Nilson, 2015).

Variations in the implementation of specifications grading are known to exist within disciplines and within institutions. Tsoi and colleagues (2019) examined the implementations by twelve instructors across four STEM disciplines at a single institution. They found variations in

implementation of specifications grading based on: 1) the utilization and/or grouping of learning objectives for determining mastery of material, 2) how the bundling of specifications aligns with course grades, 3) how individual assessments were graded, 4) the strategies used to support students (such as tokens), and 5) utilization of final exams and/or how final exams contribute to course grades. Harrington et al. (2024) examined the implementations of specifications and contract grading in a single discipline – postsecondary computer science education. They report marked differences in categorical grading systems (i.e. whether or not binary grading is used); the ability of students' to revise work, how students' can go about such revisions, and in the labeling as specifications grading, contract grading, or a combination of the two (Harrington et al., 2024).

Differences in implementations of specifications grading are also present within the chemistry education community. In one large-enrollment, organic chemistry laboratory course, the way quizzes factor into the bundling of grades is points-based, with the minimum score for earning a C being 75%, and either a B or an A being 85% (McKnelly et al., 2023). An additional 1,000+ student enrollment general chemistry course instead assigns a cut-off of 80% on their quizzes as meeting the expectations for the assignment; students must then meet this expectation on a certain number of quizzes in order to earn the different letter grades (Yik et al., 2024). Notably, this is only one example of many present in the literature. The variations continue with some chemistry instructors implementing a solely specifications-based grading scheme (Bunnell et al., 2023; Howitz et al., 2023; Howitz, McKnelly, & Link, 2021; Hunter, Pompano, & Tuchler, 2022; Kelz et al., 2023; McKnelly et al., 2023; Ring, 2017; Saluga et al., 2023) whereas other instructors combine specifications grading with another alternative grading scheme (Toledo & Dubas, 2017). Moreover, some chemistry instructors choose to implement a hybrid specifications grading scheme

by combining specifications grading and traditional grading (Ahlberg, 2021; Donato & Marsh, 2023; Houseknecht & Bates, 2020; Martin, 2019; Noell et al., 2023).

Prior research on the adoption of evidence based instructional practices has demonstrated that variations in implementing a practice can affect the fidelity of implementation (Andrews et al., 2011; Chase, Pakhira, & Stains, 2013; Stains & Vickrey, 2017; Turpen & Finkelstein, 2009). This is particularly salient in specifications grading as aspects of implementation have been hypothetically associated, though not empirically linked, to specific anticipated outcomes (Table, 1; Nilson, 2015). Indeed, in a recent work investigating students' perception of specifications grading, it was found that a specifications grading scheme which closely follows Nilson's described characteristics did not achieve all of the tested outcomes (Yik et al., 2024). While there was a decrease in anxiety and an increased understanding of the expectations as compared to traditionally graded courses, the students indicated a decreased motivation to learn (Yik et al., 2024). Such work highlights the potential necessity of associating different characteristics of implementation with the observed outcomes. However, the variety of methods of implementation of specifications grading is not yet understood.

A recent study looked at the implementations of specifications grading in higher education to understand how it is being implemented. The authors examined published accounts of specifications grading in the literature in a variety of disciplines. They report on publication trends; impacts on student performance, stress and anxiety, and attitudes; themes such as time investment required; and characterized four primary components: "grade bundles, rubrics with specifications and defined passing thresholds, opportunities to revise and resubmit work, and a token system" (Howitz, McKnelly, & Link, 2025). They found that grade bundles can be developed according to different approaches to learning outcomes and whether there was a combination of core learning

outcomes and additional learning outcomes, whether all learning outcomes were treated equally, whether learning outcomes were grouped into modules, and whether learning outcomes needed to be demonstrated multiple times (Howitz, McKnelly, & Link, 2025). Additionally, when examining the use of rubrics with specifications and defined passing thresholds, the authors found that a binary, two-level system was predominant with more complex systems existing in the literature but being less common (Howitz, McKnelly, & Link, 2025). Finally, their literature review examined the ability to revise and resubmit work as facilitated by token systems, noting that students were not reported to run out of tokens (Howitz, McKnelly, & Link, 2025). While this study is invaluable as a means of gaining knowledge about implementations of specifications grading, the scope of the study is limited to accounts of specifications grading that are published in the literature. This can mean that certain aspects of implementation are not reported on, and thus not able to be analyzed. Additionally, Howitz et al.'s recent study is an exhaustive review of specifications grading across disciplines, meaning that implementations specifically within chemistry have not been compared. We aim to further this work through an in-depth analysis of chemistry instructors' implementations of specifications grading through soliciting chemistry instructors to provide their course artifacts (e.g., syllabi, rubrics, token system outlines). This analysis will enable a detailed examination of implementations previously unreported in the literature. Additionally, this work represents what is to the best of our knowledge the first expansive investigation into the implementation of specifications grading targeted towards chemistry courses. Through this work, we aim to clarify the implementation of specifications grading in chemistry higher education. Such insight will facilitate future work investigating best practices for the implementation of this grading strategy. We thus approached this project while guided by the following research question:

1. How is specifications grading implemented in chemistry courses?

Methods

The course artifacts analyzed in this study are part of a larger interview-based study on instructors' experiences with specifications grading. This study was conducted under Protocol #5936 as reviewed and approved by the University of Virginia Institutional Review Board for the Social and Behavioral Sciences.

Data Collection

Course artifacts were collected from instructors in the United States who use specifications grading in their chemistry courses. Participants were identified through publications, conference abstracts, social media, and personal communications, recruited via email invitation, and provided informed consent before providing course artifacts. Our study sample consists of 50 syllabi from 29 instructors at 24 institutions. These institutions include 9 bachelor's degree-granting institutions, 2 master's degree-granting institutions, and 13 doctoral degree-granting institutions. Many of these instructors implement specifications grading in more than one course. Some instructors shared syllabi for multiple courses and/or for previous course iterations, which contained significant changes compared to their current use of specifications grading. Most instructors included additional course artifacts that they provide to students to help describe their specifications grading scheme. The collection of course artifacts analyzed in this study include syllabi, final grade calculations, grade trackers, token system descriptions, and grading rubrics; the collection of these artifacts will collectively be referred to as "course artifacts" hereafter. In total, our sample includes 50 courses. 36 sets of course artifacts were from lower-division courses (e.g., general and organic chemistry), 12 from upper-division courses (e.g., analytical and inorganic chemistry), and 2 from graduate-level courses. The samples also consists of 20 lecture-only

courses, 11 lab-only courses, and 19 combined lecture-lab courses (i.e., a single course catalog number that has lecture and laboratory components).

Data Analysis

Data were analyzed by two chemistry education researchers knowledgeable about specifications grading. Authors H.M. and B.J.Y. read all course artifacts. Each author independently coded the syllabi based on (1) the presence or absence of rationale for using specifications grading, (2) explicit connections between learning objectives and assignments/assessments, (3) use of a token system, (4) methods for assignment revisions or retakes, (5) number and types of assessment categories, (6) number and descriptors of specifications levels and thresholds, (7) use of term and/or final exams, (8) method of final course grade determination, and (9) additional unique and/or notable characteristics. Following independent analysis of all syllabi, consensus was reached for all codes between the two raters. Then, author H.M. created further categorizations. The additional categorizations included (1) the total number of specification level used with subdivisions for levels of credit compared to levels of marks, (2) whether assessments aligned with learning objectives at the question-level, assessment-level, or bundle-level, and (3) whether and how a final exam was used to adjust a final grade. Authors H.M. and B.J.Y. reached consensus on these further categorizations.

Results and Discussion

The results herein provide a glimpse into the vast array of implementation methods of specifications grading in chemistry courses. To answer the research question, *How is specifications grading implemented in chemistry courses?,* we analyzed features that are a hallmark of specifications grading (threshold for specifications levels, number of specifications levels, nomenclature of specifications levels, and revision opportunities), features that may be a

departmental requirement (final exams), and features that must be considered when designing any chemistry course (final letter grade determination and alignment of assignments to learning objectives). The threshold for meeting specifications is first discussed as this metric is fundamental to course design in a specifications-graded course. Next, the levels of credit that students can earn on assignments are covered, as our research and the literature indicate notable deviations from Nilson's proposed pass/fail grading. We further characterize the nomenclature used by instructors to indicate the level of credit earned by students as we note drastic differences even within specifications grading schemes using the same number of levels. The different approaches to revisions in chemistry courses are covered in detail and include both token and non-token systems. We then examine how final exams are used in specifications-graded courses before characterizing how these chemistry courses determine a final letter grade. Finally, we report on the evidenced direct alignment between specifications grading schemes and a course's learning objectives. Emergent from these primary results are minor trends within chemistry course types (lecture, lab, and combined lecture-lab) which are also explored herein. Together, these results are indicative of how specifications grading is implemented in chemistry courses and provide insight into this complex landscape.

Threshold for meeting specifications

The heart of specifications grading is determining what the specifications are. Indeed, the bar for meeting specifications is primarily responsible for upholding high academic standards. Through only providing credit on assignments when students achieve at or above a traditional B-level of work, students will not succeed in courses without producing high-quality work (Nilson, 2015). Furthermore, this high bar is expected to motivate students to excel in their courses, as they will recognize the need to perform well on each assignment (Nilson, 2015).

The methods of establishing a B-level threshold vary and include students earning a points-based score of 80% on an assignment, students meeting the specifications set for 80% of the items on a provided rubric, and students meeting the individual specifications set for each assignment (Table 22). Notably, there is a fourth method which was postulated by Nilson wherein students would receive credit on an assignment if students performed on a B level for each category in a provided rubric (Nilson, 2015); this designation was not apparent in our sample. Courses which used the 80% score convention are able to easily incorporate traditional assignment into their specifications grading scheme, such as quizzes. This is also a method which we believe will be easily understood by students as it is directly related to traditionally-graded courses. The convention of meeting expectations for 80% of the items on a rubric is rarer in chemistry courses and requires that each item on a rubric carry the same weight or importance; additionally, each rubric item must be evaluated on a pass/fail basis. The other common method of establishing a B-level threshold is through the use of individual specifications that are specific to an assignment. This method enables instructors to provide detailed specifications which can vary based on the assignment type. Thus, the use of individual specifications gives instructors more flexibility in determining what is necessary to receive credit on an assignment and enables adaptations for more complex assignments. Indeed, such individual specifications would facilitate what Howitz et al. noted in their review of the specifications grading literature, where "instructors may also set some specifications as 'required' so that the assignment does not earn credit if those 'required' specifications are not met, regardless of how many others are met" (2025). While the method of establishing a B-level threshold varied from course to course, and even from one assignment type to another, there was nearly always evidence of how this high academic standard would be upheld in specifications-graded chemistry courses.

Table 22. Methods for determining the threshold of specifications on individual assignments

B-level Threshold	Definition	n Courses	Notes
80% percentage score	The assignment is scored based on points; a 80% score or greater is required to receive credit for the assignment	15	Used only on quizzes; course does not have term exams
		6	Used only on quizzes; course has both term exams (specifications or points-based) and quizzes
		1	Used on both homework and quizzes
		1	Used on quizzes and exams (pass = 70%; High pass = 90%)
		1	Used on lab reports with a threshold of 75%
80% of items on rubric	The assignment is graded on a rubric. 80% of the items on a rubric must be met to receive credit on the assignment	1	A detailed rubric should be provided to students before they complete the assignment
Each rubric item completed to a B-level	The assignment is graded on a rubric. Each rubric item must be completed to a B-level	Postulated by Nilson (2015)	A detailed rubric should be provided to students before they complete the assignment. If one rubric item is not completed to a B-level or greater, the assignment does not receive credit
Individual Specs.	The assignments have individual designations for what is required to meet specifications/pass.	27	The threshold is different for each or many assignment(s), particularly across assignment types

Two lecture courses use a combination of an 80% threshold and individual specifications; one lab course does not grade using specifications; two lecture-lab courses use a combination of an 80% threshold and individual specifications; one lecture-lab course does not specify their B-level threshold.

Levels of specifications in grading scheme

Assignments in specifications grading are assigned marks which describe whether the students' work has met the instructor's expectations. We refer to these hierarchical marks as 'levels.' The simplest possible scheme has 2-levels where assignments meet the instructor's expectations (level 1) or they do not (level 2). The 2-level system is recommended by Nilson as she directly relates it to four of the fifteen expected outcomes of specifications grading (upholding high academic

standards, saving faculty time, agreement between different graders, and being simple) (Nilson, 2015). Furthermore, a clear, binary grading system aids in clearly defining the course standards. The prevalence of 2-level systems is seen in a review of specifications grading across higher education (Howitz, McKnelly, & Link, 2025).

The majority of our sample uses such a system (70%). However, some courses in our sample have more complex systems, which vary in their intent and resulting effect on student grades and require a distinction between total number of levels and total number of credit levels, another trend which has been noted in the literature (Howitz, McKnelly, & Link, 2025). The 3-level system with 2 levels of credit as well as the 4-level system with 2 levels of credit are functionally the same as a standard 2-level system. However, the additional levels in the specifications grading scheme enable distinction, though not differing credit, for either exemplary assignments or assignments which were not attempted. For example, a course may have the specifications levels of 'incomplete,' 'needs revision,' and 'meets specifications.' Within this scheme an 'incomplete' represents assignments that are not turned in or not completed by the students and therefore do not receive credit, whereas a 'needs revision' designation is an assignment which was completed below the set standards and thus did not receive credit. Even when isolating our analysis to the 2-level systems, the nomenclature used to describe the two possible marks varies greatly.

Some courses use true multi-level systems including three-, four-level systems. One combined lab-lecture course uses a 3-level system on all assignments; this enables partial credit to be granted to students. Similarly, two chemistry courses (one lecture and one lab) use a 4-level specifications system to allow for varying amounts of partial credit to be awarded. The one lecture course, which has a 5-level system with 4 levels of credit, functions in kind but also has an additional highest level to acknowledge exemplary work without providing more credit to the

students. Interestingly, 16% of the specification graded chemistry courses incorporate a mixture of specification grading levels which varies based on assignment type. While this is most commonly seen in lecture courses (n=5), it is also present in lab (n=1) and combined lab-lecture (n=2) courses. Notably, one laboratory course, while using a specification grading system, does not grade individual assignments based on specifications. Rather, it uses traditional, points-based systems to grade assignments. These graded assignments are then bundled into the students' final letter grades based on the instructor's pre-determined cut-offs. Such sdviations from the binary scale enable instructors to provide partial credit on some, or all, assignments. Our sample is not unique in this feature, as it has been reported how specifications grading schemes can provide partial credit on some but not on all assignments (Bunnell et al., 2023; Kelz et al., 2023). As previously reasoned in the literature, it is possible that partial credit is being introduced into specifications grading due to some assignments remaining relatively high-stakes in the grading scheme (McKnelly, Morris, & Mang, 2021) or to enable differentiation between students with a limited number of assignments (Howitz et al., 2023). Thus, specifications grading can be implemented in courses even if there are concerns about students solely being graded on a pass/fail basis for all assignments.

Table 23. Number of levels in specifications grading schemes

Specification Grading Levels		Courses (n=50)	Notes
2-levels		**70%**	Frequently paired with revision opportunities
3-levels	*3-levels of credit*	**2%**	Enables partial credit
	3-levels with 2 levels of credit	**2%**	Enables distinction, though not differing credit, for exemplary assignments or assignments not attempted
4-levels	*4-levels of credit*	**4%**	Enables partial credit
	4-levels with 2 levels of credit	**2%**	Enables distinction, though not differing credit, for exemplary assignments and/or assignments not attempted
5-levels	*5-levels with 4 levels of credit*	**2%**	Top level is acknowledgement of exemplary work but provides same amount of credit as 2nd highest level. Additionally, not assessable receives minimal credit compared to no credit for no report
Mix of levels	*2-level and 3-level*	**8%**	Often used to give partial credit, which is not normally provided, on high-stakes assignments
	2-level and 4-level	**6%**	
	3-level and 4-level	**2%**	
No specifications grading		**2%**	Specifications grading is not used on individual assignments

Marks ascribed to levels of specifications met

In addition to characterizing the levels in specifications-graded courses, it is important to examine the marks ascribed to said levels of achievement. The nomenclature used to describe student's work is a crucial component of grading schemes as it directly influences how students understand their performance. For example, a label of "pass/fail" provides students with immediate feedback as to whether they have met the specifications set for an assignment. However, when choosing the marks used, instructors must balance this clarity with their intentions. If a grading scheme enables revisions, then a mark of "fail" does not communicate to the students that they can still earn credit. In such instances, a mark of "needs revision" may be appropriate. The decision regarding nomenclature is one that instructors must decide based on their course aims and contexts. Herein we report on the nomenclature ascribed to marks in specifications-graded chemistry

courses, which highlights the inconsistency of specifications grading within the chemistry discipline.

The marks ascribed to levels of specification vary widely. For example, there are 10 passing phrases and 12 failing phrases used just in the systems with 2-levels of specifications (n=35). Interestingly, there are even two courses, each with a different instructor and course type, which use two sets of passing/failing phrases in their syllabi. Despite the wide range, there are more common terminologies. The most common names for a passing mark are *pass, satisfactory, mastery*, and *competency*; and the failing marks are more commonly called *unsatisfactory* and *needs revision*. Among the 35 courses using a 2-level system, there are three distinct categories of marks used to indicate that an assignment has not met the instructor's requirements: unspecified, specified, and specified with implied improvement. 17 courses do not specify what the mark is for a failing assignment in any of their provided course materials. This can potentially be a hinderance to the representative marks of alternative grading; however, there may be a representative mark ascribed in practice, even if this is not apparent in the course artifacts. In the specified group, 8 different marks are used to indicate that an assignment failed (e.g., *unsatisfactory, incomplete, fail*, etc.). However, these are distinct from the marks that incorporate an implied ability of the students to improve. In the specified with implied improvement group, phrases such as "*not yet"* are used at the beginning of the mark (i.e., not yet mastered) and indicate that the students can still learn the material needed in the course. Further examples of this implication is seen with the marks "*Try again*" and "*Needs Revision*." The specification grading systems with more than two levels are also widely varied in their terminology (Table 24). Such variation means that instructors must be clear in communicating their intended message when introducing specifications grading, as

students will not have a pre-developed understanding of the marks system, even if they have had a different specifications-graded course previously.

Table 24. Nomenclature for multi-level systems in specifications grading schemes

	Scheme	Courses	Scheme	Courses	Scheme	Courses
3-level	Good; Acceptable; Unacceptable	n=1	Mastery; Emerging; Not assessable	n=1	1; 0.5; 0	n=1
	High pass Low pass Needs revision	n=3	High pass Pass Needs revision	n=1		
	Scheme	**Courses**	**Scheme**	**Courses**	**Scheme**	**Courses**
4-level	Exemplary; Proficient; Satisfactory; Not specified	n=1	Exceeds standards; Meets standards; In development; Incomplete	n=1	Excellent Meets expectations Revision needed Not assessable	n=1
	Very good; Satisfactory; Unsatisfactory; Missing	n=2	2; 1.5; 1; 0	n=1	Excellent; Good; Revision required; Incomplete	n=1
	Scheme	**Courses**				
5-level	Excellent/exemplary; Meet expectations; Revision needed; Not assessable; No report/not specified	n=1				

Reattempts and revisions

The ability to revise work is a core component of specifications grading. In revising their work, students are expected to learn from their mistakes and make use of their instructor's helpful feedback (Nilson, 2015). Nilson further connects the ability for students to revise work with reducing student stress, discouraging cheating, minimizing conflict between students and instructors, and providing feedback to students that they will use (Nilson, 2015).

The vast majority of specifications graded chemistry courses incorporate some form of reattempt on assignment(s) (90%, Table 25). A large percentage of courses solely use tokens (32%) to grant reattempts on assignments. While this is lower than the 51% of higher education courses analyzed in the literature this does provide further evidence that the revision through tokens

component of specifications grading is reflective of the general implementations in higher education (Howitz, McKnelly, & Link, 2025). In our sample, the token systems are somewhat evenly split between courses which give tokens to the students (8% of total sample), which have the students earn tokens (14% of total sample), and which utilize a combination of earned and freely given tokens (10% of total sample). Tokens which are freely given to students may be provided at the start of course, in set intervals throughout a term, or after preliminary assignments (e.g., group-formation questionnaires, syllabi quizzes). Tokens which are earned may be done so through homework completion/grades, completion of surveys, a period of time without lab safety violations in the section, a period of time without incidents of academic dishonesty in the section, etc. Interestingly, 56% of the syllabi from chemistry courses ensure reattempts without grade penalties without using a token system. Of these, the vast majority utilize an automatic reattempt system wherein students are provided with a set number of reattempts on assignments which are often pre-scheduled. For example, a course can have biweekly quizzes on Tuesdays, where students are able to reattempt to previous quiz each 'off-week.' Far fewer courses use an "unlimited"-reattempt system, and none of chemistry lab courses do so. In such unlimited systems, students are able to reattempt assignments as many times as they wish. This can bring practical concerns, particularly in regards to demands on instructors' assessment design and grading time, which may account for the rarity of such revision systems. Notably, even these "unlimited" systems typically still have some boundaries on student revisions. There may be a restriction of a certain number of reattempts per week, and exceptions to unlimited reattempts must be made for activities assigned at the end of the term. Thus, there are numerous avenues through which instructors provide their students with revision opportunities within specifications grading, further highlighting the potential adaptability of this grading scheme.

Table 25. Approaches to assignment revisions and reattempts

Revision Method		Definition	Courses (n=50)	Notes
Revision/ Reattempt in exchange for token (revision may not be only use of token(s))	**Given**	A set number of tokens are given to students after completing initial assignment(s)	**8%**	A set number of tokens are given to students at the start of term . These may be automatically given or be dependent on initial assignment(s) (syllabi quizzes, group formation surveys, etc.)
	Earned	Students can earn up to a set number of tokens throughout the course	**8%**	Tokens are earned through homework completion/grades, completion of surveys, a period of time without lab safety violations in the section, a period of time without incidents of academic dishonesty in the section, etc.
	Given and Earned	Students are provided a set number of tokens at the start of term and are able to earn additional token(s) throughout the course	**10%**	The tokens provided at the start of term may be freely given or associated with initial assignment(s)
Automatic reattempts are built into the course structure	**Limited**	Students are automatically allowed a set number of reattempts on assignment(s)	**32%**	Exceptions to reattempts or the normal amount are reattempts may be made for assignments assigned at the end of the term
	No limit	Students are allowed unlimited attempts on assignment(s)	**6%**	There may be a restriction of a certain number of reattempts per week. Exceptions may be made for assignments assigned at the end of the term
Automatic and Exchange of Token	**Earned Tokens**	Some amount of revision is built into the course structure. Exchange tokens for reattempts beyond what is built into course structure	**10%**	Can be used to increase number of allotted attempts per week or to increase number of allotted attempts per assignment
	Given and Earned tokens		**16%**	
No Revisions		No revision opportunities on any assignments	**10%**	Not recommended under specifications grading (Nilson, 2015)

In addition to the limitations built into token systems and reattempt schedules, some chemistry courses impose additional considerations when navigating student resubmissions. Two courses, one lab and one combined lecture-lab, limit the ability to reattempt an assignment based on the quality of the original attempt. Specifically, additional attempts on assignments are not granted to students who did not submit a first attempt by the original due date. Three courses, one of each course type, implement additional token penalties depending on the original score, effort, or attendance. The distinction is made between either originally failing scores and originally absent (i.e. not turned-in) assignments or, in multi-level specifications schemes, between an assignment with an originally failing score as compared to an assignment with an originally partial-credit score. In implementation, these systems will require additional tokens to submit a reattempt that

was either not turned in or did not receive even partial credit, as compared to a good first attempt that was just short of the needed specifications. Additionally, five courses, one lecture and four combination lecture-labs, require students to complete additional assignments to receive the ability to resubmit something that did not meet specifications. These additional assignments are geared towards increasing understanding of material (e.g. reflections, re-working missed problems, attendance at tutoring hours, meeting with instructor, etc.). These additional assignments are completed in systems a variety of systems. They can be used when courses do not have tokens, be required in conjunction with exchanging a token, or be used as an additional avenue for revisions in place of a token penalty. Such additional restrictions can ensure both that the number of revisions are manageable for instructors and that the revision opportunities promote students' learning.

An interesting divide is seen among the courses which do not offer revision opportunities (n=5; taught by 4 instructors). 60% of the courses without revision opportunities move away from the 2-level specifications grading system and incorporate multiple levels of credit on at least some assignments. This may be to compensate for high-stakes assignments which students cannot revise. However, the remaining 40% of courses which do not have revision opportunities stay within the strict, binary grading associated with the specifications grading system. Understanding the reasoning behind the instructors' grading and revision scheme is a promising area of future study, especially as schemes absent of revision opportunities are present across disciplines despite being rare (Harrington et al., 2024).

Explicit alignment with learning objectives

Nilson specifies that assignments in a specifications-graded course must directly align with the course's learning objectives (LOs). Thus, grades earned in a specifications grading scheme reflect student learning outcomes, which aids in motivating students to learn the material, as opposed to

simply doing enough to pass the course (Nilson, 2015). The alignment with LOs is also the primary method of final grade analysis in Howitz et al.'s review of specifications grading in higher education (Howitz, McKnelly, & Link, 2025). This prior study identified four main LO based methods: the core and additional LOs model, where LOs are categorized as "core" (must be met to pass the course) and "additional" (must be met for higher grades); the all equal LOs model, where all LOs are weighted equally and scores are based on number passed; the modules configuration, where related LOs are grouped into modules, and students must meet a certain number of LOs in each module to both pass and earn higher grades; and the all equal LOs with repetition and/or complexity (ELORC) model, where LOs are revisited throughout a course and the number passed corresponds to a student's grade (Howitz, McKnelly, & Link, 2025). However, such categorizations can be difficult to decipher from course artifacts, and indeed, may not be commented on in published literature accounts (Howitz, McKnelly, & Link, 2025). Thus, below we have analyzed how courses in our sample explicitly design LOs into their assessments in addition to their bundling systems.

60% of our sampled courses have artifacts which show direct alignments with LOs; however, it is possible such relationships exist in the remaining 40% yet were not captured in the course artifacts. 8 courses align individual questions on assessments with learning objectives, resulting in multiple LOs being covered in a single assignment. For example, a four-question quiz may cover four LOs with each question only examining a single LO. Additionally, an entire assignment may be used to assess a single learning objective (n=19). A common example of these schemes occurs when students are assessed on each LO through a weekly, targeted quiz. These first two methods require explicit design efforts on the part of an instructor to ensure that only one LO is being assessed at a time. The final method we observe in our sample is the alignment of LO to the bundles

which determine a student's final letter grade. Thus, with this method, the higher-grade bundles indicate that a student has achieved more learning objectives as compared to the lower-grade bundles. This last type of alignment is only seen in 3 courses, all of which are lab courses. In its totality, the evidenced alignment of specifications grading schemes with LOs is promising and indicate that the adoption of specifications grading may foster alignment between course aims and students' measured outcomes.

Table 26. Evidenced and explicit alignment between learning objectives and aspects of specifications grading

Alignment to learning objectives	Definition	n Courses	Notes
Alignment with questions	Individual questions on or parts of assignments align with learning objectives	6	Multiple learning objectives are covered within a single assessment
Alignment with assessments	An assessment, as a whole, aligns with a learning objective	21	An assessment is specific to a particular learning objective
Alignment with bundles	If a final grade is determined by bundling, the higher-level bundles encapsulate more learning objectives being met as compared to lower level bundles	3	Only observed in syllabi for laboratory courses

Note that more alignments may be present than described herein as we are limited to what is encapsulated in the provided course artifacts.

Final exams

As with all higher education courses, instructors teaching chemistry using specifications grading must decide whether to incorporate a final exam into the course structure. Approximately one third of our courses do not use a final exam. Interestingly, the remaining two thirds are relatively evenly split between having a specifications-graded final exam or a traditional points-based final exam (Table 27). Final exams which are specification-graded have an associated passing threshold. Whether the specification is met then either is directly included into the bundling system or mapped onto points for a grade determination (see the below section). The

relatively even split between courses having no final exam, courses using a traditionally-graded final exam, and courses grading their final exam based on a pre-determined specification may be representative of contextual requirements or limitations. For example, some courses have a departmental or institutional requirement to incorporate a final exam into their grading structure. Additionally, some courses may either choose or are required to use a version of the ACS exam as their final. Indeed, it has been previously reported that specifications grading schemes incorporate the ACS exams as a method of comparison across cohorts (Bunnell et al., 2023). Thus, the variation seen in the approach to final exams may be representative of the adaptability of specifications grading to different contexts.

Table 27. Approaches to final exams in specifications graded courses

Approach to final exam		Definition	Courses (n=49*)
A category of assessment	*Final exam has a specification*	Final exam has a passing threshold and is either mapped onto points or included in the bundling of final grades	**34%**
	Traditional final exam	Final exam is points-based and used as a percentage of the final grade	**32%**
No final exam	The course does not have a final exam		**32%**

There was not enough information in the course artifacts to make a classification for 1 course in our sample.
Four courses utilize a dual grading system where grades are calculated using both a traditional and specifications grading system; student earn the higher grade if there is a difference. The table above references the final exam utility in the specifications grading scheme.

The final exam periods, and indeed the final exams themselves, also have varied utilities. Courses which do not have a final exam may still make use of their institution's final exam period as a final reattempt opportunity for students to meet specifications on assignments or as a final deadline for certain assignments (such as projects or presentations). Of the former, 5 courses use the exam period as a final time where students can reattempt prior assignments, and 4 have a separate final exam which can replace prior failing grades. For the latter purpose, the final exam is broken into sections, and students complete the specific sections for which they have not yet passed the associated learning objectives. When a final exam is its own category of assessment, it can be incorporated directly into the final grade determination or be used as a modifier to the grade

students earn through all other assessments. When modifying grades, final exams may be an optional opportunity for students where they can increase their final letter grade with no concern of negatively impacting their performance. One unique method of doing so occurs in lecture course which uses the final exam period for a dual purpose – the final exam counts as a traditional point-based percentage of a student's grade and can also be used to receive specifications on previously failed learning objectives. Additionally, the final exam can be a required assessment which either adjusts a base grade upwards or downwards; this is most commonly seen in courses which use plus-minus letter grade designations.

Final grade determination

Most of our sample assign final grades through some form of a bundling system. These systems are intended to motivate students to learn and to make them feel responsible for their grades (Nilson, 2015). Furthermore, the clear path to a final grade supports the simplicity of the grading system (Nilson, 2015) and contributes to the clearly defined course standards. Despite this commonality, what is bundled differs from course to course. Points achieved on different assignments can be bundled to determine final grades, specifications passed on different assignments can bundled together for final grade determination, or both point-based performance and specifications passed can be incorporated into the bundling process. The first method is only present in one lab course. The second method, the true bundling of specifications, occurs in 22% of our sample and follows the bundling method described by Nilson (2015). Last method of bundling is when the performance on specifications-graded assignments and reaching a certain point-based score on other assignments combine to determine the bundles for final grades. The bundling of points and specifications accounts for the largest percentage of our sample (46%). The common non-bundling form of final grade determination is the mapping of specifications onto

points (24% of our sample). Under this methodology, meeting specifications on assignments, or on a certain number of assignments, is associated with a point value. This point value then comprises a certain proportion of students' final weighted average. This method, as with the incorporation of both points and specifications into the bundling system, enables instructors to have different grading methods for different assignments. A combination of grading methods may be necessary due to factors such as LMS limitations as previously reported in the literature (McKnelly et al., 2023). Lastly, one instructor, who teaches three combined lecture-lab courses, uses passing specifications as a measure to pass the course. This means that if students do not meet specifications on certain assignments, they automatically fail the course; however, their course grade is determined by additional point-based assignments.

Table 28. General methods of final grade determination

Final grade determination		Definition	Syllabi All (n=50)
Specs as a measure to pass		Students must meet specifications on certain assignments to pass the course	**6%**
Mix of points-based and specifications-based	*Specifications map onto points*	Meeting specifications on assignments is assigned a point value. The point value then combines with values from other point-based assessments for the final grade calculation	**24%**
	Bundling of specifications and points	Specifications-graded assignments and reaching a certain point-based score on other assignments combine to determine the bundles for final grades	**46%**
Bundling of specifications		Specifications-graded assignments are combined into bundles to determine the final grade	**22%**
Bundling of points		Point-based performances on various assignment types are bundled to determine the final grade	**2%**

While the use of a plus-minus system is unanimous across higher education, it is an important design feature at institutions which utilize the system. Specifications-graded chemistry courses which use a plus-minus system do so through 1) having individual bundles which

correspond to D- through A+ grades, 2) through assigning plus-minus grades according to students' final point-based grade, or 3) through applying plus-minus grades based on additional considerations. The first method necessitates a final grade determined through bundling, whereas the second necessitates a non-bundling method. However, the third method can apply to any of the discussed methods of final grade determination. The additional considerations which dictate plus-minus grades include factors such as students' performance on a final exam, the results of peer reviews, and additional assignments completed beyond bundling requirements.

Table 29. Methods of final grade determination with a plus/minus system

Use of plus/minus system		Definition	Courses (n=49*)
Yes	*Bundling*	Plus/minus grades may be their own bundle or assigned based on proximity to base-grade bundles (i.e. one assignment over or short of a bundle's requirement)	**30%**
	Additional considerations	Often a final grade will be used as the base grade, and adjustments are made for a plus/minus based on certain assignments	**26%**
	Points-based	Plus/minus grades are built into the points-based system	**22%**
No		Plus/minus grades are not assigned or reported	**20%**

*There was not enough information in the course artifacts to make a classification for 1 course in our sample.

Trends within course types

The above characterizes the implementation of specifications grading in higher education chemistry courses. However, we further analyzed our sample to explore differences by course type (lecture, lab, and combined lecture-lab courses). Commonalities are readily apparent when examining the levels of specifications used, with the two-level system present in most courses throughout our sample. Additionally, no distinct trends were apparent in the nomenclature used to describe marks. Our examination of trends within course formats reveals the most similarities between lecture and lecture-lab courses with both having revisions being built into the course structure in addition to using a 2-level system. Additionally, lecture courses and lab courses both have a large proportion determine final letter grades through a bundling system. This contrasts with lecture-lab courses which either use a bundling system or map specifications onto points. The

latter option may be due to the combination of assignment types found in lecture-lab courses. Interestingly, no course format had a majority of instructors following one approach to final exams, indicating that other factors, such as departmental context, may dictate this aspect of course design. The two major trends are described below in further detail.

Revisions and reattempts

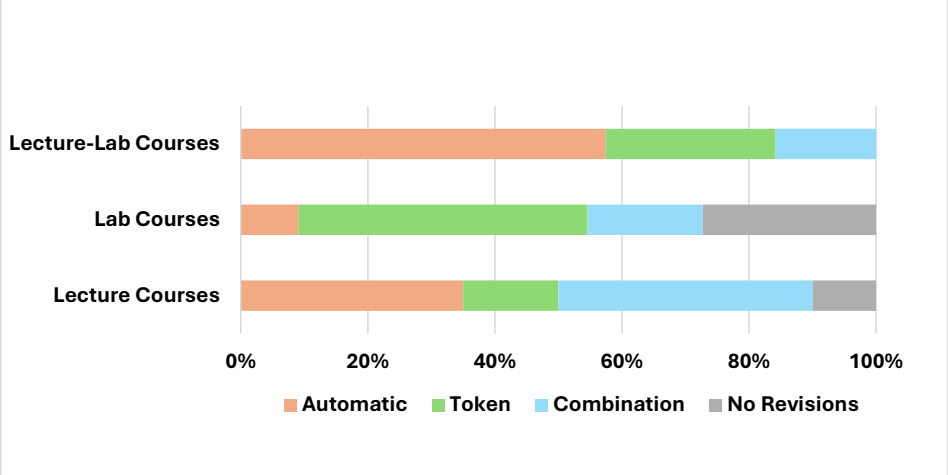

Figure 13. Approaches to revisions and reattempts by course type

Following what is recommended by both Nilson and Talbert and Clark, the lecture and combined lecture-lab chemistry courses implement revisions into their course structure (90% and 100% of courses in the samples, respectively). The chemistry lab courses have a larger percentage (27%) not offering revisions. This may be due to the difference in assignment types. Whereas common lecture-associated assignments, such as quizzes and homework, can be completed by students individually and be graded with relative ease depending on format, lab assignments are either impractical to make-up or require extended effort on the part of the grader to revisit. Indeed, many lab assignments are dependent on being in lab and completing experiments as part of a group, meaning that if lab is missed, the assignments are impossible for students to complete without

extraordinary effort on the part of the instruction team. Furthermore, many chemistry lab courses incorporate writing assignments which take more time and mental effort to grade. This potential consideration of instructors' time and energy is also seen in the lab courses which do allow for revisions, as the revisions are mostly done in exchange for a token. This exchange means that students have a set number of revisions per term and will need to selectively choose which assignments to revise. Such restrictions may be a practical requirement as a past critique of specification grading is the 'false' promise of saving instructors time (Carlisle, 2020; Hunter, Pompano, & Tuchler, 2022; Lovell, 2018; Prescott, 2015; Shields, Denlinger, & Webb, 2019; Spurlock, 2023; Vitale & Concepción, 2021).

Final grade determination

The second major distinction between course types is seen with the method for final grade determination. While lab and lecture courses showcase a wide variety of methods, combined lecture-lab courses are unique in that they always incorporate points. In these hybrid courses, specifications are either mapped onto points or the bundles designed by instructors include both point-based and specifications-graded assignments. Notably, the two syllabi which only use specifications as a measure to pass are combined lecture-lab courses (taught by the same instructor). It is possible that the multitude of assignments addressing both laboratory learning outcomes and lecture learning outcomes may be difficult to combine in a bundling system based solely on specifications-graded assignments.

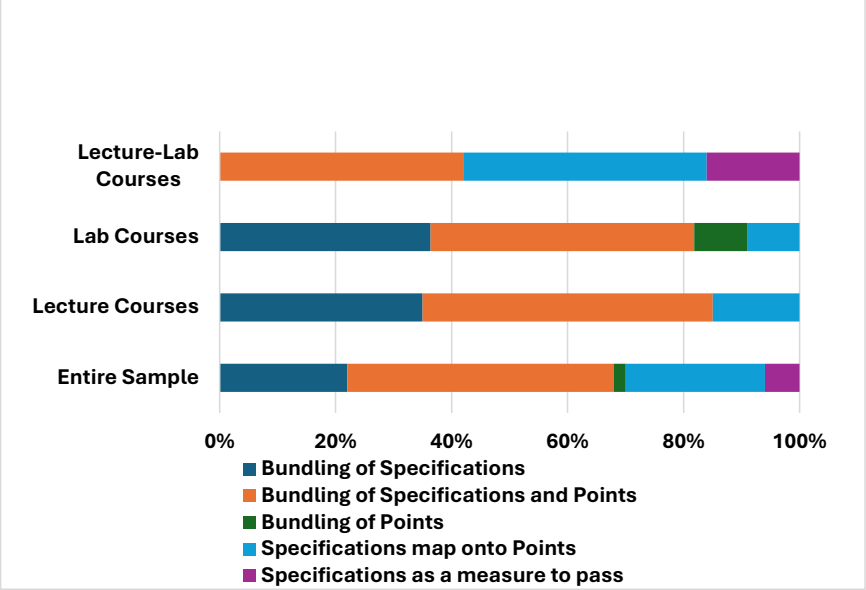

Figure 14. Methods of final grade determination by course type

Common implementations by course type

Further distinct differences were not readily apparent when segregating our sample by course type. However, it is valuable to summarize the implementation of specifications grading in chemistry courses by course type further capture how specifications grading is used.

Lecture. As with all of the course formats, the majority of chemistry lecture courses use a 2-level specifications system when assessing their students. Only 15% (n=4) have opportunities for partial credit on all assignments. However, the awarding of partial credit on select assignments is present among a large proportion (25%; one course has varying levels of partial credit depending on assignment type) of lecture courses. Chemistry lecture courses also exhibit a wide variety of revision systems. Notably, if a lecture course incorporates a token system into the larger revision system, at least some tokens will be provided to the students without being earned. Furthermore,

only 15% of the courses rely on a token system. The other 40% of our lecture sample which uses tokens pairs the exchange of a token with automatic reattempt opportunities. Automatic reattempts are also popular with another 35% of the sample, with 25% having limited automatic reattempts and 10% not placing a limit on the students, apart from institution's term duration and total number of reattempts per week. When approaching final exams, a significant portion (45%) of lecture courses do not use the category of assessment. This is the same proportion as lab courses and differs significantly from the combined lecture-lab courses which only have 11% of our sample omit a final exam. Lecture courses have varied methods of determining plus-minus grades; however, the plurality (40%) do so through incorporating the plus-minus grades into specific bundles. The bundles themselves are typically either created through combinations of passed specifications (35%) or combinations of passed specifications and performance on points-based assignments (50%).

Lab. The majority of specification-graded chemistry lab courses use a 2-level specifications grading system. Three of the remaining courses utilize a system which enables partial credit to be granted on all or a portion of assignments. Additionally, chemistry lab courses tend to have a restricted system for revisions, with 8 courses making use of a token system in exchange for revisions. Notably, only two of these courses pair the token system with a limited number of automatic revisions, and in all courses there are multiple uses of the students' tokens beyond revision opportunities (e.g., deadline extensions and excusing absences). Furthermore, 27% of the lab courses have no opportunities for revision. Interestingly, there is an even split between chemistry lab courses which use a specification-graded final exam and those which do not have a final exam as a part of the course. Lab courses also had the greatest variance in method of final grade determination. However, over 80% fell into either the bundling of specifications

(36%) meaning that the grading systems are devoid of any points-based assignments or into the bundling of specifications and points (45%) meaning that there are a mixture of both specifications- and points-based assignments. Chemistry lab courses were also the only course type wherein additional considerations were the primary method of determining plus/minus grades.

Combined lecture-lab. Combined lecture-lab courses are more homogenous than the other two formats in regards to the levels of specification. 84% use a solely 2-level specifications grading system when assessing all assignments. Combined lecture-lab courses are also the only course type in which 100% of the sample provides some measure of revisions or reattempt opportunities. Indeed, 53% have automatic reattempts built into their course structure and do not rely on an additional system (i.e. tokens) for their students to improve their work. A further 16% combine token and automatic revision systems; 27% rely solely on token systems to enable their students to revise assignments. The combined lecture-lab chemistry courses are the only course type which does not commonly omit final exams (11%). Instead, 53% of combined lecture-lab courses use a traditional final exam, and 32% use a final exam assessed against a specification. Further separating this course type is the 16% of lecture-lab courses which use specifications as a measure to pass, though all three of these courses are taught by the same instructor. The majority of lecture-lab courses do not deviate from the norm set by the other two formats with and even split of lecture-lab courses between using the bundling of specifications and points and the mapping of specifications onto points when determining a final letter grade.

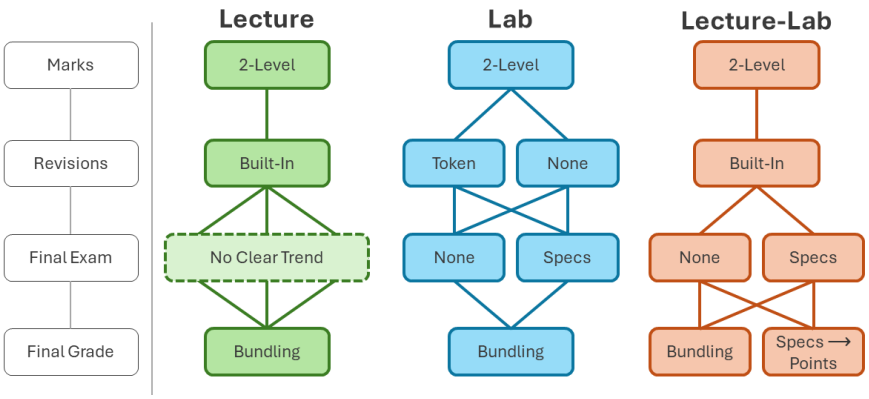

Figure 15. Implementations of specifications grading by course type

Limitations

While efforts were made to solicit syllabi and course artifacts from all higher education chemistry instructors utilizing specifications grading in the US, the researchers were likely not aware of all instructors, and not all instructors contacted were able to be reached or their materials used. Furthermore, this study is limited by the instructors providing only course materials and artifacts, as documents such as syllabi may not indicate alterations made during a term, such as retroactively allowing for revisions

Conclusion and Implications

This exploration of specifications grading within higher education chemistry courses reveals a complex picture with variations in implementation even in the aspects which are fundamental to specifications grading. The method for establishing a B-level threshold varies, with some courses implementing unique systems for each assignment and others remaining grounded in a percentage score as reference. While most courses use a basic two-level grading system, some courses incorporate multi-level systems that allow for partial credit or distinguishing between exemplary

work and incomplete assignments. The most common terms used to describe passing marks include "pass," "satisfactory," and "mastery," while failing marks are often referred to as "unsatisfactory" or "needs revision." A core feature of specifications grading is the ability for students to revise their work, which is present in 90% of courses in the sample. Approximately one-third of our sample offers revision opportunities through a token system, though others rely on automatic reattempts with either unlimited revisions with certain restrictions. Despite these diverse approaches, a tenth of our sample does not offer revision opportunities, with some incorporating multiple levels of credit to replace the need for revisions.

Further differences are seen in characteristics of specifications grading which must be considered when designing any STEM course. Indeed, in specifications-graded chemistry courses, alignment with learning objectives (LOs) is a crucial aspect of grading, as it ensures grades reflect student mastery of course material. Approximately 60% of the courses in the sample explicitly align assignments with LOs, either by having individual questions address separate objectives or by using entire assignments to assess specific LOs. Additionally, a few courses use LO alignment in their grading bundles, where higher-grade bundles represent achievement of more LOs. When considering final exams, about one-third of the courses omit them, while the rest are split between specifications-graded and traditionally graded finals. Regarding final grade determination, most courses use a bundling system. However, even more common is the combination of specifications and points, such as the bundling of both point-based and specifications-graded assignments or the mapping of specifications onto point values. Indeed, this study showcases the many different implementations of specifications grading in chemistry courses which shows both the adaptability and complexity of this grading scheme.

For researchers, these findings suggest that outcomes attributed to specifications grading are closely tied to the specific implementation details rather than reflecting a one-size-fits-all model. Consequently, future studies should focus on linking observed outcomes directly to the particular characteristics of each implementation rather than attributing effects to specifications grading as a general practice. Identifying best practices for specifications grading will require nuanced understanding, acknowledging that optimal approaches may vary across different course formats. Additionally, while this study provides a valuable snapshot of current practices, it does not explore the underlying reasons for the choices made in implementation. To gain a deeper understanding, future research should include qualitative methods, such as interviews, to investigate the rationale behind different implementation strategies and their impact on educational outcomes. This comprehensive approach will help in developing more tailored and effective models of specifications grading.

These findings also underscore the need for instructors to first consider specific course goals and learning outcomes to then tailor specifications grading to their classes. Instructors should also be mindful of contextual factors or limitations and of the ability to adopt specific aspects of specifications grading that best fit their objectives when full implementation is improbable. Furthermore, it is crucial to view specifications grading as an evolving process, where adjustments to implementation can occur across multiple iterations of a course. To optimize the design and effectiveness of specifications grading, instructors are encouraged to consult a variety of perspectives and resources, recognizing that each implementation will be unique. In all, instructors adopting specifications grading will need to take an adaptable and collaborative approach when designing their courses.

Chapter 5: Students' perceptions of specifications grading: development and evaluation of the Perceptions of Grading Schemes (PGS) instrument

The following is adapted from an open-access publication with copyright retained by the authors: Yik, B. J., Machost, H., Streifer, A. C., Palmer, M. S., Morkowchuk, L., & Stains, M. (2024). Students' Perceptions of Specifications Grading: Development and Evaluation of the Perceptions of Grading Schemes (PGS) Instrument. *Journal of Chemical Education*, *101*(9), 3723-3738.

Introduction

In the United States, the now-standard traditional grading schemes consists of assigning letter grades on an A–F scale often in combination with a 100-point scale system (Bowen & Cooper, 2022; Brookhart et al., 2016; Schinske & Tanner, 2014). Course grades are generally described as an aggregate measure of student performance on individual assessments (*e.g.,* homework and exams) and behavioral components (*e.g.,* attendance and participation) (James, 2023; Lipnevich et al., 2020). Grades impact students' potential to obtain course credit, receive academic scholarships, earn a degree, access graduate or professional degree programs, and employment prospects (Rosovsky & Hartley, 2002). Consequently, students use grades to make decisions about their college major and careers (Witherspoon, Vincent-Ruz, & Schunn, 2019; Witteveen & Attewell, 2020) and grades have been shown to relate to student retention (Chen, 2013; Cromley et al., 2013; King, 2015; Ost, 2010; Rask, 2010; Witteveen & Attewell, 2020). Given the weight that grades carry in the academic career of students and in the preparation and selection of the workforce, one would assume that grades represent valid and reliable measures of student learning. Unfortunately, extensive research points to the inadequacy of the current grading system.

First, grades have inconsistent meanings (Blum, 2020). Instructors use a wide array of different assessment tools to evaluate students (Gibbons et al., 2022). Therefore, instructors may

not only include different tools (*e.g.,* examinations, homework, attendance) in their course grading scheme, but also at different weights to determine a final course grade. For example, one instructor may use only examinations to determine course grades, another instructor teaching the same course at the same institution may use a combination of examinations, homework, and participation scores. Further, grades can be highly variable depending on the amount of partial credit awarded on individual assignments (Brookhart et al., 2016; Herridge & Talanquer, 2020; Mutambuki & Fynewever, 2012).

Second, grades provide no actionable feedback which is essential for student learning (Blum, 2020). Research shows that a score or letter grade on student work does little to improve learning (Page, 1958; Stewart & White, 1976). A score conveys little information beyond how many points were earned on an assignment. Students need guidance and direction from instructors on how to improve their learning (Guskey, 2019). Unsurprisingly, students indicate that detailed and actionable comments that provide guidance are the most important and useful form of feedback and note that grades are ineffective in supporting improvement (Lipnevich & Smith, 2009). Studies show that students receiving written comments or performance feedback in addition to grades have increased achievement and motivation (Koenka et al., 2021) with one study finding that descriptive feedback unaccompanied by grades yielded the highest improvement in quality of student work (Lipnevich & Smith, 2008).

Third, grades diminish students' intrinsic motivation and enhance their extrinsic motivation (Chamberlin, Yasué, & Chiang, 2023; Kohn, 2011; Schinske & Tanner, 2014). Most notably, grades influence students' decisions about their major and career, and determine students' ability to earn a degree (Rosovsky & Hartley, 2002; Witherspoon, Vincent-Ruz, & Schunn, 2019;

Witteveen & Attewell, 2020). These external rewards can therefore drive students' behaviors in the need for good grades.

Finally, extensive research has demonstrated that traditional grading systems contribute to educational inequities (Feldman, 2019a, 2019b; Link & Guskey, 2019; Matz et al., 2017). Studies have shown that grades are more a reflection of students' access to resources (*e.g.,* mental health, tutoring, financial health) and demographics (*e.g.,* zip code of their high school) than their learning (Feldman, 2019a; Johnson, Molinaro, & Motika, 2018). For example, systemic issues of implicit racial, class, and gender biases can negatively affect the grades assigned to underrepresented students (Feldman, 2019b). Beyond an educator's own biases, traditional grading can still further disadvantage specific groups of students. For example, students with unstable living situations perform worse on assessments meant to be completed outside of the classroom, regardless of what they learned (Feldman, 2019b).

In summary, traditional grading schemes do not reflect nor support student learning. Alternative grading schemes have been proposed to address these shortcomings.

Alternative Grading

Alternative grading schemes are a deliberate shift away from assigning evaluative grades on individual assignments and emphasize the learning process. In their conceptual framework of alternative grading, Talbert and Clark (2023) outline four key features (*i.e.,* pillars) of common elements of alternative grading: clearly defined standards, helpful feedback, marks that indicate progress, and reattempts without penalty.

- *Clearly defined standards.* Standards are clear and measurable actions that learners take to demonstrate their learning; thus, standards can be thought of as learning outcomes that learners must show evidence of learning for through clear and measurable tasks.
- *Helpful feedback.* Feedback is the evaluative information that results from the outcomes of those clear and measurable tasks. Helpful feedback is given often and framed in terms of the clearly defined standards and provides opportunities for growth.
- *Marks indicate progress.* Rather than points, "marks" refer to the outcome of a task and indicate progress to meeting the defined standards. Marks can be a word or short phrase that make clear to students their progress and understanding. For example, in specifications grading, marks that indicate progress can include words and phrases such as "meets specifications," "satisfactory," and "needs revision."
- *Reattempts without penalty*. Learners are allowed the opportunity to reattempt work and resubmit it for feedback without incurring any grade penalty. When reattempts without penalty are in place, learners can use the helpful feedback and marks that indicate progress they received to meet the clearly defined standards.

Together these four pillars of alternative grading are aimed at promoting growth and equity by ensuring that students have access to the same opportunities to learn and succeed by removing structural barriers in the grading system.

Specifications Grading

A detailed exemplar list of STEM courses that have implemented specifications grading published in peer-reviewed journal publications or conference papers between 2017 and 2022 can be found in McKnelly *et al.* (2023). Features of specifications grading as described in Nilson's book include: (1) individual assessments are graded on a pass/fail basis, (2) assessments are

supplemented with clear and detailed specifications of what represents passing, (3) specifications reflect at least B-level work, (4) students are allowed to revise or retake a limited number of assessments that do not meet specifications, (5) bundles of assessment that correspond with higher course grades require students to demonstrate a more advanced mastery of content and/or skills, and (6) these bundles are aligned with the course learning outcomes. In specifications grading, the instructor creates a set of specifications for each assignment that students must meet. These specifications are aligned with the course learning outcomes. Individual assignments are not given letter grades or points, but rather actionable feedback and a statement indicating whether the assignment met the stated specifications. The instructor then bundles the assignments to define expectations for each course-level letter grade. The difference between the bundles is based on the nature of the assignments and/or the number of assignments that meet specifications (Figure 16).

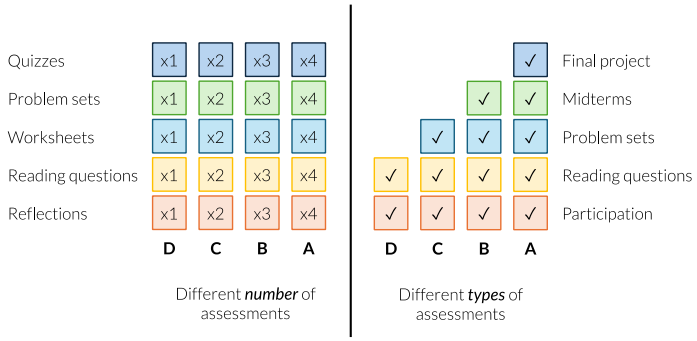

Figure 16. Examples of bundling strategy in specifications grading.
Columns represent bundles and their associated course-level letter grade, and rows represent different types of assignments.

To lower the stakes and promote mastery, students have opportunities to resubmit assignments. This process may be facilitated using a token system. Students are provided with and/or can earn a limited number of tokens. They can use these tokens in exchange for a deadline

extension, revising and resubmitting an assignment, or any other non-official accommodation the instructor allows.

Outcomes of specifications grading

Specifications grading is relatively new, and thus peer-reviewed research on specifications grading is still emerging. Most publications to date are descriptions of implementation, personal experiences, and students' satisfactions with specifications grading (e.g., Bunnell et al., 2023; Houseknecht & Bates, 2020; Martin, 2019; McKnelly, Morris, & Mang, 2021). Recently, there has been a growing number of studies that explore the effectiveness of specifications grading in chemistry with empirical evidence (Ahlberg, 2021; Mary E. Anzovino et al., 2023; Bunnell et al., 2023; Closser, Hawker, & Muchalski, 2024; Donato & Marsh, 2023; Howitz, McKnelly, & Link, 2021; Katzman et al., 2021; McKnelly et al., 2023; Moster & Zingales, 2024; Noell et al., 2023). These studies investigate the effectiveness of specifications grading typically through a comparison of final exam scores and/or final course grade distributions between specifications grading and traditionally taught versions of the course. It is important to note that Nilson does not claim that there should be an increase in course content knowledge nor an increase in final course grades. Instead, Nilson (2015) claims that specifications grading achieves 15 outcomes which can be roughly divided into instructor-centered and student-centered outcomes (Figure 17).

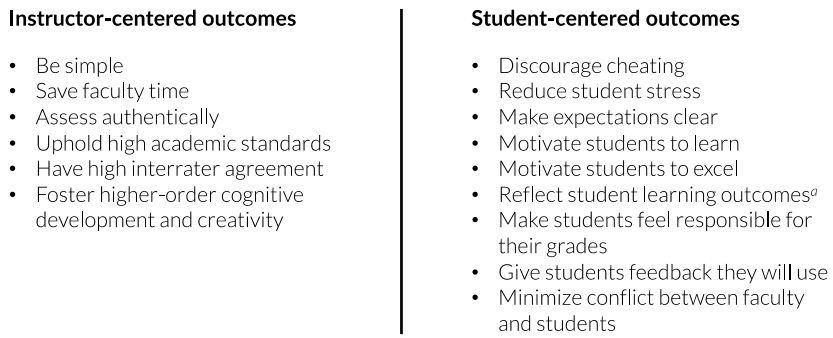

Figure 17. Theorized outcomes of specifications grading (Nilson, 2015)

With the popularity of specifications grading growing, there is a need to understand the extent to which the theorized student-outcomes are realized and thus to develop measures of these outcomes that provide valid and reliable data (Hackerson et al., 2024).

Study goals

The study herein is the first empirical exploration into understanding multiple student-centered theorized outcomes of specifications grading in chemistry. Given the nature of the student outcomes (*e.g.,* motivation, experienced level of stress, and clarity of expectations), the measure of these outcomes focuses on student perceptions. The specific goals of this study were to:

(1) Develop the Perceptions of Grading Schemes (PGS) instrument to measure student perceptions of specifications grading when compared to traditional grading.

(2) Determine the reliability and validity of the measurements made with the PGS instrument.

One specific research question of interest was:

(1) What are students' perceptions of specifications grading in a general chemistry laboratory course?

Methods

This study was conducted under Protocol #5793 which was reviewed and determined to be exempt by the University of Virginia Institutional Review Board for the Social and Behavioral Sciences.

Research Context

This study was conducted in a two-semester sequence of a general chemistry laboratory course at the University of Virginia, a large, very-high research-intensive university in the mid-Atlantic region of the United States. Author LM is the instructor and coordinator for the general chemistry laboratory sequence. Both courses use guided inquiry experiments that students complete in teams of three to four. The first semester general chemistry laboratory (*i.e.*, GC1 Lab) is only offered in the fall semester enrolling 1600–1700 students. GC1 Lab is enrolled by students with a major in the College of Arts and Sciences requiring general chemistry laboratory for their degree, and all students in the College of Engineering. The second semester general chemistry laboratory (*i.e.*, GC2 Lab) is only offered in the spring semester enrolling 800–900 students. The drop in enrollment in GC2 Lab from GC1 Lab is due to a change in the course population. Students enrolled in the School of Engineering are only required to take GC1 Lab. Therefore, very few engineering students enroll in GC2 Lab.

The specifications grading scheme implemented in these courses were inspired by and loosely modeled from the specifications grading scheme used in the organic chemistry laboratory sequence at the University of California, Irvine (Howitz, McKnelly, & Link, 2021). Course assessments are bundled together to assess student competency of student learning outcomes. Altogether, there are eight bundles with each corresponding to a student learning outcome. Each student learning outcome is assessed with a set of specific course assessments. Course assessments are evaluated on a binary scale. The threshold for sufficient student competency of learning

outcomes is set at work representative of a B grade. All assessments were graded by graduate teaching assistants (GTA) or undergraduate teaching assistants (UTA) that had been trained on using the specifications grading scheme with previous student work. Rubrics for each assessment were made available to students in the course learning management system to increase grading and expectations transparency. Students had opportunities to revise and resubmit work that did not meet the competency threshold. More information regarding the implementation can be found in Morkowchuk (2024).

Data collection

Study participants consisted of Spring 2023 GC2 Lab students and Fall 2023 GC1 Lab students. Therefore, there is no overlap of the participant population between the two semesters; however, students in GC2 Lab have likely previously experienced specifications grading in GC1 Lab. Students were recruited via an email invitation which contained a link to a Qualtrics survey where consent was obtained before participants completed study measures. Students were given ample time during their penultimate scheduled laboratory period to participate in the study if they chose to do so. The survey was left open for two weeks following this lab session to provide students who were absent during the session or had not finished the survey to fully participate in the study. There was no incentive to participate in the study. All 24 items of the pilot version of the PGS instrument were shown in a randomized order for each participant. Most participants completed the instrument within 5 minutes. After listwise deletion of participants who did not complete all PGS instrument items, the consent rate for the Spring 2023 GC2 Lab was 74.4% (n = 648; N = 871), and the consent rate for the Fall 2023 GC1 Lab was 62.0% (n = 1,031; N = 1,662). Table 9.1 shows the participant demographics of the two lab courses. Note: gender, international student status, and race/ethnicity information were obtained through a beginning-of-semester

survey assignment; participants with no data available either did not complete the assignment or enrolled in the course after the assignment was due.

Table 30. Participant demographics

	GC2 Lab (*n* = 648)	GC1 Lab (*n* = 1,031)
Gender		
Man	136 (21%)	402 (39%)
Woman	448 (69%)	577 (56%)
Non-binary or other gender identity	5 (<1%)	5 (<1%)
Prefer not to answer	1 (<1%)	7 (<1%)
Data not available	58 (9%)	40 (4%)
Generation status		
First generation	80 (12%)	176 (17%)
Continuing generation	568 (88%)	855 (83%)
International student status		
International student	8 (1%)	35 (3%)
Domestic student	583 (90%)	956 (93%)
Prefer not to answer	2 (<1%)	-----
Data not available	55 (8%)	40 (4%)
Race/ethnicity		
Black, Afro-Caribbean, or African American	31 (5%)	71 (7%)
East Asian or Asian American	84 (13%)	173 (17%)
Latino or Hispanic American	15 (2%)	36 (3%)
Middle Eastern or Arab American	18 (3%)	19 (2%)
Multiracial	53 (8%)	103 (10%)
Non-Hispanic, White, or Euro-American	286 (44%)	440 (43%)
South Asian or Indian American	79 (12%)	130 (13%)
Indigenous American	-----	3 (<1%)
Pacific Islander	1 (<1%)	-----
Other	3 (<1%)	3 (<1%)
Prefer not to answer	23 (4%)	13 (1%)
Data not available	55 (8%)	40 (4%)

Data analysis

Statistical analyses were carried out in RStudio version 2023.12.0 running R version 4.3.2 R Core Team, 2023). Unless otherwise noted, all statistical tests were carried out using base R packages and functions. Scree plot analysis, parallel analysis, the Kaiser–Meyer–Olkin measure, Bartlett's test for sphericity, and Mardia's test of multivariate normality were performed using the "psych" package (Revelle, 2024). Factor score plots were generated using the "ggplot2" package (Wickham, 2016).

Factor analyses. Factor analysis is a data reduction and interpretation technique used to describe the variability among observed and correlated variables (*i.e.,* instrument items) in terms of unobserved (or latent) variables (or constructs) that underlie the set of observed variables (Bandalos, 2018). Unobserved variables are referred to as factors. Exploratory factor analysis (EFA) is used to identify the structure of the underlying factors in the observed variables and is typically used where there has been minimal research regarding the structure of the construct of interest, such as in this case (Bandalos, 2018). Confirmatory factor analysis (CFA) is a structural equation modeling technique that tests whether a factor structure on a set of data is consistent with a specified factor structure from theory or research, such as from the results of an EFA (Bandalos, 2018).

EFA was first performed to identify the factor structure, and then CFA was used to confirm the identified factor structure. The GC2 Lab (Spring 2023) data set was split into two: a training set (n = 324) and a validation set (n = 324) from complete responses to the PGS instrument. EFA was performed on the Spring 2023 GC2 Lab training set, and was carried out using the factanal() function. CFA were performed on the Spring 2023 GC2 Lab validation set and the Fall 2023 GC1 Lab data and were carried out using the "laavan" package (Rosseel, 2012). Model parameters were estimated using the robust maximum likelihood (MLR) estimator due to multivariate non-normally distributed data. McDonald's ω coefficients were calculated using the scaleStructure() function in the "ufs" package (Peters & Gruijters, 2023), a new version of the package formerly known as "userfriendlyscience," as described in Komperda *et al.* (2018).

Model fit for factor structure was evaluated using the X^2 statistic, comparative fit index (CFI), Tucker–Lewis index (TLI), standardized root-mean square residual (SRMR), and root mean square error of approximation (RMSEA). Acceptable cutoff criteria include CFI and TLI $\geq$ 0.95, SRMR

< 0.08, and RMSEA < 0.06 (Bentler, 1990; Bentler & Bonnet, 1980; Hu & Bentler, 1999; Steiger, 1990; Tucker & Lewis, 1973).

Measurement invariance. Measurement invariance testing is a technique within a CFA framework that is used to provide validity evidence that the internal structure of an instrument holds for different groups of people (Brown, 2015; Rocabado et al., 2020). We test for measurement invariance for PGS instrument administrations among different courses (*i.e.,* GC1 and GC2 Labs) since the two courses significantly differ in the enrolment of engineering students. A series of increasingly restrictive CFA models were tested where various constraints (*i.e.,* factor structure, item loadings, item intercepts, item residuals) are set equal between groups (Dimitrov, 2010; Millsap, 2011; Rocabado et al., 2020). Testing for measurement invariance involves an additive stepwise testing process:

(1) *Configural invariance*: same factor structure between groups.
(2) *Metric invariance*: same factor structure and equal item loadings between groups.
(3) *Scalar invariance*: same factor structure, equal item loadings, and equal item intercepts between groups.
(4) *Strict invariance*: same factor structure, equal item loadings, equal item intercepts, and equal item residuals between groups.

Model fit for measurement invariance was evaluated based on the same absolute and relative fit statistics as described above for factor structure. Testing for measurement invariance ends when a CFA model fails to meet acceptable fit statistics.

Results and Discussion

Development of the Perceptions of Grading Schemes instrument

The Perceptions of Grading Schemes (PGS) instrument was developed following guidelines from the *Standards for Educational and Psychological Testing* American Educational Research Association *et al.*, 2014). Four different sources of validity evidence are presented to substantiate that the PSG instrument collects valid and reliable data: (1) test content, (2) response process, (3) internal structure, and (4) internal consistency.

Test content. Test content refers to whether instrument items capture the intended domain. Here, the intended domain is perceptions of grading schemes. We used Nilson's theorized outcomes of specifications grading (Figure 17) to identify eight student-centered outcomes to include in the domain: reflect student learning outcomes, motivate students to learn, motivate students to excel, reduce student stress, make students feel responsible for their grades, minimize conflict between faculty and students, give students feedback they will use, and make expectations clear. Discouraging cheating is the ninth student-centered theorized outcome, but we as researchers do not feel ethically comfortable asking students their perceptions of whether they have an inclination of cheating because of the grading scheme, and therefore excluded this outcome from the target domain.

We conducted a literature search to identify items and instruments that could be related to perceptions of grading but could not find instruments or items that fit this need. Therefore, we wrote our own items related to these eight theorized student-centered outcomes. The PGS instrument was developed through multiple iterations. The initial draft was written by three chemistry education researchers (authors BJY, HM, and MS) where each of the eight student-centered outcomes (*i.e.,* constructs) contained three to four items.

Testing for test content was established by consulting with experts in the target domain. The initial draft was first reviewed by a chemistry instructor who uses specifications grading (author LM) and two specifications grading experts (authors ACS and MSP). Discussions and feedback obtained were used to modify the instrument. In the initial draft, there was a broad statement that preceded each block of three to four items: *"compared to the traditional grading in my other science, technology, engineering, and mathematics (STEM) courses…"* Each item referred to specifications grading (*e.g.,* "*in specifications grading, I only want to learn what is strictly necessary*") and was measured on a five-point Likert scale ranging from strongly disagree to strongly agree. Discussions with experts centered around how the items were biased toward specifications grading. To remove this bias, we decided to change the broad statement to *"Reflect upon your experiences in other science, technology, engineering, and mathematics (STEM) courses that use traditional, point-based grading schemes, and your experiences with the specifications grading used in this course. Each statement is more characteristic of which grading scheme?"* This change removed the comparator within the item which now becomes "*I only want to learn what is strictly necessary"* and is now measured on a five-point bipolar scale ranging from "more traditional grading" on one end to "more specifications grading" on the other end with "both grading schemes equally" in the middle. This modification removes comparison between traditional and specifications grading in the item and no longer biases one grading scheme. The second draft was used subject to further expert review.

Modifications resulted in a second draft which was reviewed by a team of five chemistry education researchers. Further changes were made to the question in the statement. In particular, *"Each statement is more characteristic of which grading scheme?"* was changed to *"To what degree does each statement represent your experiences with traditional and/or specifications*

grading?" The measurement scale was also modified to "more representative of traditional grading" on one end to "more representative of specifications grading" on the other end with "equally representative" in the middle. Additionally, the experts helped narrow the number of items for each construct to three for consistency and refined the language of items; for example, the item "*I only want to learn what is strictly necessary"* was changed to *"I want to only learn what is strictly necessary to earn the grade I want."* The third draft was used to obtain response process data.

Response process. Response process refers to how instrument items and methods are interpreted by the study participants in the way that the instrument developers had intended. Testing for response process was conducted via in-person cognitive interview focus groups (Ryan, Gannon-Slater, & Culbertson, 2012; Yin, 2018). Recruitment emails were sent to all eight GC2 Lab UTAs in the Spring 2023 semester. These UTAs previously took the general chemistry lab sequence and thus had experience as students with specifications grading and were now using the specifications grading scheme as instructors. However, after multiple attempts, no participants from this sample were able to be recruited. Therefore, a convenience sample of six GC2 Lab GTAs were recruited. While this sample of GTAs are not the intended participants of the PGS instrument, these GTAs have first-hand knowledge of and experience grading using specifications grading with their undergraduate students.

Cognitive interview focus groups were conducted in three rounds with two GTA participants in each round. Verbal consent was obtained, and the cognitive interview focus groups were audio-recorded. The focus groups lasted 30–60 minutes and engaged the participants in a think-aloud process for their interpretation and understanding of PGS instrument items and response options. Feedback was obtained and items were revised between each round. Cognitive

interview focus groups led to further minor refinements of the PGS instrument from which data were collected. For example, feedback from the first focus group resulted in defining what is meant by traditional and specifications grading. Additionally, the item *"I want to only learn what is strictly necessary to earn the grade I want"* was changed to *"I am motivated to only learn what is strictly necessary to earn the grade I want"* to align the item with motivation. Feedback from the second focus group led to further refinement of the item to its final wording, *"I am motivated to only learn course material that is strictly necessary to earn the grade I want"* to explicitly state the subject is the course material. Figure 18 provides an exemplar summary of the modifications made to the original draft yielding the final version of the PGS instrument.

Original Draft

Compared to the traditional grading in my other science, technology, engineering, and mathematics (STEM) courses…

	strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
In specifications grading, I only want to learn what is strictly necessary.	○	○	○	○	○

After Expert Panel (Test Content)

Reflect upon your experiences in other science, technology, engineering, and mathematics (STEM) courses that use traditional, point-based grading schemes, and your experiences with specifications grading in this course.

To what degree does each statement represent your experiences with traditional and/or specifications grading?

	More representative of traditional grading		Equally representative		More representative of specifications grading
I want to only learn what is strictly necessary to earn the grade I want.	○	○	○	○	○

After Cognitive Interview Focus Groups (Response Process) – Final Version

The next section asks you about your experiences with different grading schemes.

Traditional grading is what is typically used where each assignment is given points or a letter grade.

Specifications grading is what is used in this course where each assignment is given either a "Mastered" or "Not Yet Mastered" grade.

Reflect upon your experiences in other science, technology, engineering, and mathematics (STEM) courses that use traditional grading schemes, and your experiences with the specifications grading scheme used in this course.

To what degree does each statement represent your experiences with **traditional** (i.e., other STEM courses) and/or **specifications** grading (i.e., this course)?

	More representative of **traditional grading**	Slightly more representative of **traditional grading**	Equally representative	Slightly more representative of **specifications grading**	More representative of **specifications grading**
I am motivated to only learn course material that is strictly necessary to earn the grade I want.	○	○	○	○	○

Figure 18 Exemplar summary of PGS instrument modifications

The 24-item pilot version of the Perceptions of Grading Schemes (PGS) instrument was used for data collection. All 24 items are provided in Appendix G.

Validity and reliability of the Perceptions of Grading Schemes instrument

Validity of measurements obtained from the PGS instrument was shown through internal structure studies. Internal structure of an instrument refers to evidence of the relationship between items and constructs intended to be measured. Internal structure was investigated through exploratory and confirmatory factor analyses.

Exploratory factor analysis. The GC2 Lab (Spring 2023) training set (n = 324) was used for EFA to determine a factor structure for the data. Prior to conducting the EFA, the data sets were checked for suitability using the Kaiser–Meyer–Olkin (KMO = 0.86) measure of sampling adequacy and Bartlett's test of sphericity ($p < 0.001$). Principal axis factoring (PAF) methods were used on the EFA data set. To determine the number of latent factors that underlie the data, Kaiser's criterion, scree analysis, and parallel analysis were used alongside theoretical considerations. Typically, Kaiser's criterion is usually wrong (Fabrigar & Wegener, 2012) and parallel analysis is most accurate (Velicer & Fava, 1998). Scree analysis can be used to supplement the choice in the number of factors, but theoretical considerations must also be included (Gorsuch, 1983). No method has been found to correctly identify the number of factors (Gorsuch, 1983; Pett, Lackey, & Sullivan, 2003); therefore, it is necessary to use multiple methods to carefully identify the most appropriate factor solution (Gorsuch, 1983; Loehlin & Beaujean, 2017; Pett, Lackey, & Sullivan, 2003; Watkins, 2018). Kaiser's criterion and parallel analysis suggested a six-factor solution; however, scree analysis and theoretical considerations suggested a five-factor solution. Ultimately, a five-factor solution was chosen. This solution is in alignment with the scree analysis and theoretical considerations because only two items loaded onto the sixth factor when Kaiser's criterion and parallel analysis were used which yielded an unreliable factor.

Factors were theorized to be correlated, and thus oblique rotation (promax) was selected for the analysis. EFA results were used to identify possibly problematic items that should be removed before continuing into a confirmatory framework. Items were removed if there was a lack of association with any factor. Items were also removed after theoretical considerations about item wording that may not reflect the nature of the factor and additionally retained lower factor loadings (*i.e.,* < 0.50). The result is a five-factor solution (*i.e., Stressful*, *Clear Expectations*, *Reflect Student Learning Outcomes*, *Useful Feedback*, and *Promotes Intrinsic Motivation*) with three items per factor. These five factors align with five of the eight theorized student outcomes with the *promoting intrinsic motivation* factor combining the "motivation to learn" and "motivation to excel" outcomes. The "make students feel responsible for their grades" and "minimize conflict between faculty and students" outcomes could not be identified in the five-factor solution and thus cannot be measured by the PGS instrument. Definitions of the five factors are described in Table 9.3. Item pattern loadings onto the five-factor solution are provided in Table 32, and factor correlations are provided in Table 33. The final version of the instrument, which resulted from the EFA, can be found in Appendix G.

Table 31. PGS factor definitions

Factor	Definition
Stressful	The grading scheme promotes students' anxiety and stress
Clear Expectations	The grading scheme makes expectations for success in the course clearer
Reflect Student Learning Outcomes	The grades received under the grading scheme reflect students' learning
Useful Feedback	The grading scheme allows students to receive useful feedback
Promotes Intrinsic Motivation	The grading scheme promotes students' intrinsic motivation to learn course content

Table 32. EFA pattern loadings

	Factor				
Label	Stressful	Clear Expectations	Reflect Student Learning Outcomes	Useful Feedback	Promotes Intrinsic Motivation
Q21		–1.04	0.194	**0.800**	–0.106
Q22				**0.773**	
Q23				**0.687**	
Q31	0.312	0.300	–0.273		
Q32			**0.781**		
Q33			**0.761**		
Q41	0.186	0.198	0.408	0.109	
Q42	0.141	0.244		0.157	0.198
Q43			**0.548**		
Q61	**0.701**	–0.122	0.109		–0.153
Q62	**0.762**			–0.121	
Q63	**0.817**				
Q71	0.389	–0.200			
Q72		0.386			
Q73	0.185	0.334			
Q81	0.255				0.112
Q82	0.296				
Q83	0.204		–0.178		0.238
Q101			0.146		**0.671**
Q102	–0.130			–0.108	**0.801**
Q103			–0.109		**0.777**
Q111	–0.103	**0.738**			
Q112	–0.167	**0.644**			0.139
Q113		**0.885**	0.138	–0.118	

Note: Bold indicates pattern loadings > 0.500

Table 33. EFA factor correlations

Factor	Stressful	Clear expectations	Reflect student learning outcomes	Useful feedback	Promotes intrinsic motivation
Stressful	1.00				
Clear Expectations	0.18	1.00			
Reflect Student Learning Outcomes	–0.32	–0.30	1.00		
Useful feedback	0.47	0.07	–0.40	1.00	
Promotes Intrinsic Motivation	–0.51	–0.33	0.60	–0.50	1.00

Confirmatory factor analysis. The factor structure suggested by the EFA was then tested using CFA. CFA provides evidence as to whether the proposed model fits new data without a predefined model. CFA was performed using the refined factor structure from the EFA. The validation set of the GC2 Lab data (n = 324) was used for the evaluation of model fit to our data.

Mardia's test of multivariate normality indicated that the data are not multivariate normal (kurtosis > 5, $p < 0.001$); therefore, the robust maximum likelihood (MLR) estimator was used to estimate model parameters to accommodate our multivariate non-normally distributed data. Results from the CFA for the GC2 Lab data indicate that the proposed factor structure exhibits acceptable fit to our data: $X^2(80, n = 324) = 104.03$, $p < 0.001$; CFI = 0.99; TLI = 0.98, SRMR = 0.04, and RMSEA = 0.03. All fit statistics meet acceptable cutoff criteria: CFI and TLI ≥ 0.95, SRMR < 0.08, and RMSEA < 0.06 (Bentler, 1990; Bentler & Bonnet, 1980; Hu & Bentler, 1999; Steiger, 1990; Tucker & Lewis, 1973). Results from the CFA for the GC1 Lab (Fall 2023) data set also indicate that the proposed factor structure exhibits acceptable fit to our data: $X^2(80, n = 1{,}031) = 112.75$, $p < 0.01$; CFI = 0.99; TLI = 0.99, SRMR = 0.02, and RMSEA = 0.02. All fit statistics meet acceptable cutoff criteria as previously described.

Nearly all items yielded high standardized factor loadings (all loadings ≥ 0.611), indicating a strong relationship between each item and the corresponding latent construct. Table 6 shows the five-factor PGS instrument with corresponding item statements and factor loadings for GC2 Lab. Figure 19 shows the Spring 2023 GC2 Lab data fit to a CFA model with standardized factor loadings. Similar factor loadings were obtained for the Fall 2023 GC1 Lab data; the CFA model with standardized factor loadings for GC1 Lab can be found in Appendix G.

Table 34. Factors and items for the PGS instrument

Factor	Label	Item	Factor loading
Stressful	Q61	I am anxious about my final letter grade.	0.778
	Q62	I am anxious about receiving a bad grade on individual assignments.	0.878
	Q63	I am anxious about making mistakes on individual assignments.	0.836
Clear Expectations	Q111	I understand what is required to achieve a particular final letter grade.	0.785
	Q112	I understand the expectations of each course assignment.	0.764
	Q113	I understand the expectations for success in the course.	0.801
Reflect Student Learning Outcomes	Q21	My grades on assignments represent what I understand about the course topics.	0.815
	Q22	My grades on assignments capture my understanding of the course material.	0.820
	Q23	How much I learned is reflected in my grades on assignments.	0.814
Useful Feedback	Q101	I pay attention to the written feedback I receive on my assignments.	0.711
	Q102	The written feedback I receive on my assignments is helpful to my learning.	0.799
	Q103	I am able to use the written feedback I receive on my assignments to improve my future work.	0.697
Promotes Intrinsic Motivation	Q32	I am motivated to learn as much course material as possible.	0.751
	Q33	I am motivated to thoroughly understand the course material.	0.817
	Q43	I am motivated to do my best on each assignment.	0.611

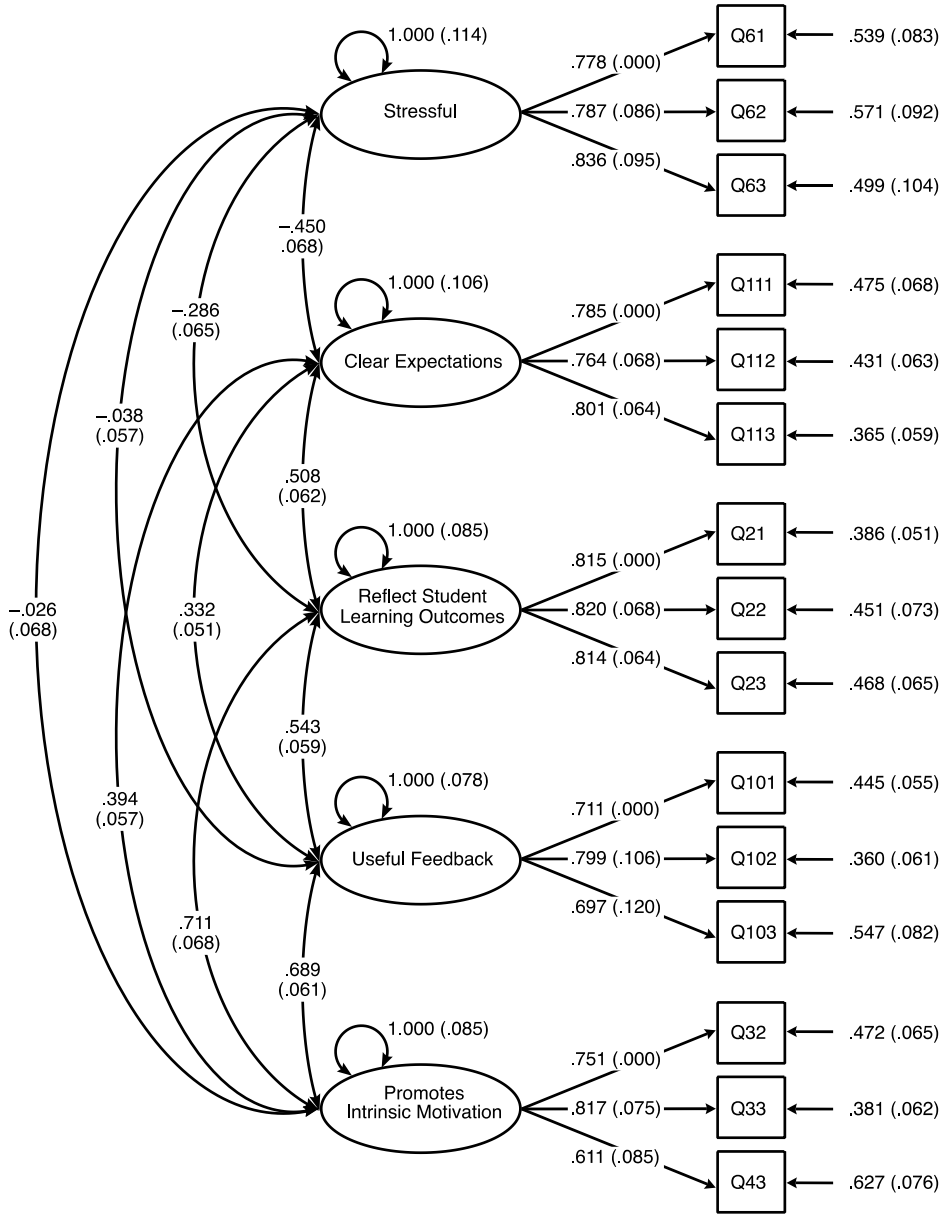

Figure 19. CFA model of Spring 2023 GC2 Lab with standardized factor loadings

Reliability. Reliability refers to the extent to which a measure (*i.e.,* construct) yields the same score each time it is administered. One measure of reliability is internal consistency. Internal consistency refers to how well a set of items that describes the same measure relate to one another. Cronbach's alpha (α) is the most typically reported measure of reliability; however, Cronbach's α assumes equal factor loading for all items (Kline, 2016; McDonald, 1981). McDonald's omega (ω) is another measure of internal consistency that is similar to Cronbach's α but is a more appropriate reliability measure when item loadings are unequal (Komperda, Pentecost, & Barbera, 2018). To determine single-administration reliability of the responses for each PGS factor, McDonald's ω coefficients were calculated and is interpreted similarly to Cronbach's α where a coefficient closer to 1 indicates a more reliable measurement. For the GC2 Lab data, McDonald's ω coefficients for each factor exceed the recommended cutoff criterion of 0.70 (Table 35) (Komperda, Pentecost, & Barbera, 2018; McDonald, 1999). For the GC1 Lab data, McDonald's ω coefficients for each factor also exceed the recommended cutoff criterion (Table 35). These reliability coefficients provide evidence for reliability measures and together these findings support internal consistency and reliability of the data generated from the PGS instrument with students in a specifications-graded general chemistry laboratory course.

Table 35. McDonald's ω coefficients for factors in the PGS instrument

Factor	GC2 Lab (n = 324)	GC1 Lab (n = 1,031)
Stressful	0.84	0.82
Clear Expectations	0.82	0.81
Reflect Student Learning Outcomes	0.86	0.84
Useful Feedback	0.78	0.82
Promotes Intrinsic Motivation	0.78	0.81

Course measurement invariance. Measurement invariance is evidence to suggest that the same latent constructs are being measured across a grouping variable (Bandalos, 2018; Rocabado et al., 2020). Typically, this grouping variable is a demographic variable (*e.g.,* gender, major) or

time points of instrument administration. In this study, there is no theoretical or empirical evidence to suggest that students of differing demographic variable would respond differently to instrument items. However, the student population differs significantly between GC1 Lab and GC2 Lab since all engineering students only take GC1 Lab and do not continue to GC2 Lab. Therefore, in this study, the grouping variable is by course (*e.g.,* GC2 Lab versus GC1 Lab) to further provide evidence that the PGS instrument is measurement invariant across courses. Establishing measurement invariance is essential to make group comparisons when using the PGS instrument (American Education Research Association *et al.*, 2014). Table 36 shows the stepwise measurement invariance model fit statistics and model comparison between GC2 Lab and GC1 Lab.

Strict invariance was established for course groups. This evidence is indicated by the no change in most fit statistics (ΔCFI = 0.00, ΔTLI = 0.00, ΔRMSEA = 0.00, and ΔSRMR = 0.00) which are within acceptable change cutoff thresholds to establish strict invariance: ΔCFI $\leq$ 0.01, ΔRMSEA $\leq$ 0.015, and ΔSRMR $\leq$ 0.01 (Chen, 2007). Strict invariance implies that there is evidence to warrant the comparison of groups using composite scores taken directly as an average of the observed items (Gregorich, 2006; Sass, 2011).

Students' perceptions in a general chemistry laboratory course sequence

Evidence of valid and reliable data produced from the PGS instrument suggests that it is reasonable to produce factor scores for interpretation. Two standard methods of computing scores from items in a factor exist: (1) take the average of items in a factor and (2) use the measurement model to estimate values based on item factor loadings and latent variable correlations. Both methods for calculating factor scores have their own advantages and value (see McAlpin et al., 2022).

Strict invariance was achieved in measurement invariance studies which allows for the comparison of groups using composite scores. Therefore, in this study, we will use the term "factor score" to represent the average value of the items in a factor when the scale is centered and standardized. That is, PGS instrument items are centered on the measurement scale where "equally representative" becomes the value 0, and the scale is standardized where "more representative of traditional grading" becomes the value –1 and "more representative of specifications grading" becomes the value +1. Then, the average of the three items in a factor are averaged. This method of calculation is used for its simplicity and accessibility in interpretation to a broader audience of potential users of the PGS instrument. The centering and standardization of the scale adds to the interpretability such that the scale is bound from –1 to +1, a negative value represents that students' perception of the factor leans toward traditional grading, and a positive value represents that students' perception of the factor leans toward specifications grading.

Table 36. Course measurement invariance fit information and model comparisons

Model	X^2	*df*	*p*-value	ΔX^2	Δdf	*p*-value	CFI	ΔCFI	TLI	ΔTLI	RMSEA	ΔRMSEA	SRMR	ΔSRMR
Baseline – GC2 Lab	104.0	80	<0.001	—	—	—	0.99	—	0.98	—	0.03	—	0.04	—
Baseline – GC1 Lab	112.8	80	<0.001	—	—	—	0.99	—	0.99	—	0.02	—	0.02	—
Configural	217.3	160	0.002	—	—	—	0.99	—	0.99	—	0.03	—	0.03	—
Metric	232.6	170	0.001	15.3	10	0.12	0.99	0.00	0.99	0.00	0.03	0.00	0.03	0.00
Scalar	255.3	180	<0.001	22.7	10	0.01	0.99	0.00	0.99	0.00	0.03	0.00	0.03	0.00
Strict	282.0	195	<0.001	26.7	15	0.03	0.99	0.00	0.99	0.00	0.03	0.00	0.03	0.00

Students' perceptions of specifications grading in a general chemistry laboratory course versus traditional grading in other STEM courses can be analyzed through factor scores from the

PGS instrument. The 95% confidence intervals for the mean factor score of each of the five factors for the Spring 2023 GC2 Lab course is shown in Figure 20 and for the Fall 2023 GC1 Lab course in Figure 21. Confidence interval overlapping with zero (*i.e.,* "equally representative") indicates that the mean factor score is not statistically significantly different from zero. Confidence intervals that are negative and do not contain zero indicate a "lean" of students' perceptions of that factor toward traditional grading in other STEM courses; confidence intervals that are positive and do not contain zero indicate a "lean" of students' perceptions of that factor toward specifications grading in this course. For both courses (*i.e.,* GC1 Lab and GC2 Lab), the mean values of the factor scores are similar with the 95% confidence intervals indicating that students are perceiving the course's implementation of specifications grading similarly for each factor. A *negative* score for the *Stressful* factor and a *positive* score for the four other factors (*i.e., Clear Expectations*, *Reflect Student Learning Outcomes*, *Useful Feedback*, and *Promotes Intrinsic Motivation*) indicate that specifications grading has achieved that outcome in the implementation.

Data obtained using the PGS instrument from the GC1 and GC2 Lab course suggest that these implementations of specifications grading may not be achieving the theorized student outcomes (Figures 20 and 21). In both courses, students perceive traditional grading to be more stressful but also better at promoting intrinsic motivation when compared to the course's specifications grading scheme. However, students perceive that the specifications grading scheme provides clearer expectations when compared to traditional grading schemes. Students also perceive that the traditional and specifications grading schemes equally reflected student learning outcomes and yielded useful feedback.

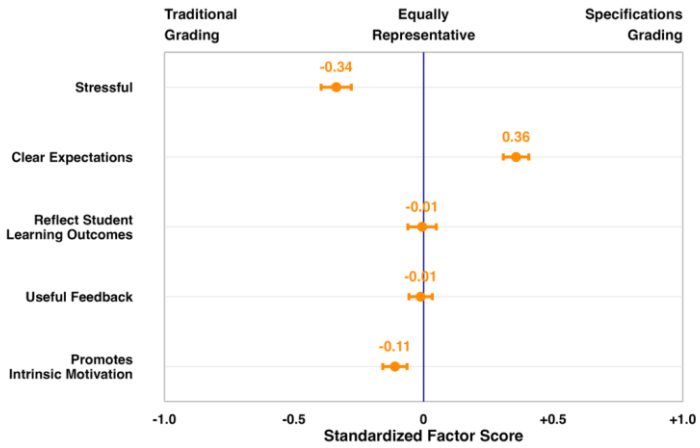

Figure 20. Student perceptions in the Spring 2023 GC2 Lab (n = 324).
95% confidence intervals of the mean factor scores are shown.

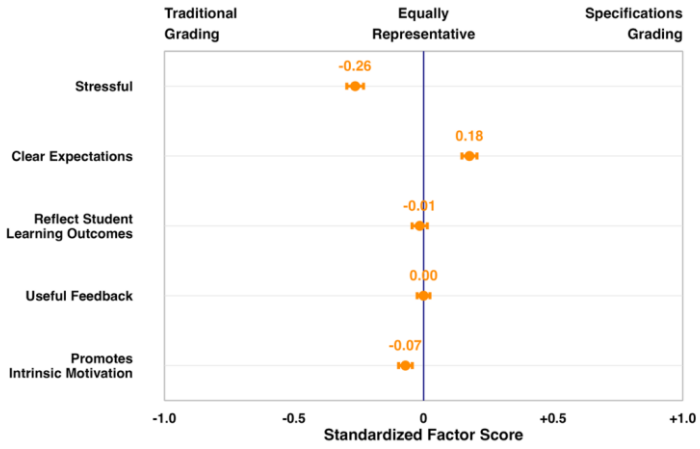

Figure 21. Student perceptions in the Fall 2023 GC1 Lab (n = 1,031).
95% confidence intervals of the mean factor scores are shown.

Limitations

There are several limitations of note for this work. First, the Perceptions of Grading Schemes (PGS) instrument does not capture all of Nilson's theorized student-centered outcomes. Namely, the PGS instrument cannot provide a measure of *making students feel responsible for their grades* and *minimizing conflict between faculty and students* as these outcomes could not be identified as factors in the EFA. Future iterations of the PGS instrument may attempt to measure these two student-centered theorized outcomes.

Second, we demonstrate that the PGS instrument provides valid and reliable data in two sequential general chemistry laboratory courses that use virtually identical specifications grading schemes and are coordinated by the same instructor at a single institution. The chemistry education literature, and more broadly the education literature, on specifications grading indicates that instructors' implementations of specifications grading vary greatly (Tsoi et al., 2019). Thus, there is currently no evidence to suggest that the PGS instrument will provide valid and reliable data outside of a general chemistry laboratory course; future studies should test the instrument in other chemistry and STEM courses including course types (*e.g.,* lecture-only, and combined lecture and lab courses).

Third, we do not have evidence for external validity or relations to other variables. In other words, we do not have other measures of similar variables to compare with PGS data. For example, measures exist for anxiety (e.g., Hensen & Barbera, 2019) and could be used as measures of convergent validity and criterion-related concurrent validity if the PGS instrument and anxiety measures are captured at the same point in time (American Educational Research Association *et al.*, 2014). Convergent validity refers to how closely related a test of a measure of a construct is related to a similar test of the same construct. Criterion-related concurrent validity refers to how

well one measure of a construct predicts a known measure of the same construct when administered at the same time. Providing evidence for external validity or relations to other variables would add to the validity of the PGS instrument.

Implications

The Perceptions of Grading Schemes (PGS) instrument can be used to provide an understanding of student perceptions of different implementations of specifications grading. The chemistry education literature on specifications grading detail vastly different implementations even within the same disciplinary course (cf. Bunnell et al., 2023; Noell et al., 2023; Tsoi et al., 2019). Comparison of PGS factor scores allows for an easy comparison of the impact of different implementations of specifications grading on stressfulness, clear expectations, reflecting student learning outcomes, useful feedback, and promoting intrinsic motivation when compared with traditional grading in other STEM courses. The PGS instrument could also aid in informing understanding of what facets of specifications grading implementations yield the theorized outcomes through an analysis of PGS factor scores and descriptions of the implementation (*e.g.,* course syllabus). While the PGS instrument is intended to be used with specifications grading, the PGS items and scales could be adapted for other alternative grading schemes and across disciplines.

The PGS instrument can also be useful for instructors. The short length of the PGS instrument (< 5 min.) makes it simple to administer during scheduled class time or as an out-of-class assignment. Factor scores can also be easily determined to yield insight into students' perceptions of the course's implementation of specifications grading. This insight can provide instructors with an understanding as to whether their implementation is achieving the theorized student outcomes. Administering the PGS instrument during different terms (*e.g.,* on-sequence

versus off-sequence) can additionally provide information with different populations of students. If instructors revise and refine their specifications grading scheme, longitudinal administration of the PGS instrument can provide further understanding about whether the refinements better achieve the theorized student outcomes. Overall, the PGS instrument can be used to formatively and iteratively inform implementation.

Conclusions

The Perceptions of Grading Schemes (PGS) instrument has been developed and is shown to produce valid and reliable data with multiple sources of evidence including test content, response process, internal structure, and internal consistency. The theorized student-centered outcomes presented in Nilson's (2015) book (Table 1) were used as a theoretical framework to guide the development of the instrument. The evaluation of this instrument was conducted with a yearlong general chemistry laboratory sequence that uses specifications grading (*i.e.,* GC1 and GC2 Labs) to show that it yields data with reproducible psychometric properties. The 15-item PGS instrument can be used to efficiently evaluate student perceptions of specifications grading in terms of stressfulness, clear expectations, reflecting student learning outcomes, useful feedback, and promoting intrinsic motivation. Evaluation of factor scores from the PGS instrument in a general chemistry laboratory sequence demonstrate that implementations of specifications grading may not be achieving some of Nilson's (2015) theorized student-centered outcomes and warrants further investigation.

Part 2 Conclusions and Future Directions

Conclusions

The collective findings from part 2 highlight that specifications grading in higher education chemistry courses is highly variable. Instructors' adoption of specifications grading is influenced by perceived benefits, including flexibility and enhanced student learning, despite concerns over instructor workload and student resistance. While some common trends in implementation are identified, such as the frequent use of two-level grading systems and revision opportunities, substantial variation remains, particularly in how final grades are determined. This suggests that specifications grading should not be viewed as a set grading system, but rather one that requires adaptation to the specific context of each course. This is furthered by the work regarding the Perceptions of Grading Schemes (PGS). Indeed, the developed instrument is effective in evaluating student perceptions; yet its application suggests that some anticipated student outcomes may not be fully realized. Thus, it is necessary to establish best practices within specifications grading to enable fidelity of implementation of this alternative grading scheme in chemistry courses. Together, these studies on specifications grading in chemistry higher education emphasize the importance of investigating educational innovations to understand both their methods of propagation and effectiveness when implemented.

Future Directions

The efforts towards understanding chemistry instructors' usage of specifications grading detailed in this thesis are the beginning steps of a much larger project. Future research will aim to characterize the attributes and experiences of students who thrive under specifications grading, as well as those who do not. By collecting both registrar data (e.g., race and ethnicity, gender, grades, first-generation status) and survey data from students enrolled in courses taught by the chemistry

instructors who use specifications grading, we will be able to capture students' characteristics such as caregiving responsibilities, employment, and other personal circumstances. These data will be used in conjunction with student grades to determine students' personal factors that may affect the effectiveness of specifications grading in chemistry classrooms. Additionally, chapter 5 analysis and registrar data will be used to determine which features of specifications grading are most effective at optimizing student outcomes. This will allow us to understand which aspects of the implementation can minimize opportunity gaps and maximize student-focused outcomes, providing insights into how to tailor specifications grading practices to specific contexts.

Finally, this project will lead to a participatory action research study wherein instructors are actively engaged with education researchers to improve their specifications grading practices. We will collaborate with instructors to 1) design a refinement plan for their courses, based on the insights gained from student data and course-level analyses, and 2) develop an evaluation plan to monitor the effects of these refinements on student outcomes. The research team will provide ongoing support in data collection and analysis, including interviews, registrar data, and surveys. The outcome of this participatory action research study will be a set of evidence-based recommendations for instructors seeking to optimize their specifications grading implementation. Additionally, the study will offer valuable insights into the opportunities and challenges that both instructors and students encounter when adopting specifications grading.

These three future directions aim to advance the field of chemistry education by offering a deeper understanding of students' perspectives on and experiences with specifications grading and its impact on their learning. Additionally, the study will provide a set of features and recommendations for effective specifications grading implementation, tailored to different

learning environments and grounded in the lived experiences and needs of both instructors and students.

Overarching Themes of Dissertation

Several common themes emerge from the work described in this dissertation, reflecting the complexities and challenges faced by STEM instructors in higher education. These themes center on the importance of providing support for instructors, from their usage of pedagogical practices to their ability to adapt to various educational environments.

Firstly, the work described herein points to the role of instructors as vital agents in pedagogical innovation, whether in the form of reflective practice or the adoption of new grading systems. Chapter 3 highlights how instructors choose to implement specifications grading based on their perceptions of its benefits and challenges, underlining instructors' agency in adopting or adapting new practices. Chapters 1 and 2 reaffirm this sense of agency, as instructors reflect on their experiences and emotions, which may shape their approach to future teaching situations. This common theme reinforces the idea that teaching is not a passive practice, but rather a dynamic process where instructors must have agency to make informed, thoughtful decisions.

Another critical theme is the need to support instructors in their chosen pedagogical practices, whether through fostering reflective thinking, addressing emotional experiences, or adopting new grading methods. Chapter 1 highlights that physics and astronomy instructors with limited teaching experience struggle to reflect at a level that promotes growth, suggesting a need for structured support to enhance their reflective practices. Similarly, chapter 2 identifies an array of emotional experiences among physics instructors, including many negative emotions and negative views of ones' self. The emotional challenges faced in the classroom signal a need for better preparation and holistic support for novice instructors. Chapters 3 through 5, examining chemistry instructors' adoption of specifications grading, suggests that instructors should be equipped with the necessary resources and guidance to understand and implement grading systems

that may differ from traditional methods. This is of particular importance with new innovations due to the unknown effects such pedagogical methods can have on students. Across these findings, it is apparent that supporting instructors – emotionally, pedagogically, and technically – is crucial for the advancement of STEM higher education.

A further recurring theme is the complexity and variability of pedagogical practices. This idea of variability is seen in chapter 1 where different instructors who are supplied with the same writing prompt, reflect on different levels; this then influences the potential for growth resulting from reflective practices. Chapter 2, while focused on emotions, also hints at this variability. Emotions and teaching strategies are influenced by individual circumstances and contexts, suggesting that a blanket approach to addressing emotions or teaching strategies may not be effective. Indeed, chapter 3 exemplifies how different instructors have differing motivations, even if they lead to the same pedagogical innovation. Chapter 4 emphasizes that there is no single approach to the implementation of pedagogical innovations; rather, the practice must be tailored to specific course types, student populations, and instructional goals. The recognition of this complexity calls for more individualized and context-sensitive approaches to teaching and pedagogical innovation.

Closely linked to the theme of variability is that of adaptability. Instructors in STEM higher education must be able to adapt their practices in order to facilitate learning. Chapter 4 highlights the need for instructors to adapt specifications grading to their particular course context, noting that the implementation of this grading system may evolve over time. This adaptability extends to the way instructors respond to challenges or unexpected emotional experiences in the classroom. Further, instructors themselves showed how they must adapt in the classroom through their plans

to address future situations, as detailed in chapter 1. Instructors must remain flexible to be effective in the classroom, making adjustments and improvements over time.

Together, these studies converge on the importance of supporting instructors through professional development, offering flexibility in the application of teaching practices, recognizing the complexity of those practices, and acknowledging the emotional and reflective dimensions of teaching. Addressing these interrelated factors will help foster more effective and transformative teaching in STEM education.

Appendix A. Part 1 IRB and instrument documents

Appendix A.1: Distributed Online Survey

Study Title: Collection and Evaluation of Reflective Writings of STEM Instructors
UVA IRB-SBS #: 5248

Start of Block: Introduction

Cycle Thank you for agreeing to reflect on your teaching! We know it is not easy but we also know from experience and the literature that this can be transformative for you and your students!

We will help you write your reflections by asking you guiding questions over the next few pages.

End of Block: Introduction

Start of Block: Critical Incident

CI explanation Think of a teaching situation.

When beginning the process of reflection, it is helpful to first identify a particular teaching situation on which to reflect.
For example, Think of a situation when you felt uncomfortable, unprepared, unqualified, or regretful as an instructor.
Think of a time where you felt dissatisfied, particularly with the outcome of your actions or with the effects of your words.
Recall situations that took you by surprise or simply made you pause to think in the moment.
Thinking of such situations may bring feelings of discomfort, especially during the examination of a challenging past experience. If you are uncomfortable with completing this activity, you can withdraw at any time.

End of Block: Critical Incident

Start of Block: Description of experience

Description - intro
What's the situation?

Describe the situation focusing on the facts of what occurred and what was said; feelings will be described in the next step.

Questions to consider include: *When and where did the situation occur?* *Who was present?* *What happened?*

☐ Example (2)

Display This Question:

If What's the situation? Describe the situation focusing on the facts of what occurred and what wa... = Example

Description - Exampl *I teach a general chemistry course. Yesterday, after an out-of-class review session before the midterm, a student came up to me. Everyone else had left the room, and it was just the two of us. She asked me what an intermolecular force (IMF) was, which is a subject covered in the first month of the course. I asked her which force she was talking about – London dispersion, dipole-dipole, or H-bonds – to which she replied that she didn't know what any of those were. I told her that she should already know this or have come to me earlier than two days before the test. Her eyes became wide, and she was very quiet while I explained what IMFs are and the different types. She then left without saying anything else. This morning, she did not come to class, which was the final review before the midterm on Friday.*

Description - Essay Please provide your description below.

_____

_____

_____

_____

_____

End of Block: Description of experience

Start of Block: Initial response

Initial response
How did you feel?

Now that you have described the facts of the situation, recall your feelings and thoughts that you had during the experience. Also try to incorporate the feelings of others involved as well as the impact this may have had on them.

Questions to consider include: *How did you feel before, during, and after the situation? What were you thinking during the situation? What do you think the other participants felt before, during, and after the situation?*

☐ Example (2)

Display This Question:

If How did you feel? Now that you have described the facts of the situation, recall your feelings an... = Example

Initial response-ex *Right before my interaction with this student, I was actually pretty happy. The review session had gone well. When the question was asked, I was initially confused because I didn't understand how she didn't address foundational topic before. I was a little bit shocked when she said that she had no idea what IMFs were in general. I think my blurted-out statement probably made them feel embarrassed or like they were going to fail the upcoming test. At the time, I was not concerned with what I said, as I was mainly worried about her possibly failing the course, and I also was frustrated with them for not seeking help before it was too late. After seeing that she chose not to come to class today, I am really worried that I may have discouraged her from the subject all together. I hope she isn't going to drop the class. If she does, I feel like it would be partially my fault.*

Q37 Please provide your response below.

_____

_____

_____

_____

End of Block: Initial response

Start of Block: Relate

relate - intro Has something similar happened before?

Compare the described situation to a previous experience that you've had. If no such prior experience exists, then simply type "N/A" in the text box below.

Questions to consider include: *Have you seen or experienced something similar before? If so, what was similar or different between other situation and the one you have described during this exercise?*

☐ Example (4)

Display This Question:

If Has something similar happened before?Compare the described situation to a previous experience th... = Example

relate - example *Weirdly, this is similar as to when I was working with a post-doc I hired a few years ago. They were international and had missed a deadline for filing for their Visa, and when they approached me to get help with this problem, the first words out of my mouth were "How could you miss the deadline?" It was a similar situation in that I spoke without thinking, and my concern for the other person involved in the conversation took over my thought processes to the detriment of my brain-to-mouth filter. This then resulted in me giving a response which was completely unhelpful and only served to increase another person's anxiety or feelings of "I messed up." However, with the post doc, I was speaking to an adult aged 28 who had just seriously jeopardized their job. Additionally, while I was their boss, we were close to being peers in both age and experience level. This is a direct contrast to the student who was either 18 or 19 and may not have even wanted to pursue STEM. They were also my student which forces an unfortunate power dynamic into the situation. I think the common factor between these two situations is that when my brain goes into "panic mode" I say whatever is on my mind, and even*

I myself do not always agree with those initial, panicky thoughts. I have the knowledge about how to correct this, but I need to work on making "think before you speak" a habit when I become frazzled rather than just a habit during more normal conversations.

---

relate - answer Please provide your answer below.

_____

End of Block: Relate

Start of Block: Evaluation

evaluation-intro Why were the outcomes as described?

Explore why certain aspects went well while others did not. Consider whether you had the adequate knowledge and skills to handle the situation. Finally, consider what someone who has experience with this type of situation would have done.

Questions to consider include: *Why did things go well or poorly? Did you feel equipped to handle the situation (at the time you experienced it)? How would have someone with experience in this type of situation handled things?*

☐ Example (2)

Display This Question:

If Why were the outcomes as described? Explore why certain aspects went well while others did not. C... = Example

evaluation- ex *When speaking with my student, it was good that they approached me to get help, and I explained the concept well. However, I made her, most likely, feel insecure and judged by my comment. Her not coming to the review the following day was likely due to my actions. I know my mentors from both undergrad and grad school would have first explained the concepts and then patiently asked their student if they were all right and if there were any extenuating circumstances that they needed an extension for. They would have approached with understanding rather than disbelief. I have the skills necessary to do the same thing, but*

apparently not the impulse control. As I think about it, I may have discouraged my student from the subject completely. Our department sees too few female applicants, and I hate to lose those that do choose to come here, especially due to my dumb, thoughtless comment.

Eval answer Please provide your evaluation below.

End of Block: Evaluation

Start of Block: Conclusion

conclusion-intro What will you do going forward?

Consider what you learned through this experience, particularly how you would react to similar situations in the future. Plan how you will develop the skills and/or knowledge you need to better handle future, similar situations.

Questions to consider include: *What did I learn from this situation? What skills or knowledge, if any, do I need to develop? How will I do this? How would I respond to similar situations in the future?*

☐ Example (2)

Display This Question:

If What will you do going forward?Consider what you learned through this experience, particularly ho... = Example

Q53 *I have a problem with blurting out my initial thoughts when I am surprised. I need to learn how to delay my reactions to unexpected situations. As a next step, I will become more mindful of thinking before speaking in all conversations to hopefully force that action to be an ingrained habit. In the future, I will be open to people coming to me with any level of question and will specifically phrase my words to not imply a negative judgment. Something I read about in a journal was the need for more formative feedback for teachers. I may have students give anonymous questions or comments part way through the semester, rather than just the end of*

course evaluations, to try and catch gaps in understanding like what occurred with this student.

Q54 Please provide your conclusion below.

_____

_____

_____

_____

_____

End of Block: Conclusion

Start of Block: Contact information

Q32 Please provide your contact information so that we can email you a copy of these reflections.

- First name (1) _____
- Last name (2) _____
- Email address (3) _____
- Name of your institution (4) _____

End of Block: Contact information

Appendix A.2: Recruitment Email for Collection of Written Reflections

Dear [Title. Name]:

I am an education researcher conducting a study on educators' reflective writings. Reflecting on one's own instructional practices has been identified as one of the most effective practices to grow as an instructor. Teaching reflections are also becoming more common requirements in promotion and tenure procedures. The goal of this study is to gather teaching reflections in order to understand how we can support instructors in writing reflections and analyze the content of said reflective writings.

If you are willing to participate in this study, please use the link below. Participation in this study is completely voluntary, and should you choose to participate, will take approximately 25 minutes of your personal time. There are no direct benefits should you participate in this study, and you may experience discomfort as a result of examining a challenging past experience. As a whole, the study will provide valuable insight into teacher-thinking.

Study Title: Collection and Evaluation of Reflective Writings of STEM Instructors
UVA IRB-SBS #: 5248

Link to Study: https://virginia.az1.qualtrics.com/jfe/form/SV_eya4XNWr1hSx5We

If you have any questions or would like to withdraw from the study after completing the survey, please contact:

Dr. Marilyne Stains

Professor; Dept. of Chemistry

University of Virginia

mstains@virginia.edu

Haleigh Machost

PhD Candidate; Dept. of Chemistry

University of Virginia

hrm6cw@virginia.edu

Thank you for your time!

Sincerely,

Haleigh Machost

Appendix A.3: Informed Consent Agreement for Collection of Written Reflections

Study Title: Collection and Evaluation of Reflective Writings of STEM Instructors
UVA IRB-SBS #: 5248

Please read this consent agreement carefully before you decide to participate in this study.

Purpose of the research study: Reflecting on one's own instructional practices has been identified as one of the most effective practices to grow as an instructor. Teaching reflections are also becoming more common requirements in promotion and tenure procedures. The goal of this study is to gather teaching reflections in order to understand how we can support instructors in writing reflections.

What will you do in the study: You will answer an online survey wherein you are prompted to write your reflections about a particular situation of your choosing. You should complete the survey on your own time or when prompted as a part of a course or workshop. This survey should *not* be completed during your teaching time. If you are participating in conjunction with a workshop or course, other reflective writings collected during the workshop or course will also be analyzed.

Time Required: This study will require about 25 minutes of your time.

Risks: If you participate in this in this study, there is the risk of experiencing discomfort resulting from the examination of a challenging past experience.

Benefits: There are no benefits to your participation in this study.

Confidentiality: Your identifying information will be gathered to aid in analysis and will be collected on the last page of the survey. Your name will be replaced with a pseudonym. All analysis will take place using the de-identified pseudonyms, and the data containing your true will be deleted.

Voluntary Participation: Your participation in this study is completely voluntary. If you are tasked with taking this survey as a part of a workshop or a course, your progress in the workshop or course will not be penalized should you choose to not take part in this study.

Right to withdraw from the study: You have the right to withdraw from the study at any time without penalty.

How to withdraw from the study: You may contact Dr. Marilyne Stains (mstains@virginia.edu) or Haleigh Machost (hrm6cw@virginia.edu) directly to withdrawal from the study.

There is *no penalty* for withdrawing from this study.

Once a request is received by Dr. Stains or Ms. Machost, all of your data will be deleted.

Payment: You will receive no payment for participating in the study.

Using data beyond this study: Data will not be shared beyond the research group, and will be de-identified before both analysis and sharing within the research group. Data may be used for different research endeavors such as case study. The identifiable data you provide in this study will be stored in UVA Box and in external hard drives securely stored in the research team office (in a locked storage container) per IRB requirements. All files used for analysis (with pseudonyms rather than identifiers) will be stored on a cloud system used by the research team. All raw data will be retained in a secure manner for 15 years and then destroyed.

If you have questions about the study, contact:
Dr. Marilyne Stains
Department of Chemistry, University of Virginia
409 McCormick Rd, Charlottesville, VA 22903
Telephone: 434-243-6430
email address: mstains@virginia.edu

To obtain more information about the study, ask questions about the research procedures, express concerns about your participation, or report illness, injury or other problems, please contact:
Tonya R. Moon, Ph.D.
Chair, Institutional Review Board for the Social and Behavioral Sciences
One Morton Dr Suite 500
University of Virginia, P.O. Box 800392
Charlottesville, VA 22908-0392
Telephone: (434) 924-5999
Email: irbsbshelp@virginia.edu
Website: https://research.virginia.edu/irb-sbs
Website for Research Participants: https://research.virginia.edu/research-participants UVA IRB-SBS # 5248

Study Agreement:

Option A: I agree to participate in the research study described above

Option B: I DO NOT agree to participate in the research study described above

Appendix B. Supplementary information for chapter 1

Appendix B.1: Inter-rater reliability process

Table B1.1. Summary of iterative inter-rater reliability for codebook creation and validation

Iteration	Documents	Summary of changes post-inter-rater analyses
Round 1	108, 121, 135, 147	• 'general' subcode deleted from 'communication with students' parent planning code • Small discrepancies between names of codes in written codebook and NVIVO file remedied
Round 2	131, 104, 151	• Renamed 'implement proven practices' planning code to 'implement successful strategies' • Altered definition of 'implement successful strategies' to remove intent and instead focus solely on practices/actions of instructors • Renamed 'student indirect complaints/evaluations' topic code to 'student indirect negative feedback'
Round 3	117, 126, 139	• Updated the codebook with more reader-friendly definitions of Larrivee for practical purposes. The old definitions, which are still in the codebook, are verbatim from the 2008 paper. An additional practical definition was discussed and agreed upon • The caveat that each category does not need to be coded for was clarified (i.e. reflections could not include any planning codes) • Definition of 'reexplain course material' planning code was expanded to explicitly include utilizing a different method or approach to the course material • Renamed the plan code 'no plan'; subcode: 'no desire' to 'no plan'; subcode: 'not responsible' • The example given for the planning code 'withdraw from students' was expanded to explicitly reference a change from a prior, more-invested state • Expanded the definition of 'meet students where they are' planning code to include provision of supplementary materials by instructor to help student reach a higher starting point • Verbally clarified that 'provide external resources' planning code is specific to resources outside of the instructor or instructor-provided materials. Includes things such as counselling services and the writing center
Round 4	111, 132, 151	• No changes
Round 5†	134, 141, 142	• No changes

†; Planning codes were not analyzed in round 5 due to high inter-rater reliability scores at the end of round 4

Appendix B.2: Data analysis

Table B2.1. Full codebook for plans described in instructors' reflections

Type of plan	Subcode	Definition	Prevalence of subcode (n)
Self-reliant	Communication with students – establish clear expectations	Instructor will formally establish the classroom norms and what is expected from students.	18
	Communication with students – communications from instructor	Instructor will give an explanation to the students, clarify a different communication, or otherwise address their students.	11
	Explicitly acknowledge or correct mistake	If the instructor makes an error, they will let their students know and/or will correct their mistake.	11
	Meet students where they are	Instructors will meet students where they are academically rather than having prior knowledge expectations.	9
	Alter coursework or objectives	Instructor will change an assignment, end class early, change course objectives (remove a topic from the test and therefore course), create alternative activity, etc.	8
	Provide external resources	Instructor will learn about school resources (counselling, tutors, writing center) as well as soft skills to provide help to students or to be able to point students in the direction of resources.	7
	Re-explain course material	When the whole or majority of the class is struggling, the instructor will explain the material, often starting from basic concepts or using a different method/approach.	6
	Offer extra help	Instructors will offer additional academic help to those in their classes who are struggling with the course material.	6
	Discuss privately/small group	To address the issue of concern, the instructor will speak to students privately or in a small group.	5
Self-preserving	Personal grace	Instructor will have more patience with students, have more emotional control, and be graceful in handling their own mistakes.	30
	Pre-planning	Instructor will better prepare for class/lab/office hours; Instructor will prepare for the unexpected situations.	24

Type of plan	Subcode	Definition	Prevalence of subcode (n)
Seeking knowledge	Implement successful strategies	Instructor will utilize proven strategies (EBIPS/group work/ assessment for learning/explanations other professors give students, etc.) Implementation is not intent related beyond the rationale that the strategy is known to be successful in some context.	12
	Communication with peers	Instructor will get advice from peers or solicit peers' opinions.	10
	Participate in professional development	Instructor will attend workshops, seek online resources, read the academic teaching and learning literature, etc. outside of what their department provides or requires.	8
	Communication with students – solicit student feedback	Instructor will solicit students' opinions on coursework/teaching styles verbally or through instructor created surveys. Instructor will be the one to prompt feedback from their students.	6
No plan	Did not write a plan	Instructor does not address the prompt asking for them to plan for future situations nor write any plan in other sections of the reflection.	9
Other types of plan	More evaluations or types of evaluations	Instructor will implement more evaluations to determine where their students are at and/or utilize additional assessment formats.	4
	Instructor starts from basics	If the instructor doesn't know how to answer a question, the instructor will start from the basics or foundational concepts and work on finding the answer from that point.	4
	Withdraw from students	The instructor will withdraw from the students and will only act in a strictly professional manner as required by the job description. This is in contrast to prior behavior where they were more invested.	4
	Gain teaching experience	Instructor plans to improve with experience; there is no mention of professional development.	4
	No plan – not responsible	The instructor does not plan for future situations because they believe it is not their responsibility to prevent or address such situations.	3
	No plan – no idea	The instructor does not plan for future situations because of a self-identified lack of	3

Type of plan	Subcode	Definition	Prevalence of subcode (n)
		knowledge about how to handle a future similar situation.	
	Participate in departmental initiatives	Instructor will participate in departmental initiatives for professional development, diversity equity and inclusion, curriculum design, etc.	2
	Gather evidence	If students cheat in the future, the instructor will gather evidence before acting.	1
	Administrative action	The instructor will take administrative action, such documenting incident(s), informing students' parents (if student is underaged), and working with the administration for the handling of a student's actions.	1
	Real-time aid from students	If an instructor cannot answer a question, the instructor will solicit or accept help from a student.	1

Table B2.2: Types of plan developed by instructor based on level of reflection.

Percentages represent the proportion of instructors on a specific level of reflection who utilized a specific combination of planning code categories. As the combinations are mutually exclusive, the sum is 100% for each level.

	Pre-reflection (n=23)	Surface (n=59)	Pedagogical (n=9)	Critical (n=7)
Self-reliant only	26%	17%	22%	43%
Self-preserving only	9%	19%	---	---
Seeking knowledge only	13%	7%	---	14%
Self-reliant AND Self-preserving	4%	29%	11%	43%
Self-reliant AND Seeking knowledge	13%	8%	33%	---
Self-preserving AND Seeking knowledge	4%	7%	11%	---
Self-reliant, Self-preserving AND Seeking knowledge	---	2%	22%	---
No plan	17%	8%	---	---
Other types of plan	13%	3%	---	---

Appendix C. Supplemental information for chapter 2

Appendix C1. Participant Pool

Of the 62 instructors who attended the July 2022 workshop, 52 submitted a written response to the survey for analysis, and 46 met the inclusion criteria. Among the 106 instructors who attended the June 2023 workshop, 57 submitted a written response to the survey for analysis, and 52 met the inclusion criteria. Finally, of the 60 workshop attendees from November 2023, 33 submitted a written response to the survey for analysis, and 27 met the inclusion criteria. In total, 125 instructors of physics or astronomy participated in this study.

Appendix C2. Data Analysis

Table C2.1. Codebook used in secondary coding of emotions expressed by instructors

Code	Definition ("Merriam Webster,")	Other emotions captured within the code
Hopeful	a: full of hope, wherein hope is to cherish a desire with anticipation: to want something to happen or be true	
Doubtful	a: lacking a definite opinion, conviction, or determination b: uncertain in outcome: undecided	
Confused	a: being perplexed or disconcerted b: disoriented with regard to one's sense of time, place, or identity	uncertain, unsure
Anxious	a: characterized by extreme uneasiness of mind or brooding fear about some contingency	panicked, concerned
Nervous	timid, apprehensive	stress, worried, hesitant, unsure, flustered, tense
Angry	a: feeling or showing anger, wherein anger is a strong feeling of displeasure and usually of antagonism.	annoyed, frustrated, exasperated, aggravated, furious, irritated, appalled, indignant
Happy	a: enjoying or characterized by well-being and contentment	content, eased, excited
Guilty	a: suggesting or involving guilt, wherein guilt is feelings of deserving blame especially for imagined offenses or from a sense of inadequacy.	regret, shame
Relieved	expressing or showing relief especially from anxiety or pent-up emotions, wherein relief is a removal or lightening of something oppressive, painful, or distressing.	
Surprised	feeling or showing surprise because of something unexpected, wherein surprise is the feeling caused by something unexpected or unusual	horrified, shock

Empathetic	involving, characterized by, or based on empathy, wherein empathy is the action of understanding, being aware of, being sensitive to, and vicariously experiencing the feelings, thoughts, and experience of another	
Confident	having or showing assurance and self-reliance	proud
Conflicted	experiencing or marked by ambivalence or a conflict especially of emotions, wherein conflict is the mental struggle resulting from incompatible or opposing needs, drives, wishes, or external or internal demands.	uncertain
Betrayed	treacherously abandoned, deserted, or mistreated	
Exhausted	depleted of energy: extremely tired	
Defensive	serving to defend or protect, wherein defend is to maintain or support in the face of argument or hostile criticism	offended
Disappointed	defeated in expectation or hope	dismay, displeased, discouraged
Overwhelmed	completely overcome or overpowered by thought or feeling	
Embarrassed	feeling or showing a state of self-conscious confusion and distress	humiliated
Sad	affected with or expressive of grief or unhappiness	disheartened, unhappy
Uncomfortable	causing discomfort or annoyance	awkward
Helpless	marked by an inability to act or react	desperate
Fearful	full of fear, where fear is anxious concern	frightened, afraid, dread

Table C2.2. Codebook used in secondary coding of reasons for emotions as detailed by instructors.

The experienced reasons correspond to causes that are known to have occurred. The anticipated reasons correspond to causes of emotions that are speculated to happen by instructors.

Experienced Reasons	
Code	**Definition:** ***An instructor has an emotional response…***
Instructor's actions	as a result of their own actions or inactions, such as words spoken to students or lack of preparation
Student's actions	as a result of a student's actions or inactions
Student Academic Challenges	because a student is struggling academically (in terms of grades or understanding) in their course
Feedback	to feedback on their pedagogical practices
Instructor effect on students	in regards to their effect on students, such as hindering their students learning, causing the students to feel poorly about themselves, or causing positive feelings towards the course
Contextual/policy requirements	mediated by the department or course having certain requirements that must be met by the instructors in classes (e.g. disability accommodations) or in order for the instructors to take punitive actions (e.g. cheating)
Conflicting student performance/behavior	because student performance is not aligned with their behavior; OR student performance on some assignments is not aligned with other assignments
Opposing needs of students	to students having opposing needs, resulting in some being harmed while others were being helped
Instructor's plan fails	in regards to their pedagogical actions taken or currently being taken failing or not going as planned
Unknown/unclear solution or path	due to not knowing what actions to take or what to say in a situation
Lack of student prior knowledge	due to student's lack prior knowledge relating to past courses or basic information
Instructor unanticipated adjustment	because they have to adjust their teaching when they were expecting not to.
Instructor's inability to answer or teach	because they are unable to answer a question or teach a concept. They may attempt to do so, but do so poorly

Commented [MS1]: This is not clear

Instructor ability	related to viewing themselves able to handle a situation and/or support their students OR unable to do more/an adequate job in responding to the situation or to support their students
Student's circumstances or context	related to their knowledge or understanding of the student's circumstances or personal context
TA's actions	to the TA's actions or inactions
Prior unawareness of academic needs	due to not knowing that their students needed more academic help
Student's opinion	due to how they believe their students feel about them.
Overarching outcome	due to how they view the totality of the situation being described
Anticipated Reasons	
Code	**Definition:** ***An instructor has an emotional response…***
Consequences for instructor	to possible consequences as a result of the situation/experience
Feedback	to possible feedback on their pedagogical practices
Instructor effect on students	in regards to their potential effect on students, such as hindering their students learning, causing the students to feel poorly about themselves, or causing positive feelings towards the course
Opposing needs of students	to the possibility that students may have opposing needs which can result in an intervention harming others while helping some
Student's action	as a result of anticipated student actions or inactions
Peer opinion	due to how their peers may perceive them, their actions, or their ability
Instructor ability	related to anticipation of either handling a situation well or not handling a situation well
Student Academic Challenges	because they anticipate that a student is potentially struggling in their course, but this has not been confirmed by the instructor
Confrontation	due to an anticipation that an interaction with their students will become confrontational or be awkward or uncomfortable
Contextual/policy requirements	due to potentially not meeting requirements for the instructors in classes (e.g. disability accommodations) or for the instructors to take punitive actions (e.g. cheating)
Student's circumstances or context	related to their student's potential circumstances or personal context
Student effect on TA	due to the potential impact that a student can have on a TA
Instructor's plan fails	in regards to their pedagogical actions taken or currently being taken potentially failing or not going as planned

Student opinion	due to how they anticipate their students will feel about them

SEANCE Analysis:

Outputs of SEANCE include metrics from the General Inquirer (GI) dictionary lists, the Lasswell dictionary lists, ScenticNet, EmoLex, the Geneva Affect Label Coder (GALC), Affective Norms for English Words (ANEW), the two Hu-Liu polarity lists, Valence Aware Dictionary for sEntiment Reasoning (VADER), and SEANCE component scores (Scott A Crossley, Kristopher Kyle, & Danielle S McNamara, 2017). Of these, the GI, Lasswell, ScenticNet, and SEANCE component scores were not utilized in this work. Additionally, we utilized the negation option in SEANCE, where a word is not counted towards analysis if a negation word (e.g., 'not') is used in the prior three words. This setting was considered for EmoLex, GALC, ANEW, and Hu and Liu polarity scores; VADER scores automatically consider such negations present in text.

Hu and Liu Polarity

In 2004, Hu and Liu created two polarity lists for sentiment analysis wherein the positive list includes over 2,000 words and the negative list includes over 4,500 words (Hu & Liu, 2004; Liu, Hu, & Cheng, 2005). The indices provided by SEANCE are provided on the document level, and the decimals represent the prominence of categorized words that are either positive or negative, meaning the two numbers add to 1. Also provided is a ratio between the two metrics, (positive proportion)/(negative proportion) (Scott A Crossley, Kristopher Kyle, & Danielle S McNamara, 2017). To better represent our data, we chose the inverse of this quantity to be (negative proportion)/(positive proportion). We classify strongly negative sentiments as ratios greater than 1.50, negative as those between 1.50 and 1.26, neutral as ratios between 1.25 and 0.75, positive sentiments as between 0.76 and 0.66, and strongly positive sentiments as ratios smaller than 0.66. If the ratio was 1:0 (meaning the response only had a non-zero output for the Hu and Liu negative

score), the results were classified as strongly negative. If the ratio was 0:1 (meaning the response only had a non-zero output for the Hu and Liu positive score), the results were classified as strongly positive.

VADER

VADER (Hutto & Gilbert, 2014) scores are derived from a list of words which were given sentiment ratings. Each of the over 7,500 words classified in the VADER sentiment lexicon is assigned a rating from negative to positive four (extremely negative [-4], very negative [-3], moderately negative [-2], slightly negative [-1], neutral or N/A [0], slightly positive [1], moderately positive [2], very positive [3], extremely positive [4]). The SEANCE outputs for VADER include indices for positive, negative, and neutral sentiments for each document. A larger value indicates a greater prevalence of that sentiment in the text. The individual scores of positive, negative, and neutral categories were not used in this analysis; rather we utilized the VADER compound score that is produced for each document. Notably, while the scores attached to each word in the VADER tool range from negative to positive four, the VADER compound scores, which are utilized herein and describe entire documents, range from negative to positive one (Scott A Crossley, Kristopher Kyle, & Danielle S McNamara, 2017). Following prior literature, we set the neutral threshold to be between -.05 to +.05 (Lazrig & Humpherys, 2021), with larger positive scores representing positive sentiments and larger negative scores representing negative sentiments. However, we further delineated the VADER scores as strongly negative (less than -0.6), negative (-0.6 to -0.06), positive (0.06 to 0.6), and strongly positive (greater than 0.6). Importantly, VADER compound scores are inherently modulated by five rules which pertain to punctuation (namely exclamation points), capitalization (especially all-capitalized emphasis), nearby intensifiers, contrastive

conjunctions (e.g., but, however), and negation that occurs within three words of the classified word (Hutto & Gilbert, 2014).

ANEW

ANEW (Bradley & Lang, 1999) is a database of over 1,000 words and their associated values for the valence, arousal, and dominance when each classified word is used. In ANEW, values are given a one to nine scale, where five represents neutrality in regards to one of the three dimensions. Valence is a measure of affect wherein a pleasure dimension is created between happiness and sadness. High numbers represent a closer association to happiness whereas low numbers represent a closer association to sadness. Arousal is indicative of a physical or mental response where large numbers represent agitation, responsiveness, or being excited. Dominance is a dimension which measure the degree of control that is felt where larger numbers represent being in control of and having a handle on a situation (Bradley & Lang, 1999). A single score for each of the three domains is provided for each document; these indices are also on the same 1-9 scale as used in the tool creation (Scott A Crossley, Kristopher Kyle, & Danielle S McNamara, 2017). Our interpretation of the ANEW scores is a modification of prior work in the literature. Delatorre et al. classified ANEW outputs into negative (scores between 1 and 4), neutral (scores between 4 and 6), and positive (scores between 6 and 9) (Delatorre et al., 2019). We first altered the 'negative' and 'positive' denominations to 'low' and 'high' to signify where on the dominance, arousal, and valence scale the response s lied without another sentiment being attached. We then expanded upon the score delineation by narrowing the 'neutral' range and distinguishing between strong scores and near-neutral scores of the three dimensions. Thus, we classified the ANEW scores as follows: Low (1-2.99), Low-mid (3.00-4.49), Neutral (4.5-5.5), High-mid (5.51-7.00), or High (7.01-9).

EmoLex

EmoLex (Mohammad & Turney, 2010; Mohammad & Turney, 2013) classifies emotions into ten categories: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, positive, and negative. Notably, lists of words are categorized based upon the emotion they are associated with as well as whether they are generally associated with either negative or positive emotional experiences. Each category includes at least 500 different words, with the largest list containing over 3,000 (Mohammad & Turney, 2010; Mohammad & Turney, 2013). The scores provided are decimals where a larger value indicates a larger prevalence of the emotion or sentiment, and zero indicates the absence of the emotion in the text. An index score is provided for each of the ten EmoLex categorizations per document (Scott A Crossley, Kristopher Kyle, & Danielle S McNamara, 2017). The EmoLex outputs are primarily utilized in determining whether the eight emotions and two sentiments characterized are present or absent from the analyzed text (Arias-Cabarcos, Khalili, & Strufe, 2022; Păvăloaia et al., 2019).

GALC

As with the tool EmoLex, GALC (Scherer, 2005) establishes categories based upon emotions and the general positive and negative emotive states. However, GALC has 36 emotion categories and two additional classes (Table S3, (Scherer, 2005)).

Table C2.3. GALC categories

Admiration	Amusement	Anger	Anxiety	Being Touched
Boredom	Compassion	Contempt	Contentment	Desperation
Disappointment	Disgust	Dissatisfaction	Envy	Fear
Feeling Loved	Gratitude	Guilt	Happiness	Hatred
Hope	Humility	Interest/Enthusiasm	Irritation	Jealousy
Joy	Longing	Lust	Pleasure	Pride
Relaxation	Relief	Sadness	Shame	Surprise
Tension	Positive	Negative		

The outputs themselves are decimals where the larger value indicates a greater prevalence of the emotion, and zero represents the absence of that emotion as classified under GALC. As with EmoLex, the GALC scores are used to determine either the presence or absence of the emotion categories. Notably, due to GALC's relatively small lexicon as compared to EmoLex, 0 is a more common value in the index scores. Each document is assigned an index score for each of GALC's 38 categories (Scott A Crossley, Kristopher Kyle, & Danielle S McNamara, 2017).

Main unused components of SEANCE:

The GI pulls from the Harvard IV-4 dictionary lists and was originally created by Stone and coworkers in 1966 (Stone, Dunphy, & Smith, 1966). GI has a wide scope, including upwards of 11,000 words which span 119 dictionary lists. These lists are additionally categorized into 17 different classes (Scott A Crossley, Kristopher Kyle, & Danielle S McNamara, 2017). However, due to the relative difficulty of accessing the original dictionary lists and their classifications, the GI was not utilized for sentiment analysis in this project.

The Lasswell dictionary lists were originally created by Lasswell and Namenwirth (Lasswell & Namenwirth, 1969) and were later expanded upon by Namenwirth and Weber (Namenwirth & Weber, 2016). Overall, the Lasswell analysis utilizes 63 different word lists which are categorized into nine different classes (power, rectitude, respect, affection, wealth, well-being, enlightenment, and skill). (Scott A Crossley, Kristopher Kyle, & Danielle S McNamara, 2017). Notably, Lasswell was also excluded from utilization in this project due to the difficulties in accessing the original source lists and their categorizations.

ScenticNet is a newer tool, created by Cambria in 2010 (Cambria, Havasi, & Hussain, 2012; Cambria et al., 2010). However, the ScenticNet outputs provided by SEANCE are not

wholly transparent. The associated scores are produced from semi-supervised algorithms; thus, the outputs incorporate machine learning and the associated black box introduced by such a method. This is not a comment on the reliability or utility of ScenticNet outputs; however, as this is an exploratory study, the authors elected to only utilize methods which can be clearly explained and understood in detail by experts outside of the machine learning fields. The insight provided by ScenticNet, other semi-supervised programs, and even artificial intelligence engines is a promising area of future study once the exploratory work has been conducted.

SEANCE component scores were the last option not utilized. The component scores are unique to the SEANCE program and provide metrics based upon the outputs for the other models contained within the SEANCE program. While the component scores out-performed sentiment analyses by a common tool (LIWC), the microfeatures outperformed the SEANCE component scores. (Scott A Crossley, Kristopher Kyle, & Danielle S McNamara, 2017). Thus, the authors chose to focus on the microfeatures (EmoLex, GALC, ANEW, Hu and Liu Polarity, and VADER) and their scores as provided by the SEANCE program.

Appendix C3: Additional Results

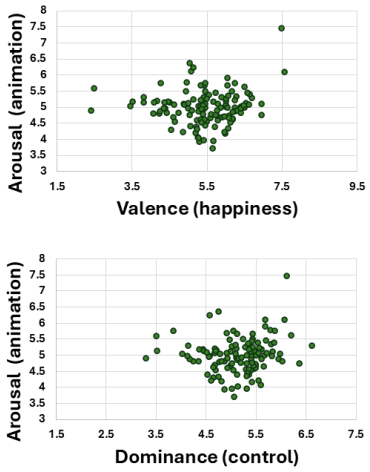

Figure C3.1. Graphs showing lack of correlation between ANEW metrics

Appendix D. Part 2 IRB and Instrument documents

Appendix D1. Instructor documents

Appendix D1.1: Pre-Interview Survey

Pre-Interview Survey

Start of Block: Teaching Experience

How would you describe your current academic position?

- ○ Professor
- ○ Professor of Teaching/Practice
- ○ Associate Professor
- ○ Associate Professor of Teaching/Practice
- ○ Assistant Professor
- ○ Assistant Professor of Teaching/Practice
- ○ Lecturer or Instructor
- ○ Postdoctoral Instructor
- ○ Graduate Student Instructor or Teaching Assistant
- ○ Other _____

What institution are you currently at?

_____

What is your tenure status at your institution?

- ○ Tenured
- ○ On tenure track, but not yet tenured
- ○ Not on tenure track, but my institution has a tenure system
- ○ No tenure system at my institution

Do you have the opportunity for promotion that comes with increased security of employment (e.g., longer contracts)?

- ○ Yes, and I have received such a promotion
- ○ Yes, and I have not received such a promotion
- ○ No opportunity exists

Are you considered a full-time employee of your institution for at least nine (9) months of the current academic year?

- ○ Yes
- ○ No

What is your typical teaching load (i.e., how many course sections do you teach) during an academic year? If you oversee a laboratory course with multiple sections without a designated lecture section, please count that as one for each term.

_____

How long have you been teaching? (in years)

_____

What courses have you taught?

_____

_____

_____

_____

_____

Have you ever participated in any of the following types of teaching-related professional development?

	No	Yes
Half-day workshop(s)	○	○
Full-day or longer workshop(s)	○	○
Attending a teaching-focused conference	○	○
Attended teaching-related presentations at conference not solely dedicated to teaching	○	○
Regular meetings as part of a formal program (e.g., learning community)	○	○
New faculty experience at my institution	○	○
New faculty workshop external to my institution (e.g., Cottrell Scholars Collaborative - CSC NFW, Project NExT, Project ACCESS, Physics New Faculty Workshop)	○	○
Other:	○	○

End of Block: Teaching Experience

Start of Block: Specs & Course Artifacts

Do you use (or have you used) specifications grading in at least one of the courses you teach (or have taught)?

- ○ Yes
- ○ No

How many unique courses do you (or have you) used specifications grading in?

_____

In total, how long have you been using specifications grading? Please include either semesters or quarters in your response (e.g., 2 semesters, 3 quarters).

_____

What is the name of the course that you have used specifications grading in the longest?

_____

On average, what is the typical enrollment for this course?

_____

[Required - Syllabus] Please upload the latest course syllabus from the course that you have used specifications grading the longest.

[Optional - Document 1] Please upload any applicable course documents explaining specifications grading (e.g., assignment specifications, grading scheme, final grade calculations, tokens, revisions) that are not captured by your course syllabus.

[Optional - Document 2] Please upload any applicable course documents explaining specifications grading (e.g., assignment specifications, grading scheme, final grade calculations, tokens, revisions) that are not captured by your course syllabus.

[Optional - Document 3] Please upload any applicable course documents explaining specifications grading (e.g., assignment specifications, grading scheme, final grade calculations, tokens, revisions) that are not captured by your course syllabus.

If you have additional applicable course documents that cannot be uploaded into this form, please email them to Dr. Brandon Yik (byik@virginia.edu) at least 48 hours before your interview.

End of Block: Specs & Course Artifacts

End of Survey

Thank you for filling out this form. Your response has been recorded.

Please follow this link to schedule your approximately 60 minute interview and 2 minutes for a brief post-interview survey: https://calendly.com/vikbrand/specsgradinginterview

Your responses to this form and calendly scheduling will be reviewed shortly. If you requested a virtual interview, a Zoom link will be added within 24 hours to the calendar invite you should have received.

If you have any questions or concerns, please email byik@virginia.edu.

Appendix D1.2: Interview Protocol

Specifications Grading – Interview Protocol – Instructors

Hello [participant's name]. Thank you so much for taking the time to meet with me today.

Pass out study consent form and state that the participant has previously acknowledged consent form. Ask all participants to review the consent form again thoroughly. Then ask for questions, before asking for verbal consent. Let participants know they can keep the consent form.

State that we are going to begin recording. Once the recording has started, ask for verbal confirmation of their consent to be recorded. State the day and time the interview is taking place.

Today we're going to have a conversation about your role as an instructor that uses specifications grading. I would like you to provide honest answers to the questions that I will ask you about.

1. How did you learn about specifications grading?
    a. What resources did you use to learn (more) about specifications grading?
2. Why did you decide to use specifications grading?
    a. What other goals did you have when deciding to use specifications grading in this course?
    b. *Switchers:* What drove you away from traditional grading? Why did you decide to use specifications grading in this course?
    New: Why did you decide to use specifications grading in this course?
    c. Before implementing specifications grading, what benefits did you see in specifications grading?
    d. Before implementing specifications grading, what challenges or worries did you see in specifications grading?

Interviewer space to write down goals and motivations for using specifications grading:

[Section I: For those that *switched to* specifications grading from traditional grading.]

I would like to switch gears to your experiences in designing this course that uses specifications grading.

3. What was your process of switching to using specifications grading?
    a. What were the concrete steps you took when making this switch?
        i. Did you need to change your assessments? How?
        ii. Did you need to change your teaching practices? How?
        iii. Did you need to change your learning outcomes? How?
    b. I see your syllabus has learning outcomes. Can you tell me more about these learning outcomes? Why did you decide to include them in the syllabus?
    c. How did you change your grading scheme?
        i. How did you decide how to bundle your final grades?
4. I see from your course syllabus that you chose to use [X] levels of specifications (e.g., "meets specifications", "not meet specifications", and etc.)? Can you elaborate on this decision?
    a. Some instructors use three or more levels – what do think about this?
    b. Some people use wording such as "meets specifications" and "not yet meets specifications" – what do you think about this?
5. Looking at your syllabus, it looks like you [use/don't use] tokens in your course. Can you tell me more about that?
    a. How did you decide the starting number of tokens? What was the reason for this decision?
    b. Can students earn tokens in your course? Why/why not?
    c. Can students use tokens in your course? Why/why not?

[Section II: For those that *started using* specifications grading with a *new course*.]

I would like to switch gears to your experiences in designing this course that uses specifications grading.

3. What was your process of designing a new course that uses specifications grading?
    a. What were the concrete steps you took when designing this course?
        i. How did you develop your syllabus?
        ii. How did you develop your assessments?
        iii. How did you develop your teaching practices?
        iv. How did you develop your learning outcomes?
    b. I see your syllabus has learning outcomes. Can you tell me more about these learning outcomes? Why did you decide to include them in the syllabus?
    c. How did you develop your grading scheme?
        i. How did you decide how to bundle your final grades?
4. I see from your course syllabus that you chose to use [X] levels of specifications (e.g., "meets specifications", "not meet specifications", and etc.)? Can you elaborate on this decision?

    a. Some instructors use three or more levels – what do think about this?
    b. Some people use wording such as "meets specifications" and "not yet meets specifications" – what do you think about this?
5. Looking at your syllabus, it looks like you [use/don't use] tokens in your course. Can you tell me more about that?
    a. How did you decide the starting number of tokens? What was the reason for this decision?
    b. Can students earn tokens in your course? Why/why not?
    c. Can students use tokens in your course? Why/why not?

Next, I would like to talk about your experiences in using specifications grading.

6. How much effort did it take to design the course before the semester?
7. How much effort did it take while teaching the course during the semester?
8. *<u>Switchers:</u> How does this effort compare with your previous effort in traditional grading?*
9. Has the feedback you give to students change compared with using traditional grading? How?
    a. Do students have an opportunity to go about revising and resubmitting their work?
        i. How does this work?
        ii. Do students get feedback in between? What kind of feedback is provided?
        iii. Is there a limit to how many times a student may revise and resubmit their work? Why?
10. Do you have any concerns about academic integrity in your course?
    a. *<u>Switchers:</u>* Have you received more or less requests for points/extra credit, or a bump in a final letter grade from students compared to the traditionally-graded version?
    b. *<u>New:</u>* Have you received more or less requests for points/extra credit, or a bump in a final letter grade from students compared to other traditionally-graded courses you've taught?

I want to circle back to your intentions before using specifications grading and the actual outcomes.

11. You mentioned that your motivations in switching to specifications grading were [Response to #2]. Do you have any evidence of attaining these goals?
12. Do you have any evidence of improved student outcomes (e.g., final grades)?
    a. Why do you think this is?
13. Are there any other things you took into consideration before implementing specifications grading?
    a. An expected outcome of specifications grading is more equitable student outcomes. Is this something you considered?
        i. Were you able to look deeper into a breakdown by gender, race/ethnicity, etc.?

[Section III: For those that *stopped using* specifications grading.]

A. Why did you choose to stop using specifications grading in this course?
B. Would you consider using specifications grading for a different course you teach?
    a. What information would you need to make this decision?
    b. What steps would you take?

[Section IV: For those that *currently use* specifications grading.]

A. Since using specifications grading, what modifications to your course or grading scheme have you made (or intend to make)?
    a. What are your motivations behind these modifications?
B. Do you plan on implementing specifications grading in other courses you teach?

---

14. Can you provide the names other instructors that you know who use specifications grading?

That's all the questions that I have for you. Do you have any questions for me?

Thank you for having this conversation with me today. I really appreciate you helping us obtain a better understanding of assessing with specifications grading.

Appendix D1.3: Post-Interview Survey

Post-Interview Survey

Start of Block: Demographics

The following survey questions will capture data regarding your demographics. We will only use answers to the following questions to help contextualize study findings and to describe the research sample at the aggregate level. No analyses will be conducted based on these characteristics. Your answers to these questions will not be linked to any other information provided in the study.

What is your age?

_____

My gender or gender identity is best described as (select all that apply):

- [ ] I am agender.
- [ ] I am a man.
- [ ] I am nonbinary.
- [ ] I am gender nonconforming.
- [ ] I am genderqueer or genderfluid.
- [ ] I am questioning.
- [ ] I am transgender.
- [ ] I am a woman.
- [ ] My gender or gender identity is best described as: _____
- [ ] I prefer not to disclose my gender or gender identity.

My racial or ethnic background is best described as (select all that apply):

- ☐ **Black, Afro-Caribbean, or African American** (e.g., Jamaican, Nigerian, Haitian, Ethiopian, etc.)
- ☐ **Asian or Asian American** (e.g., Chinese, Japanese, Filipino, Korean, South Asian, Vietnamese, etc.)
- ☐ **Indigenous American or Alaska Native** (e.g., Navajo Nation, Blackfeet Tribe, Inupiat Traditional Gov't., etc.)
- ☐ **Latino/a/e/x or Hispanic American** (e.g., Puerto Rican, Mexican, Cuban, Salvadoran, Colombian, etc.)
- ☐ **Middle Eastern, North African, or Arab American** (e.g., Lebanese, Iranian, Egyptian, Moroccan, Israeli, Palestinian, etc.)
- ☐ **Native Hawai'ian or Pacific Islander** (e.g., Samoan, Guamanian, Chamorro, Tongan, etc.)
- ☐ **Non-Hispanic White or Euro-American** (e.g., German, Irish, English, Italian, Polish, French, etc.)

☐ **My race or ethnicity is best described as:**

_____

☐ **I prefer not to disclose my racial or ethnic background.**

End of Block: Demographics

Appendix D1.4: Instructor Recruitment Email for Specification Grading Project

Dear [Title. Name]:

You have been identified as an instructor that may use specifications grading in your teaching.

Specifications grading has recently gained popularity, but as of yet, its implementation is under studied. As such, we are team of education researchers and educational developers that aim to characterize instructors' motivations and experiences using specifications grading.

What you will do in the study:

(1) Take a **10-minute online survey** in which you will answer questions about your teaching experience and use of specifications grading. You will also provide course artifacts from your specifications-graded course(s). Course artifacts include any documents related to explaining components of specifications grading, and may include but are not limited to course syllabi, course grading descriptions, specifications for assignments, etc.
(2) Participate in a **60 minute in-person or virtual interview**. Interviews will be audio-recorded. The interview will how you came about specifications grading, goals behind and experiences in using specifications grading, design of your specifications-graded course, and outcomes of using specifications grading. You will be asked to take a brief 2-minute online survey immediately following the interview.

If you are willing to participate in this study, please use the link below to access the consent form, pre-interview survey, and to schedule your interview. Participation in this study is completely voluntary, and should you choose to participate, will take about 10 minutes of your time to complete the pre-interview survey, and about 60 minutes of your time for the interview and post-interview survey.

There are no direct benefits to you for participating in this research study. The study will provide valuable insight to the motivations, barriers, and experiences of instructors that implement or have implement specifications grading. Findings will inform how to better provide professional development workshops and characterize how instructors implement specifications grading.

Overall, this study may provide valuable insight into the implementation of specifications grading.

Study Title: Instructors' Experiences with Specifications Grading
UVA IRB-SBS #5936

Link to consent form, upload course artifacts, and schedule interview:

If you have any questions or would like to withdraw from the study, please contact either of the following team members:

Marilyne Stains, Ph.D.
Professor; Dept. of Chemistry, University of Virginia
mstains@virginia.edu

Brandon Yik, Ph.D.
Postdoctoral Research Associate; Dept. of Chemistry, University of Virginia
byik@virginia.edu

Sincerely,
Dr. Brandon Yik

Appendix D1.5: Instructor Informed Consent Agreement for Specifications Grading Study

Study Title: Instructors' Experiences with Specifications Grading
Protocol #: 5936

Please read this consent agreement carefully before you decide to participate in the study.

Purpose of the research study: The purpose of the study is to characterize and understand instructors' motivations, goals, values, and experiences using specifications grading.

What you will do in the study:

(3) Take a 10-minute online survey in which you will answer questions about your teaching experience and use of specifications grading. You will also provide course artifacts from your specifications-graded course(s). Course artifacts include any documents related to explaining components of specifications grading, and may include but are not limited to course syllabi, course grading descriptions, specifications for assignments, etc.
(4) Participate in a 60 minute in-person or virtual interview. Interviews will be audio-recorded. The interview will how you came about specifications grading, goals behind and experiences in using specifications grading, design of your specifications-graded course, and outcomes of using specifications grading. You will be asked to take a brief 2-minute online survey immediately following the interview. This survey will be sent to your email.

Time required: The study will require about 10 minutes of your time to complete the pre-interview survey, and about 60 minutes of your time for the interview and post-interview survey.

Risks: You may feel slight discomfort associated with some survey and/or interview questions. You can refuse to answer questions you do not wish to answer.

Benefits: There are no direct benefits to you for participating in this research study. The study may provide valuable insight to the motivations, barriers, and experiences of instructors that implement or have implement specifications grading. Findings will inform how to better provide professional development programming centered around specifications grading and characterize how instructors implement specifications grading.

Confidentiality: The information that you give in the study will be handled confidentially. Your information will be assigned a unique code. The list connecting your name to this code will be kept in a locked file. When the study is completed and the data have been analyzed, this list will be destroyed along with your audio recording and transcription. Your name or any identifying information will not be used in any report. All raw data, including interview audio recordings and transcripts, will be retained in a secure manner for 10 years and then destroyed.

Voluntary participation: Your participation in the study is completely voluntary. If you are a University of Virginia employee, your decision to participate will have no effect on employment or university services.

Right to withdraw from the study: You have the right to withdraw from the study at any time without penalty. Withdrawing will not affect your experience as a university employee. Your interview audio recording will be destroyed should you decide to withdraw.

How to withdraw from the study: If you want to withdraw from the study, please tell the interviewer to stop the interview. After the interview, you may contact Dr. Marilyne Stains (mstains@virginia.edu) or Dr. Brandon Yik (byik@virginia.edu) to withdraw from the study. Once a request is received, all of your data will be destroyed.

Payment: You will receive no payment for participating in the study.

Using data beyond this study: Data will be de-identified before both analysis and sharing within the research team. Data may be used for different research endeavors such as case study. The identifiable data you provide in this study will be stored in UVA Box. All files used for analysis (with unique code rather than personal identifiers) will be stored on a cloud system used by the research team. All raw data will be retained in a secure manner for 10 years and then destroyed.

If you have questions about the study, contact:
Marilyne Stains, Ph.D.
Department of Chemistry, University of Virginia
409 McCormick Rd, Charlottesville, VA 22903
Telephone: (434) 243-6430
Email: mstains@virginia.edu

To obtain more information about the study, ask questions about the research procedures, express concerns about your participation, or report illness, injury or other problems, please contact:
Tonya R. Moon, Ph.D.
Chair, Institutional Review Board for the Social and Behavioral Sciences
One Morton Dr Suite 400
University of Virginia, P.O. Box 800392
Charlottesville, VA 22908-0392
Telephone: (434) 924-5999
Email: irbsbshelp@virginia.edu
Website: https://research.virginia.edu/irb-sbs
Website for Research Participants: https://research.virginia.edu/research-participants

UVA IRB-SBS #5936

Electronic Signature Agreement:
I agree to provide an electronic signature to document my consent.

○ I agree to provide an electronic signature to document my consent.

○ I DO NOT agree to provide an electronic signature to document my consent.

Participants will provide electronic signature in Qualtrics.

Study Agreement:
I agree to participate in the research study described above.

○ I agree to participate in the research study described above.

○ I DO NOT agree to participate in the research study described above.

You may print a copy of this consent for your records.

Appendix D2. Student documents

Appendix D2.1. Focus Group Recruitment

Dear [Title. Name]:

We are a team of education researchers who are studying the impact and student perceptions of specifications grading. Specifications grading has recently gained popularity, but as of yet, its implementation is under studied. As such, we aim to gather data concerning students' experiences with specifications grading, as well as information regarding life events, identities, and extenuating circumstances that may have affects students' experiences with specifications grading.

We are writing to you to solicit your participation in a focus group to review the design of our survey. You will not need to prepare anything in advance of the focus group. If you are willing to participate in this study, please contact Dr. Brandon Yik (information below). Participation is completely voluntary, and should you choose to participate, will take approximately 60 minutes of your personal time. The focus group will be in-person. There are no direct benefits should you participate in this study. You may experience discomfort as a result of reflecting on your experiences, you may leave the focus group at any time.

As a whole, the study will provide valuable insight into the implementation of specifications grading.

Study Title: Understanding Effects of Specifications Grading: Student Experiences and Perceptions
UVA IRB-SBS #5793

To participate, please reply to this email, or email Dr. Brandon Yik stating your interest in participating in a focus group interview.

If you have any questions or would like to withdraw from the study after completing the focus group interview, please contact any of the following team members:

Dr. Marilyne Stains
Professor; Dept. of Chemistry
University of Virginia
mstains@virginia.edu

Dr. Brandon Yik
Postdoctoral Research Associate; Dept. of Chemistry
University of Virginia
byik@virginia.edu

Haleigh Machost
PhD Candidate; Dept. of Chemistry
University of Virginia
hrm6cw@virginia.edu

Thank you for your time!
Sincerely,
Dr. Brandon Yik

Appendix D2.2. Focus group consent

Please read this consent agreement carefully before you decide to participate in this focus group.

Purpose of the research study: Specifications grading is an alternative to the traditional, points-based grading system, focuses on student mastery of course objectives, growth-mindset processes, and transparency of expectations. Specifications grading has recently gained popularity, but as of yet, its implementation is under studied. Our goal is to evaluate the implementation of specifications grading and its impact on students and student outcomes.

What will you do in the study: You will participate in a focus group with 1 to 2 other teaching assistants to provide valuable feedback on our survey. The survey itself was designed to probe the effects of specifications grading on students as well as student factors which may affect the potential benefits of specifications grading. In the focus group, you will be provided with a copy of the survey to review and provide feedback. You will not need to prepare anything in advance of the focus group. The focus group session will be audio recorded, and notes will be taken by two members of the research team as the focus group is held.

Time Required: This focus group will require about 60 minutes of your personal time.

Risks: You may feel uncomfortable answering questions as part of a focus group. You may leave the group at any time.

Benefits: There are no benefits to your participation in this study. As a whole, the study will help us understand the impact of specifications grading on students and student outcomes. Study findings will have practical implications for the continuation, improvement, and implementation of specifications grading in classroom environments.

Confidentiality: The information that you give in the study will be handled confidentially. Your name and other information that could be used to identify you will not be collected or linked to the data. Because of the nature of the data, it may be possible to deduce your identity; however, there will be no attempt to do so and your data will be reported in a way that will not identify you.

Voluntary Participation: Your participation in this focus group is completely voluntary. Your decision to participate will have no effect on grades or school services.

Right to withdraw from the study: You have the right to withdraw from the study at any time without penalty or consequence.

How to withdraw from the study: You can decline to participate at any point before the focus group beings. You can leave the room where the focus group is taking place at any time. After the focus group has concluded, you may contact Dr. Marilyne Stains (mstains@virginia.edu), Dr. Brandon Yik (byik@virginia.edu), or Haleigh Machost (hrm6cw@virginia.edu) directly to withdrawal.

There is *no penalty* for withdrawing from this study.

Once a request is received by Dr. Stains, Dr. Yik or Ms. Machost, all of your individual data will be deleted. Due to the nature of audio recording, we may be unable to delete sections of the recording where your voice is recorded. However, your portions will be removed from transcripts created from the audio recording.

Payment: You will receive no payment for participating in the study.

Using data beyond this study: Data will be de-identified before both analysis and before sharing within the research team. Data may be used for different research endeavors such as case study. The identifiable data you provide, if any, in this study will be stored in UVA Box. All files used for analysis will be stored on a cloud system used by the research team. All raw data will be retained in a secure manner for 15 years and then destroyed.

If you have questions about the study, contact:
Dr. Marilyne Stains
Department of Chemistry, University of Virginia
409 McCormick Rd, Charlottesville, VA 22903
Telephone: 434-243-6430
email address: mstains@virginia.edu

To obtain more information about the study, ask questions about the research procedures, express concerns about your participation, or report illness, injury or other problems, please contact:
Tonya R. Moon, Ph.D.
Chair, Institutional Review Board for the Social and Behavioral Sciences
One Morton Dr Suite 500
University of Virginia, P.O. Box 800392
Charlottesville, VA 22908-0392
Telephone: (434) 924-5999
Email: irbsbshelp@virginia.edu
Website: https://research.virginia.edu/irb-sbs
Website for Research Participants: https://research.virginia.edu/research-participants

UVA IRB-SBS #5793

You may keep this copy for your records.

Appendix D2.3. Focus group protocol

The following protocol will be used for each of the three focus groups. Each focus group will include two members of the research team and between 2-4 participants. The participants will be purposefully invited from a pool of teaching assistants.

1. Begin by introducing the two members of the research team as well as the purpose of the overall study as well as the specific purpose of this focus group (i.e., to provide valuable feedback on the developed survey).
2. Pass out study information sheets. Ask all participants to thoroughly review the sheet. Then ask for questions, before asking for asking for verbal consent. Let participants know they can keep the study information sheet.
3. State that we are going to begin recording. Once the recording has started, ask for verbal confirmation of their consent to be recorded.
4. Briefly cover the expectations of this focus group:
    a) Provide honest feedback about your interpretations and opinions of the survey.
    b) Respect everyone's opinions in the room, including not talking over each other.
    c) If you need a break, you can step out at any time for any reason. You are welcome to eat/drink during this focus group session.
5. Ask the participants to read the first section of the survey. Then the research team members will:
    a) Ask about any confusion resulting from the word used in survey items
    b) Ask about any needed clarifications
    c) Ask about the presentation of the Likert scale
    d) Ask for their overall impressions of the survey questions
    e) For questions involving a Likert scale, ask if the scaling of the Likert scale (i.e. 5 versus 7 point options) makes sense for what is being asked
    f) Ask for any additional feedback that may be helpful
6. Repeat Step 5 for all sections of the survey. The first section contains 15 yes/no or numeric responses, and the following sections covers approximately 3-4 questions to be answered with a Likert scale.
7. Thank the survey participants for participating. Answer any questions they may have about the study or the focus group.

Appendix D2.4. Student recruitment

Dear [Title. Name]:

We are a team of education researchers who are studying the impact and student perceptions of specifications grading. Specifications grading has recently gained popularity, but as of yet, its implementation is under studied. As such, we aim to gather data concerning students' experiences with specifications grading, as well as information regarding life events, identities, and extenuating circumstances that may have affects students' experiences with specifications grading.

If you are willing to participate in this study, please use the link below to access the consent form and continue on to take the survey. Participation in this study is completely voluntary, and should you choose to participate, will take approximately 20 minutes.

There are no direct benefits should you participate in this study. You may experience discomfort as a result of reflecting on your experiences, you can skip any questions you do not want to answer. You are also being asked for permission to use course artifacts such as course syllabi, course notes, course materials, course homework, course assignments as implemented as required by the course, and grades to be used for analysis in this study. Survey responses, course artifacts, and course gradebook will be connected for analysis.

As a whole, the study will provide valuable insight into the implementation of specifications grading.

Please note that your instructor, Dr. Morkowchuk, is a member of the study team. Dr. Morkowchuk will not know who is participating in this study until after final grades are submitted.

Study Title: Understanding Effects of Specifications Grading: Student Experiences and Perceptions
UVA IRB-SBS #5793

Link to study:

If you have any questions or would like to withdraw from the study after completing the survey, please contact any of the following team members:

Dr. Marilyne Stains
Professor; Dept. of Chemistry
University of Virginia
mstains@virginia.edu

Dr. Brandon Yik
Postdoctoral Research Associate; Dept. of Chemistry

University of Virginia
byik@virginia.edu

Haleigh Machost
PhD Candidate; Dept. of Chemistry
University of Virginia
hrm6cw@virginia.edu

Thank you for your time!
Sincerely,
Dr. Brandon Yik

Appendix D2.5. Student consent

Study Title: Understanding Effects of Specifications Grading: Student Experiences and Perceptions
UVA IRB-SBS #5793

Please read this consent agreement carefully before you decide to participate in this study.

Purpose of the research study: Specifications grading is an alternative to the traditional, points-based grading system, focuses on student mastery of course objectives, growth-mindset processes, and transparency of expectations. Specifications grading has recently gained popularity, but as of yet, its implementation is under studied. Our goal is to evaluate the implementation of specifications grading and its impact on students and student outcomes.

What will you do in the study: You will answer an online survey wherein you are prompted to answer both questions about your individual circumstances (i.e., your commute time to campus, employment status, etc.) and your perceptions of specifications grading. You should complete the survey when prompted as a part of a course. This survey should *not* be completed during other obligations. You are also being asked for permission to use course artifacts such as course syllabi, course notes, course materials, course homework, course assignments as implemented as required by the course, and grades to be used for analysis in this study.

Time Required: This study will require about 20 minutes of your time.

Risks: Participants may feel slight discomfort associated with recalling information to answer the survey questions asked. You can skip any question you don't want to answer.

Benefits: There are no benefits to your participation in this study. As a whole, the study will help us understand the impact of specifications grading on students and student outcomes. Study findings will have practical implications for the continuation, improvement, and implementation of specifications grading in classroom environments.

Confidentiality: Your identifying information will be gathered to aid in analysis and will be collected on the last page of the questionnaire. Your name will be replaced with a unique random number. All analysis will take place using the de-identified unique random numbers and the data containing your true identity will be deleted. The course instructor, Dr. Morkowchuk, is a member of the research team. Dr. Morkowchuk will not have knowledge if you participate in this study. Dr. Morkowchuk will only have access to data after final grades are submitted and after your name has been de-identified and replaced with a unique random number.

Voluntary Participation: Your participation in this study is completely voluntary. Your decision to participate will have no effect on grades or school services.

Right to withdraw from the study: You have the right to withdraw from the study at any time without penalty.

How to withdraw from the study: You may contact Dr. Marilyne Stains (mstains@virginia.edu), Dr. Brandon Yik (byik@virginia.edu), or Haleigh Machost (hrm6cw@virginia.edu) directly to withdrawal from the study.

There is *no penalty* for withdrawing from this study.

Once a request is received by Dr. Stains, Dr. Yik, or Ms. Machost, all of your data will be deleted.

Payment: You will receive no payment for participating in the study.

Using data beyond this study: Data will be de-identified before both analysis and sharing within the research team. Data may be used for different research endeavors such as case study. The identifiable data you provide in this study will be stored in UVA Box. All files used for analysis (with unique random number rather than personal identifiers) will be stored on a cloud system used by the research team. All raw data will be retained in a secure manner for 15 years and then destroyed.

If you have questions about the study, contact:
Dr. Marilyne Stains
Department of Chemistry, University of Virginia
409 McCormick Rd, Charlottesville, VA 22903
Telephone: 434-243-6430
email address: mstains@virginia.edu

To obtain more information about the study, ask questions about the research procedures, express concerns about your participation, or report illness, injury or other problems, please contact:
Tonya R. Moon, Ph.D.
Chair, Institutional Review Board for the Social and Behavioral Sciences
One Morton Dr Suite 500
University of Virginia, P.O. Box 800392
Charlottesville, VA 22908-0392
Telephone: (434) 924-5999
Email: irbsbshelp@virginia.edu
Website: https://research.virginia.edu/irb-sbs
Website for Research Participants: https://research.virginia.edu/research-participants UVA IRB-SBS # 5793

Electronic Signature Agreement:

I agree to provide an electronic signature to document my consent.

Participants will provide electronic signature in Qualtrics.

Appendix D2.6. PGS instrument survey

What is your age?

_____

Including this current semester, how many semesters (Fall and Spring) have you been in college? Include any prior semesters spent at a community college or university.

_____

How many credits are you enrolled in this semester?

_____

Are you a current recipient of any type of financial aid (e.g., scholarship or grant) that requires you to maintain a certain minimum grade point average (GPA)?

- ○ Yes
- ○ No
- ○ Not Sure

Display This Question:

If Are you a current recipient of any type of financial aid (e.g., scholarship or grant) that requir... = Yes

Of all your financial aid, what is the highest GPA requirement?

_____

Being a first-generation college student means that your parent(s) or guardian(s) did not complete a four-year college or university degree, regardless of other family member's level of education.

Are you a first-generation college student?

- ○ Yes
- ○ No

Are you a UVA student–athlete? This does not include participation in club or intramural sports.

- ○ Yes
- ○ No

Display This Question:

If Are you a UVA student–athlete? This does not include participation in club or intramural sports. = Yes

To what extent has your student–athlete responsibility negatively affected your ability to complete work for this course?

- ○ Not at all
- ○ To a small extent
- ○ To a moderate extent
- ○ To a large extent

How long is your average commute to campus/Grounds? This is a **ONE-WAY TRIP** and not a round trip.

- ○ Less than 15 minutes
- ○ 15 to 29 minutes
- ○ 30 to 44 minutes
- ○ 45 to 59 minutes
- ○ More than 60 minutes

Some people provide care or assistance to a family member (e.g., child, sibling, parent, grandparent, relative, etc.) or friend. Assistance can range from cleaning, shopping, and cooking, to providing medical or personal care. Providing care may include management of a health condition, long-term illness, or disability.

Do you consider yourself to be a caregiver to a family member or friend this semester?

- ○ Yes
- ○ No

Display This Question:

If Some people provide care or assistance to a family member (e.g., child, sibling, parent, grandpar... = Yes

To what extent has your care-giving responsibility negatively affected your ability to complete work for this course?

- ○ Not at all
- ○ To a small extent
- ○ To a moderate extent
- ○ To a large extent

To what extent your physical health negatively affected your ability to complete work for this course?

- ○ Not at all
- ○ To a small extent
- ○ To a moderate extent
- ○ To a large extent

To what extent your mental health negatively affected your ability to complete work for this course?

- ○ Not at all
- ○ To a small extent
- ○ To a moderate extent
- ○ To a large extent

Were you employed (i.e., have a job or internship) this semester? This does not include participation in any sports or extracurricular activities.

- ○ Yes
- ○ No

Display This Question:

If Were you employed (i.e., have a job or internship) this semester? This does not include participa... = Yes

To what extent has your employment negatively affected your ability to complete work for this course?

- ○ Not at all
- ○ To a small extent
- ○ To a moderate extent
- ○ To a large extent

Did you participate in professional organizations, Greek life, clubs, activities, volunteering opportunities, or other extracurriculars this semester?

- ○ Yes
- ○ No

Display This Question:

If Did you participate in professional organizations, Greek life, clubs, activities, volunteering op... = Yes

To what extent has your participation in these activities negatively affected your ability to complete work for this course?

- ○ Not at all
- ○ To a small extent
- ○ To a moderate extent
- ○ To a large extent

The next section asks you about your experiences with different grading schemes. Traditional grading is what is typically used where each assignment is given points or a letter grade. Specifications grading is what is used in this course where each assignment is given either a "Mastered" or "Not Yet Mastered" grade.

Reflect upon your experiences in other science, technology, engineering, and mathematics (STEM) courses that use traditional grading schemes, and your experiences with the specifications grading scheme used in this course. To what degree does each statement represent your experiences with traditional (i.e., other STEM courses) and/or specifications grading (i.e., this course)?

More representative of **traditional grading**	Slightly more representative of **traditional grading**	Equally representative	Slightly more representative of **specifications grading**	More representative of **specifications grading**
○	○	○	○	○

To what degree does each statement represent your experiences with traditional and/or specifications grading?

	More representative of traditional grading		Equally representative		More representative of specifications grading
My grades on assignments represent what I understand about the course topics.	○	○	○	○	○
My grades on assignments capture my understanding of the course material.	○	○	○	○	○
How much I learned is reflected in my grades on assignments.	○	○	○	○	○
I want to only learn what is strictly necessary to earn the grade I want.	○	○	○	○	○
I am motivated to learn as much as possible.	○	○	○	○	○
I am motivated to thoroughly understand the course material.	○	○	○	○	○
I am motivated to get the highest grade on all assignments.	○	○	○	○	○
I am motivated to aim for a higher final grade.	○	○	○	○	○

I am motivated to do my best on assignments.	○	○	○	○	○
I am anxious about my final grade.	○	○	○	○	○
I am anxious about receiving a bad grade on individual assignments.	○	○	○	○	○
I am anxious about making mistakes on assignments.	○	○	○	○	○
Having a different instructor would change my final grade.	○	○	○	○	○
It is up to me to earn the final grade I want.	○	○	○	○	○
My final grade depends on the decisions I make when completing assignments.	○	○	○	○	○
I ask for opportunities to increase my grade.	○	○	○	○	○
I contest grades on assignments.	○	○	○	○	○
I ask for regrades on assignments.	○	○	○	○	○
I pay attention to the feedback I receive on my assignments.	○	○	○	○	○

The feedback I receive on my assignments is helpful to my learning.	○	○	○	○	○
I am able to use the feedback I receive on my assignments to improve my future work.	○	○	○	○	○
I understand what is required to achieve a particular final grade.	○	○	○	○	○
I understand the expectations of the course assignments.	○	○	○	○	○
I understand the expectations for success in this course.	○	○	○	○	○

Appendix E. Supplementary information for chapter 3

Appendix E1. Additional Participant Data

Table E1.1. Number of potential participants identified through each method

Potential participant identification	Potential participants identified
Conference abstracts	26
Journal article publications	23
Snowball sampling	14
Social media posts	10
Online searches	5
Personal communications	4
Book chapters	3
Total	**85**

Table E1.2. Carnegie classifications of participants' institutions.

Classifications include baccalaureate colleges and universities (BAC), master's colleges and universities – larger programs (M1), doctoral/professional universities (D/PU), doctoral universities – high level of research activity (R2), and doctoral universities – very high level of research activity (R1).

	BAC	M1	D/PU	R2	R1	Total
Public	2	1	-	2	6	**11**
Private	7	1	2	2	1	**13**
Total	**9**	**2**	**2**	**4**	**7**	**24**

Table E1.3. Course characteristics of participants.

If instructors taught more than one specifications-graded course, they were asked to describe the course of their first implementation.

	Enrollment	Number of instructors
Lab	Small (<40)	0
	Medium (40-150)	1
	Large (150-1000)	2
	Very large (>1000)	2
	Total	**5**
Lecture	Small (<40)	4
	Medium (40-150)	4
	Large (150-1000)	2
	Very large (>1000)	1
	Total	**11**
Lab-Lecture	Small (<40)	9
	Medium (40-150)	3
	Large (150-1000)	1
	Very large (>1000)	0
	Total	**13**

Appendix E2. Data Analysis

Table E2.1. Summary of changes to the codebook during the interrater reliability process

Iteration	Documents	Summary of changes post-inter-rater analyses
Round 1	178, 179, 180, 184, 186, 188, 189, 192, 195, 197	• Clarified definition of 'reduced student proficiency' • Clarified definition of 'increased instructor workload' • Added to definition of 'unfamiliar system for students' • Verbally clarified that challenges with training TAs falls under 'increased instructor workload' • Changed 'potential negative impact on instructor' to be specific to career affects • Changed 'lack of support' to include damage professional relationships without any affect on career • Added a code 'does not accommodate different groups of students' • Added a code 'maintaining rigor'
Round 2	102, 117, 118, 119, 120, 124, 125, 129, 145, 149	• Clarified the distinction between codes 'unfamiliar system to students' and 'student resistance;' correspondingly clarified the respective definitions • Clarified 'increased tension between instructor and student' to be tension as perceived by the instructor • Removed unused code 'increased student engagement' • Incorporated the code 'multiple opportunities to revise or retake' into the code 'increased flexibility'
Round 3	152, 153, 154, 163, 170, 171, 173, 174, 177	• No further changes
Final Review		• All instances of codes altered, as described above, were revisited to ensure consistency

Table E2.2. Codebook describing perceived advantages of specifications grading that lead to their implementation

Code	Definition: *Instructors want to implement specifications grading…*	Exemplar Quotes
Accommodates different groups of students	in order to increase equitable opportunities for all students as specifics grading can enable students with different cultural backgrounds or contexts to succeed in the course (e.g. differing education backgrounds among students)	"*Students that had a strong organic background could just go through and do everything … the first try and then the students that needed more time, I would be able to give them more retries and feedback. And so, it would help me to differentiate the class a little bit more … and tailor to the different populations we have in there*." -Instructor 152
Increased opportunities for feedback	in order to increase the frequency at which their students receive feedback and/or to ensure that the feedback students receive is meaningful or actionable	*"Specs grading, you sort of get that immediate feedback on this objective or this series of objectives, this one concept. And so, you know, if I'm getting that data out of that exam … you know, out of that, then the students also can get it*." Instructor 186
Increased flexibility	in order to increase the flexibility available to their students (e.g. ability to miss assignments due to illness or sports, ability to revise or retake assignments)	"*I like that it (the token system) gives the students some built-in flexibility in the course. So, then they can request a(n) assignment due-date extension, or, you know, they really just couldn't submit something. Maybe you let them trade in two tokens at the end to make up that assignment, especially if it's a complete/incomplete assignment.*[…]" *Instructor 117*
Increased student agency over learning	in order to increase student self-regulation (e.g., autonomy, metacognition, motivation)	*""(Specs grading) improving their (students') agency over their own learning. I feel like that is a really, um, critical component of student success is having them feel like they have some control over it. And so, um, I think that was what caused me to get started on it" Instructor 102*

Grades reflect students' proficiency	in order to have final grades that are representative of students' proficiency with course material and to enable instructors to accurately track their students' understanding of the material	"*I just started thinking more about what a grade means and something more tangible to say, 'okay, a student who left my class having earned a C, they can do these core things.' And so I could say, yeah, that basically anyone who passed a class C or higher can do these core things and, you know, B students can do some number more, A students can do pretty much everything*." *Instructor 171*
Reduced student stress	in order to decrease their students' anxiety or stress	*"I think it decreased anxiety. Um, I think they were able to compartmentalize these smaller pieces of information and carry them from week to week through some of the exercises that I did as opposed to like, that cram that night before, and take the exam, and then move on and forget everything." Instructor 189*
Increased student learning gains	in order to increase student proficiency with and/or retention of the course material and related skills or to increase the rigor of the course or to increase students' focus on the learning process	"*a mechanism that says 'you're gonna try things. If you don't meet the mark, you get second chances, and that's gonna help reinforce these set of knowledge and skills you need so that you're ready for these more higher stake assessments towards the end'." Instructor 118*
Transparent expectations	because expectations will be clearer to students (e.g. what assignments are necessary, how to complete assignments)	"*I was hoping that it could clarify the expectations for students in terms of what they needed to do to, to earn whatever grade they wanted. And also my expectations of them in terms of what learning outcomes I expected them to meet on each kind of assignment in the class." Instructor 125*
De-incentivizes cheating	because it de-incentivizes students from cheating	*"And, um, again, sort of like I had said, kind of trying to come up with ways where maybe we had lower stakes assignments to, to potentially limit any, um, you know, reasons to, to possibly*

		want to cheat or, or, you know, on the longer assignments" Instructor 173
Reduced grading burden	in order to reduce the mental effort of grading on either themselves and/or the TA's	*"so I was doing a lot of grading by hand, so I was really interested in how specifications grading could help me streamline that grading process for, um, the assessments that I do. 'cause they're most mostly like project kind of things and not multiple choice assessments." Instructor 163*
No partial credit	because it does not use partial credit on individual questions or assignments which are not wholly correct or not completed to standard	"*one of the things that was appealing to me with specifications grading is getting rid of that whole idea of partial credit and having to decide, okay, you know, did this person get 9 points out of 10 and things like that*." *Instructor 173*
Alignment between course learning objectives, and assessments	because it enables better alignment of the course learning objectives and assessments	*"It made me take another look at what I was teaching. And then it was like, 'Okay, well here's my big ideas, and this big idea. I only have one module and one assessment. And for other big ideas, I might have more modules and more assessments.' So it also allowed me to reflect on that and adjust, so that if I considered something important enough to be a big idea that I was assessing it and covering it in as much detail as I would … something else. So that allowed me to go back and like look at what I was covering and reflect on how much time and emphasis I was putting on certain things." Instructor 163*
Reduces tension between instructor and student	in order to reduce instructor's perceived tension in the student-instructor relationship (e.g., move away from the instructor being a gatekeeper of the students' desired grade)	*"[specifications grading] also was a way to de-emphasize the grade focus for our pre-med students who are primarily taking this course. And it allowed them to say, 'Hey, you know, …, just focus on doing the things.' Like … we don't want you to be asking about sig figs, … we don't want you to be, you know, those are important, but that's, I don't want you to be arguing those points every single lab." Instructor 180*

Table E2.3. Codebook describing instructors' dissatisfaction with traditional grading

Code	Definition: *Instructors are dissatisfied with…*	Exemplar Quotes
Creates competition	the competition created between students in a traditionally graded course where students compete to earn higher scores	*"And then if some of them need to take it later, is that an unfair advantage or whatever. And it's like the students are pitted against each other, and constantly I'm making decisions about how to be fair, quote unquote." Instructor 117*
Grades do not reflect students' proficiency	the final grades assigned to students not being representative of their knowledge/skills (e.g., arbitrary differences in points for same proficiency)	"*We realized that several students were not actually reaching mastery or proficiency on really important concepts that we knew would be important in future chemistry courses. And our letter grades at the end of the semester were indicating that that was okay, when we knew that it actually wasn't." Instructor 184*
Lack of flexibility	the lack of flexibility (e.g., unable to make up missed assignments, unable to gain credit for learning material after original assessment)	"*then you look at your evaluation system and it's like you're penalizing them for getting things in not on time. You're penalizing them for learning things later instead of earlier and, and you're penalizing them again and again." Instructor 117*
Reduced student learning gains	students passing the course without concrete understanding of the course material or without emphasis on learning (e.g., students not focused on learning, students emphasizing grades over learning)	"*And at the end of the day, like, should there be enough points where the student literally doesn't have to pass any tests and they still get a B, a D in the class? Or a C in the class? That's probably problematic because you don't know if the students have actually learned anything. Um, so yeah. Yeah, I guess I do have a lot of issues with points-based*." Instructor 120
Unable to predict final grades	the inability to predict final grades until the end of term (e.g., need to curve thus grades may change)	*"I always feels like the grading itself is not fair. <laugh>, there's like all this different percentage together and um, I feel the number is not very meaningful. But that doesn't really show what that number means. And like every time a student ask me in the middle of the semester what their letter grade will be, I cannot really give them an answer because they don't*

		have their future exams done." Instructor 149
Increased students' stress	The stress on the students that the traditional grading causes	*"And our method of assessment is, 'okay, here's a test' and then the test should be, 'okay, I'm gonna measure and see if you actually meet these learning outcomes.' … but then that makes the class rather high stakes, right? Everything in the class, their whole grade is based on the test. Which is another issue too, right? Because then you just have all these high-stake(s) things and it's stressful." Instructor 120*
Does not accommodate different groups of students	does not allow for equitable opportunities for all students such as students with different cultural backgrounds or contexts to succeed in the course (e.g., differing education backgrounds among students)	"*I also feel like a lot of the times our traditional grading is rewarding people that can test well." Instructor 189*
Partial credit	having to give partial credit on individual questions or assignments which are not wholly correct	*"Like the spidey sense in me said like, 'no, no, no, this student did not meet learning objectives as defined in the syllabus or presented,' but the student managed to accrue some partial credit here, some partial credit there, and some homework that maybe, or maybe not, it was, uh, not fairly done, and they get a C, and they pass." Instructor 129*
Increased tension between instructor and student	the tension present in student-instructor interactions (e.g., the instructor is a gatekeeper who the student fights against to get a better grade, grade grubbing)	*"A big problem that you have if you bring points into the equation is that I have the points and I can give them to you or not. And that's about me. And like, maybe we can argue about it a little bit, but in the end, like I'm the, I'm the gatekeeper, I'm the dole-er out of points or the withholder of points. Um, and, and that's just not how most things work in the world." Instructor 117*

Table E2.4. Codebook describing perceived challenges of specifications grading

Code	Definition: *Instructors are concerned…*	Exemplar Quotes
Potential negative impact on students	about the potential negative impacts on students during the semester if specifications grading does not work or a decrease in student outcomes under specs grading	*"Now I have to teach "less", quote unquote, um, you know, and maybe students will "learn less", quote unquote, because I'm teaching less." Instructor 192*
Unfamiliar system for students	that the system is unfamiliar to the students and that the students may not be able to understand the system	"*I mean, it's complicated to explain to students who have never seen it before … and probably even students who have seen something similar before*." Instructor 145
Student resistance	that students will actively not buy-in to specs grading and will choose to resist it	"*The biggest drawback for me has been trying to establish student buy-in, just because students are habituated to other types of grading systems, most commonly points-based grading systems. And there seems to be a lot of student resistance to the new system, and they tend to blame them not doing well or them not understanding how the grading system works as well - just to the fact that it's a new system and it doesn't work very well*." Instructor 120
Increased instructor workload	about having an increased workload in terms of time commitment or mental effort (e.g., spending more time creating assignments, more time grading due to retakes, translation into letter grades)	"*So if I'm doing specifications grading, the way that I do it is with flexible deadlines. And so students are taking different quizzes at different times. You know, maintaining the confidential*

		nature of the assessment pieces was, was another concern." Instructor 174
Potential negative impact on instructor	about the potential negative impacts on themselves (e.g. not receiving tenure, negative career impacts)	*"How will it affect my tenure decision? I'm going off rails. So, uh, but thankfully I had a supportive, uh, colleagues and, and the senior colleague of my mentoring committee" Instructor 129*
Lack of support	about the lack of support they may have from their peers/chair/department/institution if they start specs grading	"*We had a little bit of concern, but not too much about how the department head or administrators might like, if there might be push-back from that. But we got, uh, we got buy-in from them relatively early in the process of planning so that it wasn't really too much of a concern*." Instructor 184
Maintaining rigor	about maintaining the rigor of their course when using specs grading	"*I think my biggest concern then, and arguably, I think now even still, is okay, at what point are we really sure after, I don't know, five retakes, right? Did the student just memorize the concept to pass the objective? Can they truly apply it? Right? Do, have they actually learned it? Can they truly apply it down the road? Is a concern.*" Instructor 186

Appendix F. Supplementary information for chapter 4

Table F.1. Detailed levels of specifications

Specification Grading Levels		Lecture (n=20)	Lab (n=11)	Lecture-Lab (n=19)	All Syllabi (n=50)
2-levels		65%	64%	89%	74%
3-levels		5%			2%
4-levels		5%	18%		6%
Mix of Levels	2-level and 3-level	10%	9%	5%	8%
	2-level and 4-level	10%		5%	6%
	3-level and 4-level	5%			2%
Individual assignments are not specs-graded			9%		2%

Table F.2. Marks nomenclature of two-level systems

Passing phrase	Syllabi	Failing phrase	Syllabi
Competency/competent	n=5	Fail	n=2
Mastery/mastery achieved	n=8	Needs revision	n=5
Meets specifications	n=1	No credit	n=2
Meets expectations	n=2	Not competent	n=1
Pass	n=12	Not mastered/mastery not achieved	n=2
Proficiency	n=4	Not meeting specifications	n=1
Satisfactory	n=8	Not specified	n=17
Successful	n=1	Not yet demonstrated competency	n=1
Complete	n=2	Not yet mastered	n=3
Credit	n=1	Try again/retry	n=2
		Unsatisfactory	n=5
		Incomplete	n=4

Table F.3. Marks nomenclature of multi-level systems

	Scheme	Syllabi	Scheme	Syllabi	Scheme	Syllabi
3-level	Good; Acceptable; Unacceptable	n=1	Mastery; Emerging; Not assessable	n=1	1; 0.5; 0	n=1
	High pass Low pass Needs revision	n=3	High pass pass Needs revision	n=1		
	Scheme	**Syllabi**	**Scheme**	**Syllabi**	**Scheme**	**Syllabi**
4-level	Exemplary; Proficient; Satisfactory; Not specified	n=1	Exceeds standards; Meets standards; In development; Incomplete	n=1	Excellent Meets expectations Revision needed Not assessable	n=1
	Very good; Satisfactory; Unsatisfactory; Missing	n=2	2; 1.5; 1; 0	n=1	Excellent; Good; Revision required; Incomplete	n=1
	Scheme	**Syllabi**				
5-level	Excellent/exemplary; Meet expectations; Revision needed; Not assessable; No report/not specified	n=1				

Table F.4. Revision Systems

Revision Method		Lecture (n=20)	Lab (n=11)	Lecture-Lab (n=19)	All Syllabi (n=50)
Revision/ Reattempt in exchange for token	**Given**		18%	11%	8%
	Earned		9%	16%	8%
	Given and Earned	15%	18%		10%
Automatic reattempts are built into the course structure	**Limited**	25%	9%	53%	32%
	No limit	10%		5%	6%
Automatic and Exchange of Token	**Earned Tokens**	25%			10%
	Given and Earned tokens	15%	18%	16%	16%
No Revisions		10%	27%		10%

Table F.5. Approaches to Final Exam

Mode of Final Exam Assessment	Lecture (n=20)	Lab (n=11)	Lecture-Lab (n=19*)	All Syllabi (n=50*)
Specifications-based final exam	25%	55%	32%	34%
Traditional final exam	30%		53%	32%
No final exam	45%	45%	11%	32%

Table F.6. Determination of Final Letter Grade

Final grade determination		Lecture (n=20)	Lab (n=11)	Lecture-Lab (n=19)	All Syllabi (n=50)
Specs as a measure to pass				16%	6%
Mix of points-based and specifications-based	*Specifications map onto points*	15%	9%	42%	24%
	Bundling of specifications and points	50%	45%	42%	46%
Bundling of specifications		35%	36%		22%
Bundling of points			9%		2%

Appendix G. Supplementary information for chapter 5

Appendix G1. Additional participant data

Table G1.1. Participant demographics

	GC2 Lab (*n* = 648)	GC1 Lab (*n* = 1,031)
Gender		
Man	21%	39%
Woman	69%	56%
Non-binary or other gender identity	1%	<1%
Prefer not to answer	1%	<1%
Data not available	9%	4%
Generation status		
First-generation	12%	17%
Continuing-generation	88%	83%
International student status		
International student	1%	3%
Domestic student	90%	93%
Prefer not to answer	<1%	-----
Data not available	8%	4%
Race/ethnicity		
Black, Afro-Caribbean, or African American	5%	7%
East Asian or Asian American	13%	17%
Latino or Hispanic American	2%	3%
Middle Eastern or Arab American	3%	2%
Multiracial	8%	10%
Non-Hispanic, White, or Euro-American	44%	43%
South Asian or Indian American	12%	13%
Indigenous American	-----	<1%
Pacific Islander	<1%	-----
Other	<1%	<1%
Prefer not to answer	4%	1%
Data not available	8%	4%

Gender, international student status, and race/ethnicity information were obtained through a beginning-of-semester survey assignment; participants with no data available either did not complete the assignment or enrolled in the course after the assignment was due. The item used for first-generation status was "Being a first-generation college student means that your parent(s) or guardian(s) did not complete a four-year college or university degree, regardless of other family

member's level of education. Are you a first-generation college student?" This language was adopted from the Center for First-Generation Student Success (firstgen.naspa.org).

Appendix G2. Data Gathering and Analysis

Table G2.1. Pilot version of the Perceptions of Grading Schemes instrument

Label	Item
Out1	My grades on assignments represent what I understand about the course topics.
Out2	My grades on assignments capture my understanding of the course material.
Out3	How much I learned is reflected in my grades on assignments.
Mot1	I am motivated to learn as much course material as possible.
Mot2	I am motivated to thoroughly understand the course material.
Mot3	I am motivated to do my best on each assignment.
Mot4	I want to only learn course material that is strictly necessary to earn the grade I want.
Mot5	I am motivated to get the highest grade on each assignment.
Mot6	I am motivated to aim for a higher final letter grade.
Anx1	I am anxious about my final letter grade.
Anx2	I am anxious about receiving a bad grade on individual assignments.
Anx3	I am anxious about making mistakes on individual assignments.
Res1	Having a different instructor grading my assignments would change my final letter grade.
Res2	It is up to me to earn the final letter grade I want.
Res3	My final letter grade depends on the decisions I make when completing assignments.
Con1	I ask for extra opportunities to increase my grade.
Con2	I contest grades on assignments.
Con3	I ask for regrades on assignments.
Fbk1	I pay attention to the written feedback I receive on my assignments.
Fbk2	The written feedback I receive on my assignments is helpful to my learning.
Fbk3	I am able to use the written feedback I receive on my assignments to improve my future work.
Exp1	I understand what is required to achieve a particular final letter grade.
Exp2	I understand the expectations of each course assignment.
Exp3	I understand the expectations for success in the course.

Table G2.2. Descriptive statistics for the pilot version of the Perceptions of Grading Schemes Instrument

Label	Mean	SD	Median	Skewness	Kurtosis
Out1	0.01	0.54	0.0	–0.01	–0.61
Out2	0.03	0.58	0.0	–0.01	–0.78
Out3	–0.02	0.58	0.0	0.07	–0.78
Mot1	–0.11	0.51	0.0	0.13	–0.03
Mot2	–0.07	0.54	0.0	0.11	–0.31
Mot3	–0.11	0.49	0.0	0.08	0.15
Mot4	0.09	0.53	0.0	–0.14	–0.41
Mot5	–0.22	0.58	0.0	0.27	–0.49
Mot6	0.00	0.56	0.0	0.02	–0.26
Anx1	–0.39	0.58	–0.5	0.64	–0.37
Anx2	–0.36	0.63	–0.5	0.62	–0.67
Anx3	–0.34	0.66	–0.5	0.69	–0.67
Res1	–0.39	0.55	–0.5	0.56	–0.33
Res2	0.21	0.51	0.0	–0.11	–0.14
Res3	0.10	0.54	0.0	–0.05	–0.27
Con1	–0.32	0.53	–0.5	0.47	–0.23
Con2	–0.11	0.49	0.0	0.02	–0.06
Con3	0.23	0.54	0.0	–0.27	–0.48
Fbk1	–0.09	0.49	0.0	0.07	0.15
Fbk2	0.00	0.48	0.0	0.07	0.11
Fbk3	0.06	0.52	0.0	–0.03	–0.29
Exp1	0.43	0.55	0.5	–0.66	–0.23
Exp2	0.32	0.50	0.5	–0.22	–0.36
Exp3	0.35	0.51	0.5	–0.32	–0.31

For the half of the Spring 2023 GC2 Lab data used for the EFA, there was no missing data from the 324 participants. The average response on the items ranged from –0.43 to +0.42 on a 5-point standardized scale from –1 to +1 with standard deviations ranging from 0.47 to 0.67. In terms of normality, the largest absolute skewness is 0.75, and the largest absolute kurtosis is 0.80; all values for skewness and kurtosis are well below the typical threshold of 2, and therefore data was not further manipulated before factor analyses.

Table G2.3. Item correlations for the pilot version of the Perceptions of Grading Schemes Instrument

Item	1	2	3	4	5	6	7	8	9	10	11
1. Out1											
2. Out2	.654										
3. Out3	.645	.634									
4. Mot1	.470	.430	.412								
5. Mot2	.474	.469	.457	.651							
6. Mot3	.334	.333	.339	.454	.510						
7. Mot4	–.251	–.176	–.240	–.247	–.263	–.224					
8. Mot5	.313	.286	.290	.407	.389	.431	–.041				
9. Mot6	.325	.365	.332	.334	.306	.264	–.097	.365			
10. AnX1	–.208	–.220	–.239	–.032	–.115	–.008	.175	.033	–.069		
11. AnX2	–.269	–.271	–.244	–.141	–.131	–.065	.253	–.012	–.091	.637	
12. Anx3	–.191	–.226	–.229	–.109	–.122	–.019	.270	.028	–.046	.638	.662
13. Res1	–.073	–.087	–.118	–.018	–.034	–.010	.123	.013	.012	.326	.382
14. Res2	.376	.407	.371	.294	.319	.234	–.070	.232	.390	–.265	–.261
15. Res3	.159	.173	.167	.170	.132	.155	.091	.141	.169	–.002	.006
16. Con1	.007	.001	–.023	.075	.027	.042	.053	.097	.089	.234	.252
17. Con2	–.015	–.018	.015	.036	.034	.047	.155	.121	.038	.216	.184
18. Con3	.029	.021	.044	–.017	.032	.041	.060	.033	.047	.077	.125
19. Fbk1	.286	.281	.249	.322	.381	.345	–.180	.274	.241	.044	–.018
20. Fbk2	.322	.326	.338	.366	.427	.308	–.217	.267	.302	–.116	–.139
21. Fbk3	.302	.308	.303	.279	.367	.283	–.163	.243	.244	–.103	–.133
22. Exp1	.315	.332	.311	.197	.262	.161	–.032	.164	.301	–.316	–.301
23. Exp2	.333	.356	.354	.236	.273	.221	–.103	.212	.347	–.349	–.308
24. Exp3	.302	.331	.335	.335	.341	.209	.002	.205	.302	–.325	–.306

Table G2.4 Continued. Item correlations for the pilot version of the Perceptions of Grading Schemes instrument

Item	12	13	14	15	16	17	18	19	20	21	22	23	24
1. Out1													
2. Out2													
3. Out3													
4. Mot1													
5. Mot2													
6. Mot3													
7. Mot4													
8. Mot5													
9. Mot6													
10. Anx1													
11. Anx2													
12. Anx3													
13. Res1	.369												
14. Res2	−.229	−.141											
15. Res3	.018	−.009	.349										
16. Con1	.171	.205	−.024	.076									
17. Con2	.198	.188	−.078	.055	.294								
18. Con3	.095	.147	−.004	.078	.153	.340							
19. Fbk1	.037	.120	.139	.086	.140	.120	.086						
20. Fbk2	−.112	.003	.229	.164	.103	.088	.066	.562					
21. Fbk3	−.086	−.005	.247	.180	.088	.106	.167	.470	.578				
22. Exp1	−.301	−.207	.428	.268	−.028	−.039	.054	.076	.213	.224			
23. Exp2	−.337	−.231	.410	.246	−.080	−.033	.105	.198	.276	.280	.551		
24. Exp3	−.257	−.240	.408	.236	−.092	−.083	.026	.108	.214	.207	.658	.607	

Table G2.5. Final version of the Perceptions of Grading Schemes instrument

Label	Item
Out1	My grades on assignments represent what I understand about the course topics.
Out2	My grades on assignments capture my understanding of the course material.
Out3	How much I learned is reflected in my grades on assignments.
Mot1	I am motivated to learn as much course material as possible.
Mot2	I am motivated to thoroughly understand the course material.
Mot3	I am motivated to do my best on each assignment.
Anx1	I am anxious about my final letter grade.
Anx2	I am anxious about receiving a bad grade on individual assignments.
Anx3	I am anxious about making mistakes on individual assignments.
Fbk1	I pay attention to the written feedback I receive on my assignments.
Fbk2	The written feedback I receive on my assignments is helpful to my learning.
Fbk3	I am able to use the written feedback I receive on my assignments to improve my future work.
Exp1	I understand what is required to achieve a particular final letter grade.
Exp2	I understand the expectations of each course assignment.
Exp3	I understand the expectations for success in the course.

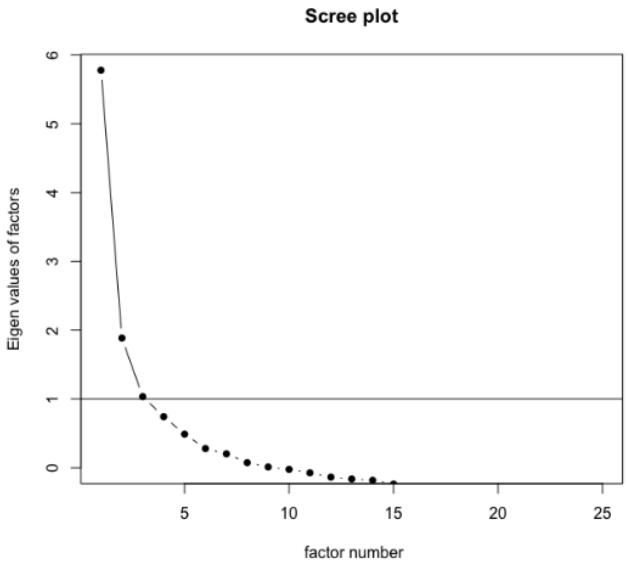

Figure G2.1. Scree plot of the EFA data

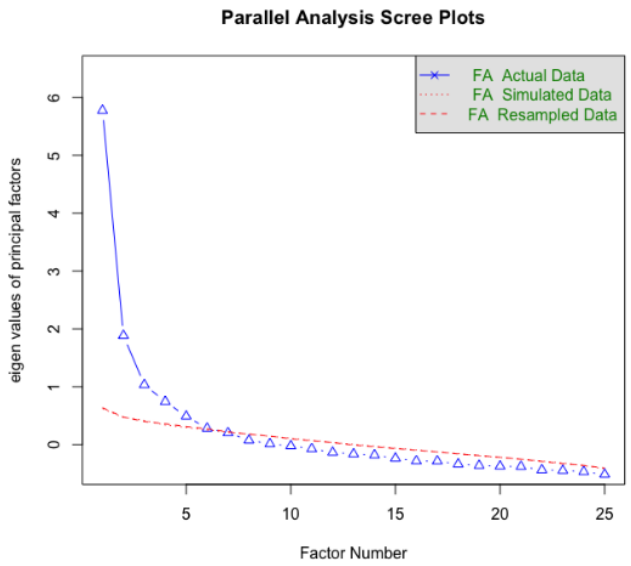

Figure G2.2. Parallel analysis of the EFA data

Parallel analysis suggests that the number of factors = 5 and the number of components = NA

Figure G2.3. CFA model for the Spring 2023 GC2 Lab withs standardized factor loadings. Gray arrows represent non-significant paths ($p > 0.05$).

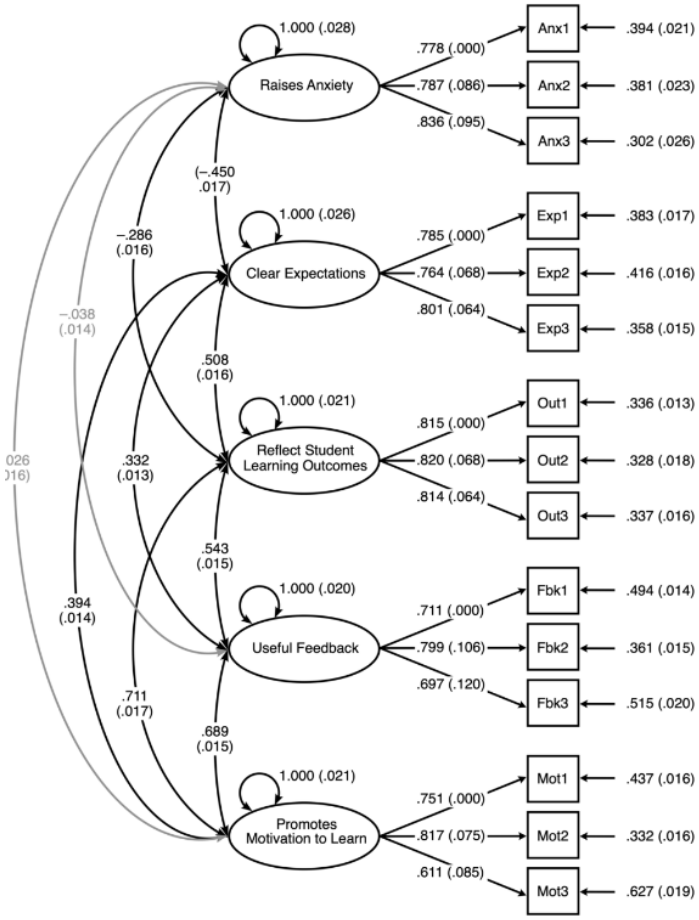

Table G2.6. CFA fit information for individual, single factor congeneric measurement models for the Spring 2023 GC2 Lab

Factor	X^2	df	p	CFI	TLI	RMSEA	SRMR
Raises Anxiety	187.53	3	<0.001	1.000	1.000	0.000	0.000
Clear Expectations	226.67	3	<0.001	1.000	1.000	0.000	0.000
Reflect Student Learning Outcomes	230.58	3	<0.001	1.000	1.000	0.000	0.000
Useful Feedback	165.04	3	<0.001	1.000	1.000	0.000	0.000
Promotes Motivation to Learn	142.89	3	<0.001	1.000	1.000	0.000	0.000

Table G2.7. CFA fit information for individual, single factor tau equivalent measurement models for the Spring 2023 GC2 Lab

Factor	X^2	df	p	CFI	TLI	RMSEA	SRMR
Raises Anxiety	5.102	2	<0.001	0.991	0.987	0.069	0.052
Clear Expectations	5.124	2	0.077	0.991	0.987	0.068	0.053
Reflect Student Learning Outcomes	2.255	2	0.324	0.999	0.999	0.020	0.033
Useful Feedback	4.135	2	0.126	0.990	0.986	0.063	0.048
Promotes Motivation to Learn	13.630	2	0.001	0.950	0.925	0.141	0.096

Table G2.8. CFA fit information for individual, single factor congeneric measurement models for the Fall 2023 GC1 Lab

Factor	X^2	df	p	CFI	TLI	RMSEA	SRMR
Raises Anxiety	582.588	3	<0.001	1.000	1.000	0.000	0.000
Clear Expectations	522.163	3	<0.001	1.000	1.000	0.000	0.000
Reflect Student Learning Outcomes	610.381	3	<0.001	1.000	1.000	0.000	0.000
Useful Feedback	530.506	3	<0.001	1.000	1.000	0.000	0.000
Promotes Motivation to Learn	503.027	3	<0.001	1.000	1.000	0.000	0.000

Table G2.9. CFA fit information for individual, single factor tau equivalent measurement models for the Fall 2023 GC1 Lab

Factor	X^2	df	p	CFI	TLI	RMSEA	SRMR
Raises Anxiety	16.574	2	<0.001	0.983	0.974	0.097	0.064
Clear Expectations	8.882	2	0.012	0.992	0.987	0.066	0.039
Reflect Student Learning Outcomes	3.550	2	0.169	0.999	0.998	0.029	0.024
Useful Feedback	0.798	2	0.671	1.000	1.000	0.000	0.013
Promotes Motivation to Learn	21.760	2	<0.001	0.974	0.961	0.113	0.074

Table G2.10. Student perceptions in the Spring 2023 GC2 Lab (n = 324)

			95% CI		
Factor	**Mean**	**SE**	**LL**	**UL**	**Lean**
Raises Anxiety	–0.34	0.03	–0.40	–0.28	Traditional grading
Clear Expectations	0.36	0.03	0.31	0.41	Specifications grading
Reflect Student Learning Outcomes	–0.01	0.03	–0.06	0.05	Equally representative
Useful Feedback	–0.01	0.02	–0.06	0.03	Equally representative
Promotes Motivation to Learn	–0.11	0.02	–0.16	–0.06	Traditional grading

Table G2.11. Student perceptions in the Fall 2023 GC1 Lab (n = 1,031)

			95% CI		
Factor	**Mean**	**SE**	**LL**	**UL**	**Lean**
Raises Anxiety	–0.26	0.02	–0.29	–0.23	Traditional grading
Clear Expectations	0.18	0.01	0.15	0.21	Specifications grading
Reflect Student Learning Outcomes	–0.01	0.01	–0.04	0.02	Equally representative
Useful Feedback	0.00	0.01	–0.02	0.02	Equally representative
Promotes Motivation to Learn	–0.07	0.01	–0.10	–0.04	Traditional grading

References

Accelerating Systemic Change Network. (2023). *Curated Teaching Evaluation Initiative Repository*. Retrieved 01/03 from https://ascnhighered.org/ASCN/evaluation_initiatives/repository.html

Ahlberg, L. (2021). Organic Chemistry Core Competencies: Helping Students Engage Using Specifications. In *Engaging Students in Organic Chemistry* (Vol. 1378, pp. 25-36). American Chemical Society. https://doi.org/10.1021/bk-2021-1378.ch003

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179-211.

Altakhyneh, B. H., & Abumusa, M. (2020). Attitudes of University Students towards STEM Approach. *International Journal of Technology in Education*, *3*(1), 39-48.

Andrews, T. C., & Lemons, P. P. (2015). It's personal: Biology instructors prioritize personal evidence over empirical evidence in teaching decisions. *CBE—Life Sciences Education*, *14*(1), ar7.

Andrews, T. M., Leonard, M. J., Colgrove, C. A., & Kalinowski, S. T. (2011). Active learning not associated with student learning in a random sample of college biology courses. *CBE—Life Sciences Education*, *10*(4), 394-405.

Ansarin, A. A., Farrokhi, F., & Rahmani, M. (2015). Iranian EFL teachers' reflection levels: The role of gender, experience, and qualifications. *The Asian Journal of Applied Linguistics*, *2*(2), 140-155.

Anzovino, M. E., Behmke, D., Villanueva, O., & Woodbridge, C. M. (2023). Specifications Grading and COVID. In *Chemical Education Research during COVID: Lessons Learned during the Pandemic* (pp. 89-105). ACS Publications.

Ardian, P., Hariyati, R. T. S., & Afifah, E. (2019). Correlation between implementation case reflection discussion based on the Graham Gibbs Cycle and nurses' critical thinking skills. *Enfermeria Clinica*, *29*, 588-593.

Arnaud, C. H. (2021). How an alternative grading system is improving student learning. *Chemical & Engineering News*, *99*(15).

Asai, D. J. (2020). Race matters. *Cell*, *181*(4), 754-757.

Bain, J., Ballantyne, R., Mills, C., & Lester, N. (2002). *Reflecting on practice: Student teachers' perspectives*. Post Pressed.

Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences*. The Guilford Press.

Beach, A. L., Henderson, C., & Finkelstein, N. (2012). Facilitating change in undergraduate STEM education. *Change: The Magazine of Higher Learning*, *44*(6), 52-59.

Beane, R., McNeal, K. S., & Macdonald, R. H. (2019). Probing the National Geoscience Faculty Survey for reported use of practices that support inclusive learning environments in undergraduate courses. *Journal of Geoscience Education*, *67*(4), 427-445.

Beatty, I. D. (2013). Standards-based grading in introductory university physics. *Journal of the Scholarship of Teaching and Learning*, 1-22.

Belvis, E., Pineda-Herrero, P., Armengol, C., & Moreno, V. (2012). Evaluation of reflective practice in teacher education. *European Journal of Teacher Education*, *35*(3), 1-14.

Belvis, E., Pineda, P., Armengol, C., & Moreno, V. (2013). Evaluation of reflective practice in teacher education. *European Journal of Teacher Education*, *36*(3), 279-292.

Bennett, J., Braund, M., & Sharpe, R. M. (2014). Student attitudes, engagement and participation in STEM subjects.

Benoit, A. (2013). Learning from the inside out: A narrative study of college teacher development.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M., & Bonnet, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588-606. https://doi.org/10.1037/0033-2909.88.3.588

Bentley, F., Kennedy, S., & Semsar, K. (2011). How not to lose your students with concept maps.

Betrabet Gulwadi, G. (2009). Using reflective journals in a sustainable design studio. *International Journal of Sustainability in Higher Education*, *10*(1), 43-53.

Black, P. (1998). Inside the black box: Raising standards through classroom assessment. *School of Education, King's College*.

Blackstone, B., & Oldmixon, E. (2019). Specifications Grading in Political Science. *Journal of Political Science Education*, *15*(2), 191-205. https://doi.org/10.1080/15512169.2018.1447948

Blum, S. D. (2020). *Ungrading: Why Rating Students Undermines Learning (and What to Do Instead)*. West Virginia University Press.

Boesdorfer, S. B., Baldwin, E., & Lieberum, K. A. (2018). Emphasizing learning: Using standards-based grading in a large nonmajors' general chemistry survey course. *Journal of Chemical Education*, *95*(8), 1291-1300.

Bowen, R. S., & Cooper, M. M. (2022). Grading on a Curve as a Systemic Issue of Equity in Chemistry Education. *Journal of Chemical Education*, *99*(1), 185-194. https://doi.org/10.1021/acs.jchemed.1c00369

Boyer Commission on Educating Undergraduates in the Research University. (1998). *Reinventing undergraduate education: A blueprint for America's research universities*. State University of New York at Stony Brook for the Carnegie Foundation for ….

Bradforth, S. E., Miller, E. R., Dichtel, W. R., Leibovich, A. K., Feig, A. L., Martin, J. D., Bjorkman, K. S., Schultz, Z. D., & Smith, T. L. (2015). University learning: Improve undergraduate science education. *Nature*, *523*(7560), 282-284. https://doi.org/https://doi.org/10.1038/523282a

Brame, C. (2016). Active learning. *Vanderbilt University Center for Teaching*.

Brems, C., Baldwin, M. R., Davis, L., & Namyniuk, L. (1994). The imposter syndrome as related to teaching evaluations and advising relationships of university faculty members. *The Journal of Higher Education*, *65*(2), 183-193.

Brookfield, S. D. (2017). *Becoming a critically reflective teacher*. John Wiley & Sons.

Brookhart, S. M. (1991). Grading practices and validity. *Educational Measurement: Issues and Practice*, *10*(1), 35-36.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the Most Common Educational Measure. *Review of educational research*, *86*(4), 803-848. https://doi.org/10.3102/0034654316672069

Brooks, A. L., Shekhar, P., Knowles, J., Clement, E., & Brown, S. A. (2024). Contextual Influences on the Adoption of Evidence-Based Instructional Practices by Electrical and Computer Engineering Faculty. *IEEE Transactions on Education*, *67*(3), 351-363.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd ed.). The Guilford Press.

Brownell, S. E., & Tanner, K. D. (2012). Barriers to faculty pedagogical change: Lack of training, time, incentives, and… tensions with professional identity? *CBE—Life Sciences Education*, *11*(4), 339-346.

Bunnell, B., LeBourgeois, L., Doble, J., Gute, B., & Wainman, J. W. (2023). Specifications-Based Grading Facilitates Student–Instructor Interactions in a Flipped-Format General Chemistry II Course. *Journal of Chemical Education*, *100*(11), 4318-4326. https://doi.org/10.1021/acs.jchemed.3c00473

Cain, J., Medina, M., Romanelli, F., & Persky, A. (2022). Deficiencies of traditional grading systems and recommendations for the future. *American Journal of Pharmaceutical Education*, *86*(7).

Campbell, C. M., & Cabrera, A. F. (2014). Making the mark: are grades and deep learning related? *Research in Higher Education*, *55*, 494-507.

Campoy, R. (2010). Reflective Thinking and Educational Solutions: Clarifying What Teacher Educators are Attempting to Accomplish. *Srate Journal*, *19*(2), 15-22.

Canning, E. A., Muenks, K., Green, D. J., & 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*(2), eaau4734.

Carlisle, S. (2020). Simple Specifications Grading. *PRIMUS*, *30*(8-10), 926-951. https://doi.org/10.1080/10511970.2019.1695238

Chamberlin, K., Yasué, M., & Chiang, I.-C. A. (2018). The impact of grades on student motivation. *Active learning in higher education*, 1469787418819728.

Chamberlin, K., Yasué, M., & Chiang, I.-C. A. (2023). The impact of grades on student motivation. *Active learning in higher education*, *24*(2), 109-124. https://doi.org/10.1177/1469787418819728

Chase, A., Pakhira, D., & Stains, M. (2013). Implementing process-oriented, guided-inquiry learning for the first time: Adaptations and short-term impacts on students' attitude and performance. *Journal of Chemical Education*, *90*(4), 409-416.

Chasteen, S. V., Perkins, K. K., Code, W. J., & Wieman, C. E. (2016). The science education initiative: an experiment in scaling up educational improvements in a research university. *Transforming institutions: undergraduate STEM education for the 21st century*, 125-139.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling*, *14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Chen, X. (2013). *STEM Attrition: College Students' Paths into and out of STEM Fields. (NCES 2014-001)*.

Clark, D., & Talbert, R. (2023). *Grading for growth: A guide to alternative grading practices that promote authentic learning and student engagement in higher education*. Taylor & Francis.

Clark, W. (2019). *Academic charisma and the origins of the research university*. University of Chicago Press.

Closser, K. D., Hawker, M. J., & Muchalski, H. (2024). Quantized Grading: An ab Initio Approach to Using Specifications Grading in Physical Chemistry. *Journal of Chemical Education*, *101*(2), 474-482.

Collins, K. H., Price, E. F., Hanson, L., & Neaves, D. (2020). Consequences of stereotype threat and imposter syndrome: The personal journey from stem-practitioner to stem-educator for four women of color. *Taboo: The Journal of Culture and Education*, *19*(4), 10.

Connor, M. C., & Raker, J. R. (2023). Measuring the Association of Departmental Climate around Teaching with Adoption of Evidence-Based Instructional Practices: A National Survey of Chemistry Faculty Members. *Journal of Chemical Education*, *100*(9), 3462-3476.

Corwin, L. A., Graham, M. J., & Dolan, E. L. (2015). Modeling course-based undergraduate research experiences: An agenda for future research and evaluation. *CBE—Life Sciences Education*, *14*(1), es1.

Council, N. R. (2001). *Knowing what students know: The science and design of educational assessment*.

Cracolice, M. S., & Deming, J. C. (2001). Peer-led team learning. *The Science Teacher*, *68*(1), 20-24.

Cromley, J. G., Perez, T., & Kaplan, A. (2016). Undergraduate STEM achievement and retention: Cognitive, motivational, and institutional factors and solutions. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 4-11.

Cromley, J. G., Perez, T., Wills, T. W., Tanaka, J. C., Horvat, E. M., & Agbenyega, E. T.-B. (2013). Changes in race and sex stereotype threat among diverse STEM students: Relation to grades and retention in the majors. *Contemporary Educational Psychology*, *38*(3), 247-258. https://doi.org/10.1016/j.cedpsych.2013.04.003

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, *49*, 803-821.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American journal of physics*, *69*(9), 970-977.

Cubukcu, F. (2013). The Significance of Teachers' Academic Emotions. *Procedia - Social and Behavioral Sciences*, *70*, 649-653. https://doi.org/https://doi.org/10.1016/j.sbspro.2013.01.105

Danielewicz, J., & Elbow, P. (2009). A unilateral grading contract to improve learning and teaching. *College Composition and Communication*, 244-268.

Davila, K., & Talanquer, V. (2010). Classifying end-of-chapter questions and problems for selected general chemistry textbooks used in the United States. *Journal of Chemical Education*, *87*(1), 97-101.

Davis, M. (2003). Barriers to reflective practice: The changing nature of higher education. *Active learning in higher education*, *4*(3), 243-255.

Day, C. (1993). Reflection: A necessary but not sufficient condition for professional development. *British educational research journal*, *19*(1), 83-93.

Del Carlo, D., & Strauss, L. (2023). Standards Based Grading Learning Objective for the Chemistry, Life, the Universe, and Everything General Chemistry Curriculum.

DeMonbrun, M., Finelli, C. J., Prince, M., Borrego, M., Shekhar, P., Henderson, C., & Waters, C. (2017). Creating an instrument to measure student response to instructional practices. *Journal of Engineering Education*, *106*(2), 273-298.

Dennin, M., Schultz, Z. D., Feig, A., Finkelstein, N., Greenhoot, A. F., Hildreth, M., Leibovich, A. K., Martin, J. D., Moldwin, M. B., & O'Dowd, D. K. (2017). Aligning practice to policies: Changing the culture to recognize and reward teaching at research universities. *CBE—Life Sciences Education*, *16*(4), es5.

Dervent, F. (2015). The effect of reflective thinking on the teaching practices of preservice physical education teachers. *Issues in Educational Research*, *25*(3), 260-275.

Dewey, J. (1933). *How we think*. Courier Corporation.

Diegelman-Parente, A. (2011). The use of mastery learning with competency-based grading in an organic chemistry course. *Journal of College Science Teaching*, *40*(5), 50.

Dimitrov, D. M. (2010). Testing for Factorial Invariance in the Context of Construct Validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121-149. https://doi.org/10.1177/0748175610373459

Dinham, J., Choy, S. C., Williams, P., & Yim, J. S. C. (2021). Effective teaching and the role of reflective practices in the Malaysian and Australian education systems: A scoping review. *Asia-Pacific Journal of Teacher Education*, *49*(4), 435-449.

Donaldson, J. H., & Gray, M. (2012). Systematic review of grading practice: is there evidence of grade inflation? *Nurse education in practice*, *12*(2), 101-114.

Donato, J. J., & Marsh, T. C. (2023). Specifications Grading Is an Effective Approach to Teaching Biochemistry. *Journal of Microbiology & Biology Education*, *24*(2), e00236-00222.

Dyment, J. E., & O'connell, T. S. (2010). The quality of reflection in student journals: A review of limiting and enabling factors. *Innovative Higher Education*, *35*, 233-244.

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. Nebraska symposium on motivation,

Ekman, P. (1992). Are there basic emotions? *Psychological review*, *99*(3), 550-553. https://doi.org/10.1037/0033-295X.99.3.550

Elkins, D. M. (2016). Grading to Learn: An Analysis of the Importance and Application of Specifications Grading in a Communication Course. *Kentucky Journal of Communication*, *35*(2), 26-48.

Ellsworth, P. C., & Scherer, K. R. (2002). Appraisal Processes In Emotion. In (pp. 572-595). Oxford University PressNew York, NY. https://doi.org/10.1093/oso/9780195126013.003.0029

Evensen, H. (2022). *Specifications Grading in General Physics and Engineering Physics Courses* 2022 ASEE Annual Conference & Exposition, Minneapolis, MN. https://strategy.asee.org/40676

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford University Pres.

Farrell, T. S. (2003). *Reflective practice in action: 80 reflection breaks for busy teachers*. Corwin Press.

Felder, R. M., & Brent, R. (1996). Navigating the bumpy road to student-centered instruction. *College teaching*, *44*(2), 43-47.

Feldman, A. (2000). Decision making in the practical domain: A model of practical conceptual change. *Science Education*, *84*(5), 606-623.

Feldman, J. (2019a). Beyond standards-based grading: Why equity must be part of grading reform. *Phi Delta Kappan*, *100*(8), 52-55. https://doi.org/10.1177/0031721719846890

Feldman, J. (2019b). What traditional classroom grading gets wrong. *Education Week*, *38*(19), 18-19.

Ferguson, J. H., & Bonner, L. A. (2024). Ungrading in organic chemistry: students assessing themselves and reflecting on their learning. Frontiers in Education,

Fernandez, T. M., Martin, K. M., Mangum, R. T., & Bell-Huff, C. L. (2020). *Whose Grade is it Anyway?: Transitioning Engineering Courses to an Evidence-based Specifications Grading System* 2020 ASEE Annual Conference & Exposition, Online. https://peer.asee.org/35512

Fox, K. R., Campbell, M., & Hargrove, T. (2011). Examining reflective practice: Insights from pre-service teachers, in-service teachers and faculty. *Journal of Research in Education*, *21*(2), 37-54.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences*, *111*(23), 8410-8415.

Frenzel, A. C., Daniels, L., & Burić, I. (2021). Teacher emotions in the classroom and their implications for students. *Educational Psychologist*, *56*(4), 250-264. https://doi.org/10.1080/00461520.2021.1985501

Galea, S. (2012). Reflecting Reflective Practice. *Educational Philosophy and Theory*, *44*(3), 245-258.

Geng, F., & Yu, S. (2024). Teacher Emotions and Instructional Practices: Evidence from a Genre Based L2 Writing Classroom. In (pp. 91-107). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-4484-8_5

Genné-Bacon, E. A., Wilks, J., & Bascom-Slack, C. (2020). Uncovering factors influencing instructors' decision process when considering implementation of a course-based research experience. *CBE—Life Sciences Education*, *19*(2), ar13.

Gervais, J. (2016). The operational definition of competency-based education. *The Journal of Competency-Based Education*, *1*(2), 98-106.

Gibbons, R. E., Reed, J. J., Srinivasan, S., Murphy, K. L., & Raker, J. R. (2022). Assessment Tools in Context: Results from a National Survey of Postsecondary Chemistry Faculty. *Journal of Chemical Education*.

Gibbs, G. (1988). Learning by doing: A guide to teaching and learning methods. *Further Education Unit*.

Gorsuch, R. L. (1983). *Factor Analysis* (2nd ed.). Erlbaum. https://doi.org/10.4324/9780203781098

Granovskiy, B. (2018). Science, Technology, Engineering, and Mathematics (STEM) Education: An Overview. CRS Report R45223, Version 4. Updated. *Congressional Research Service*.

Grant, D., & Green, W. B. (2013). Grades as incentives. *Empirical Economics*, *44*, 1563-1592.

Graves, B. C. (2023). Specifications grading to promote student engagement, motivation and learning: Possibilities and cautions. *Assessing Writing*, *57*, 100754.

Green, C., Brewe, E., Mellen, J., Traxler, A., & Scanlin, S. (2024). Sentiment and thematic analysis of faculty responses: Transition to online learning. *Physical Review Physics Education Research*, *20*(1). https://doi.org/10.1103/physrevphyseducres.20.010151

Gregorich, S. E. (2006). Do Self-Report Instruments Allow Meaningful Comparisons Across Diverse Population Groups?: Testing Measurement Invariance Using the Confirmatory Factor Analysis Framework. *Medical Care*, *44*(11), S78-S94. https://doi.org/10.1097/01.mlr.0000245454.12228.8f

Griffin, M. L. (2003). Using critical incidents to promote and assess reflective thinking in preservice teachers. *Reflective Practice*, *4*(2), 207-220.

Griffiths, M., & Tann, S. (1992). Using reflective practice to link personal and public theories. *Journal of Education for teaching*, *18*(1), 69-84.

Gross, J. J. (1998). The Emerging Field of Emotion Regulation: An Integrative Review. *Review of General Psychology*, *2*(3), 271-299. https://doi.org/10.1037/1089-2680.2.3.271

Guskey, T. R. (2019). Grades versus comments: Research on student feedback. *Phi Delta Kappan*, *101*(3), 42-47.

Guskey, T. R., & Link, L. J. (2019). Exploring the factors teachers consider in determining students' grades. *Assessment in Education: Principles, Policy & Practice*, *26*(3), 303-320.

Hackerson, E. L., Slominski, T., Johnson, N., Buncher, J. B., Ismael, S., Singelmann, L., Leontyev, A., Knopps, A. G., McDarby, A., & Nguyen, J. J. (2024). Alternative grading practices in undergraduate STEM education: a scoping review. *Disciplinary and Interdisciplinary Science Education Research*, *6*(1), 15.

Hagenauer, G., Gläser-Zikuda, M., & Volet, S. E. (2016). University Teachers' Perceptions of Appropriate Emotion Display and High-Quality Teacher-Student Relationship: Similarities and Differences across Cultural-Educational Contexts. *Frontline Learning Research*, *4*(3), 44-74.

Handal, G., & Lauvas, P. (1987). *Promoting reflective teaching: Supervision in practice*. SRHE.

Hansen, M. J., Palakal, M. J., & White, L. J. (2024). The importance of STEM sense of belonging and academic hope in enhancing persistence for low-income, underrepresented STEM students. *Journal for STEM Education Research*, *7*(2), 155-180.

Harrington, B., Galal, A., Nalluri, R., Nasiha, F., & Vadarevu, A. (2024). Specifications and Contract Grading in Computer Science Education. Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1,

Harrison, J. K., & Lee, R. (2011). Exploring the use of critical incident analysis and the professional learning conversation in an initial teacher education programme. *Journal of Education for teaching*, *37*(2), 199-217.

Hatfield, N., Brown, N., & Topaz, C. M. (2022). Do introductory courses disproportionately drive minoritized students out of STEM pathways? *PNAS nexus*, *1*(4), pgac167.

Helmke, B. P. (2019). *Specifications Grading in an Upper-Level BME Elective Course* 2019 ASEE Annual Conference & Exposition, Tampa, FL. https://peer.asee.org/33278

Henderson, C., Beach, A., & Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching*, *48*(8), 952-984.

Henderson, C., Dancy, M., & Niewiadomska-Bugaj, M. (2012). Use of research-based instructional strategies in introductory physics:<? format?> Where do faculty leave the innovation-decision process? *Physical Review Special Topics—Physics Education Research*, *8*(2), 020104.

Henderson, C., & Dancy, M. H. (2007). Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. *Physical Review Special Topics—Physics Education Research*, *3*(2), 020102.

Hensen, C., & Barbera, J. (2019). Assessing Affective Differences between a Virtual General Chemistry Experiment and a Similar Hands-On Experiment. *Journal of Chemical Education*, *96*(10), 2097-2108. https://doi.org/10.1021/acs.jchemed.9b00561

Hernandez, P. R., Schultz, P., Estrada, M., Woodcock, A., & Chance, R. C. (2013). Sustaining optimal motivation: A longitudinal analysis of interventions to broaden participation of underrepresented students in STEM. *Journal of Educational Psychology*, *105*(1), 89.

Herridge, M., & Talanquer, V. (2020). Dimensions of Variation in Chemistry Instructors' Approaches to the Evaluation and Grading of Student Responses. *Journal of Chemical Education*, *98*(2), 270-280. https://doi.org/10.1021/acs.jchemed.0c00944

Herridge, M., Tashiro, J., & Talanquer, V. (2021). Variation in chemistry instructors' evaluations of student written responses and its impact on grading. *Chemistry Education Research and Practice*, *22*(4), 948-972.

Hofmeister, E. H., Fogelberg, K., Conner, B. J., & Gibbons, P. (2023). Specifications Grading in a Cardiovascular Systems Course: Student and Course Coordinator Perspectives on the Impacts on Student Achievement. *Journal of Veterinary Medical Education*, *50*(2), 172-182. https://doi.org/10.3138/jvme-2021-0115

Holme, T., Bretz, S. L., Cooper, M., Lewis, J., Paek, P., Pienta, N., Stacy, A., Stevens, R., & Towns, M. (2010). Enhancing the role of assessment in curriculum reform in chemistry. *Chemistry Education Research and Practice*, *11*(2), 92-97.

Hora, M. T., & Anderson, C. (2012). Perceived norms for interactive teaching and their relationship to instructional decision-making: a mixed methods study. *Higher Education*, *64*(4), 573-592. https://doi.org/10.1007/s10734-012-9513-8

Houseknecht, J. B., & Bates, L. K. (2020). Transition to Remote Instruction Using Hybrid Just-in-Time Teaching, Collaborative Learning, and Specifications Grading for Organic

Chemistry 2. *Journal of Chemical Education*, *97*(9), 3230-3234. https://doi.org/10.1021/acs.jchemed.0c00749

Howitz, W. J., Frey, T., Saluga, S. J., Nguyen, M., Denaro, K., & Edwards, K. D. (2023). A specifications-graded, sports drink-themed general chemistry laboratory course using an argument-driven inquiry approach. *Journal of Chemical Education*, *100*(2), 672-680.

Howitz, W. J., McKnelly, K. J., & Link, R. D. (2021). Developing and Implementing a Specifications Grading System in an Organic Chemistry Laboratory Course. *Journal of Chemical Education*, *98*(2), 385-394. https://doi.org/10.1021/acs.jchemed.0c00450

Howitz, W. J., McKnelly, K. J., & Link, R. D. (2025). Delving into the Design and Implementation of Specifications Grading Systems in Higher Education. *Education Sciences*, *15*(1), Article 83.

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining,

Huda, M., & Teh, K. S. M. (2018). Empowering professional and ethical competence on reflective teaching practice in digital era. In *Mentorship Strategies in Teacher Education* (pp. 136-152). IGI Global.

Hunter, R. A., Pompano, R. R., & Tuchler, M. F. (2022). Alternative Assessment of Active Learning. In *Active Learning in the Analytical Chemistry Curriculum* (Vol. 1409, pp. 269-295). American Chemical Society. https://doi.org/10.1021/bk-2022-1409.ch015

Husebø, S. E., O'Regan, S., & Nestel, D. (2015). Reflective practice and its role in simulation. *Clinical Simulation in Nursing*, *11*(8), 368-375.

Inoue, A. B. (2019). *Labor-based grading contracts: Building equity and inclusion in the compassionate writing classroom*. WAC Clearinghouse Fort Collins, CO.

James, N. M. (2023). Course letter grades and rates of D, W, F grades can introduce variability to course comparisons [10.1039/D2RP00150K]. *Chemistry Education Research and Practice*, *24*(2), 526-534. https://doi.org/10.1039/D2RP00150K

James, N. M., Kreager, B. Z., & LaDue, N. D. (2022). Predict-observe-explain activities preserve introductory geology students' self-efficacy. *Journal of Geoscience Education*, *70*(2), 238-249.

Jay, J. K., & Johnson, K. L. (2002). Capturing complexity: A typology of reflective practice for teacher education. *Teaching and Teacher Education*, *18*(1), 73-85.

Jensen, S. K., & Joy, C. (2005). Exploring a model to evaluate levels of reflection in baccalaureate nursing students' journals. *Journal of Nursing Education*, *44*(3), 139-142.

Jiang, H., Wang, K., Wang, X., Lei, X., & Huang, Z. (2021). Understanding a STEM teacher's emotions and professional identities: a three-year longitudinal case study. *International Journal of STEM Education*, *8*(1). https://doi.org/10.1186/s40594-021-00309-9

Jiang, Z. (2021). The relationships between teacher emotions and classroom instruction: Evidence from senior secondary mathematics teachers in China. *International journal of educational research*, *108*.

Johnson, T., Molinaro, M., & Motika, M. (2018). *Exploring Factors in Course Grade Equity (and what we might do about it).*

Johri, A., Rangwala, H., Lester, J., & Almatrafi, O. (2017). Board# 65: Retention and persistence among stem students: A comparison of direct admit and transfer students across engineering and science. 2017 ASEE Annual Conference & Exposition,

Joseph, M. L., Miller, S. W., Diec, S., & Augustine, J. M. (2023). Successes and challenges in implementing specifications grading in skills-based laboratory courses: Experiences at two colleges of pharmacy. *Currents in Pharmacy Teaching and Learning*, *15*(2), 186-193.

Katzman, S. D., Hurst-Kennedy, J., Barrera, A., Talley, J., Javazon, E., Diaz, M., & Anzovino, M. E. (2021). The Effect of Specifications Grading on Students' Learning and Attitudes in an Undergraduate-Level Cell Biology Course. *Journal of Microbiology & Biology Education*, *22*(3), e00200-00221. https://doi.org/10.1128/jmbe.00200-21

Kelz, J. I., Uribe, J. L., Rasekh, M., Link, R. D., McKnelly, K. J., & Martin, R. W. (2023). Implementation of specifications grading in an upper-division chemical biology course. *Biophysical Journal*, *122*(3), 298a. https://doi.org/10.1016/j.bpj.2022.11.1684

Kiefer, S. F., & Earle, A. J. (2023). *Work in Progress: Specifications Grading in a System Modeling Course* 2023 ASEE Annual Conference & Exposition, Baltimore, MD. https://peer.asee.org/44356

King, B. (2015). Changing College Majors: Does it Happen More in STEM and Do Grades Matter? *Journal of College Science Teaching*, *44*(3), 44-51. http://www.jstor.org/stable/43631938

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Koenka, A. C., Linnenbrink-Garcia, L., Moshontz, H., Atkinson, K. M., Sanchez, C. E., & Cooper, H. (2021). A meta-analysis on the impact of grades and comments on academic motivation and achievement: a case for written feedback. *Educational Psychology*, *41*(7), 922-947. https://doi.org/10.1080/01443410.2019.1659939

Koester, B. P., Grom, G., & McKay, T. A. (2016). Patterns of gendered performance difference in introductory STEM courses. *arXiv preprint arXiv:1608.07565*.

Kohn, A. (2011). The case against grades. *Educational leadership*, *69*(3), 28-33.

Komperda, R., Pentecost, T. C., & Barbera, J. (2018). Moving beyond Alpha: A Primer on Alternative Sources of Single-Administration Reliability Evidence for Quantitative Chemistry Education Research. *Journal of Chemical Education*, *95*(9), 1477-1491. https://doi.org/10.1021/acs.jchemed.8b00220

Kordts-Freudinger, R. (2017). Feel, think, teach – Emotional Underpinnings of Approaches to Teaching in Higher Education. *International Journal of Higher Education*, *6*(1), 217. https://doi.org/10.5430/ijhe.v6n1p217

Kothiyal, A., Majumdar, R., Murthy, S., & Iyer, S. (2013). Effect of think-pair-share in a large CS1 class: 83% sustained engagement. Proceedings of the ninth annual international ACM conference on International computing education research,

Kraft, A. R., Atieh, E. L., Shi, L., & Stains, M. (2024). Prior experiences as students and instructors play a critical role in instructors' decision to adopt evidence-based instructional practices. *International Journal of STEM Education*, *11*(1), 18.

Kryshko, O., Fleischer, J., Grunschel, C., & Leutner, D. (2022). Self-efficacy for motivational regulation and satisfaction with academic studies in STEM undergraduates: The mediating role of study motivation. *Learning and Individual Differences*, *93*, 102096.

Kulik, C.-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of educational research*, *60*(2), 265-299.

Lake, D. A. (2001). Student performance and perceptions of a lecture-based course compared with the same course utilizing group discussion. *Physical therapy*, *81*(3), 896-902.

Lane, A. K., Earl, B., Feola, S., Lewis, J. E., McAlpin, J. D., Mertens, K., Shadle, S. E., Skvoretz, J., Ziker, J. P., & Stains, M. (2022). Context and content of teaching conversations: exploring how to promote sharing of innovative teaching knowledge between science faculty. *International Journal of STEM Education*, *9*(1), 1-16.

Larrivee, B. (2000). Transforming Teaching Practice: Becoming the critically reflective teacher. *Reflective Practice*, *1*(3), 293-306. https://doi.org/DOI: 10.1080/14623940020025561

Larrivee, B. (2005). *Authentic classroom management: Creating a learning community and building reflective practice*. Allyn & Bacon.

Larrivee, B. (2008a). Development of a tool to assess teachers' level of reflective practice. *Reflective Practice*, *9*(3), 341-360. https://doi.org/https://doi.org/10.1080/14623940802207451

Larrivee, B. (2008b). Meeting the challenge of preparing reflective practitioners. *The New Educator*, *4*(2), 87-106.

Laverty, J. T., Underwood, S. M., Matz, R. L., Posey, L. A., Carmel, J. H., Caballero, M. D., Fata-Hartley, C. L., Ebert-May, D., Jardeleza, S. E., & Cooper, M. M. (2016). Characterizing college science assessments: The three-dimensional learning assessment protocol. *PloS one*, *11*(9), e0162333.

Lee, H.-J. (2005). Understanding and assessing preservice teachers' reflective thinking. *Teaching and Teacher Education*, *21*(6), 699-715.

Lee, S. J. C., & Abdul Rabu, S. N. (2022). Google Docs for higher education: Evaluating online interaction and reflective writing using content analysis approach. *Education and Information Technologies*, *27*(3), 3651-3681.

Lewis, D. (2020). Student anxiety in standards-based grading in mathematics courses. *Innovative Higher Education*, *45*(2), 153-164.

Link, L. J., & Guskey, T. R. (2019). How Traditional Grading Contribute to Student Inequities and How to Fix It. *Curriculum in Context*, *45*(1).

Lipnevich, A. A., Guskey, T. R., Murano, D. M., & Smith, J. K. (2020). What do grades mean? Variation in grading criteria in American college and university courses. *Assessment in*

Education: Principles, Policy & Practice, *27*(5), 480-500. https://doi.org/10.1080/0969594X.2020.1799190

Lipnevich, A. A., & Smith, J. K. (2008). RESPONSE TO ASSESSMENT FEEDBACK: THE EFFECTS OF GRADES, PRAISE, AND SOURCE OF INFORMATION. *ETS Research Report Series*, *2008*(1), i-57. https://doi.org/10.1002/j.2333-8504.2008.tb02116.x

Lipnevich, A. A., & Smith, J. K. (2009). "I really need feedback to learn:" students' perspectives on the effectiveness of the differential feedback messages. *Educational Assessment, Evaluation and Accountability*, *21*(4), 347-367. https://doi.org/10.1007/s11092-009-9082-2

Loehlin, J. C., & Beaujean, A. A. (2017). *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis* (5th ed.). Routledge.

Lönngren, J., Bellocchi, A., Berge, M., Bøgelund, P., Direito, I., Huff, J. L., Mohd-Yusof, K., Murzi, H., Farahwahidah Abdul Rahman, N., & Tormey, R. (2024). Emotions in engineering education: A configurative meta-synthesis systematic review. *Journal of Engineering Education*, *113*(4), 1287-1326. https://doi.org/https://doi.org/10.1002/jee.20600

Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American journal of physics*, *74*(2), 118-122. https://doi.org/10.1119/1.2162549

Loughran, J. J. (2002). *Developing reflective practice: Learning about teaching and learning through modelling*. Routledge.

Lovell, M. D. (2018). *Defining and Assessing Competencies in an Undergraduate Reinforced Concrete Design Course* 2018 ASEE Annual Conference & Exposition, Salt Lake City, UT. https://peer.asee.org/30034

Lubbe, W., & Botha, C. S. (2020). The dimensions of reflective practice: a teacher educator's and nurse educator's perspective. *Reflective Practice*, *21*(3), 287-300.

Lund, T. J., & Stains, M. (2015). The importance of context: an exploration of factors influencing the adoption of student-centered teaching among chemistry, biology, and physics faculty. *International Journal of STEM Education*, *2*(1), 1-21.

Machost, H., & Stains, M. (2023). Reflective Practices in Education: A Primer for Practitioners. *CBE—Life Sciences Education*, *22*(2), es2. https://doi.org/10.1187/cbe.22-07-0148

Madsen, A., McKagan, S. B., & Sayre, E. C. (2017). Resource letter Rbai-1: research-based assessment instruments in physics and astronomy. *American journal of physics*, *85*(4), 245-264.

Mahmood, M. A., Tariq, M., & Javed, S. (2011). Strategies for active learning: An alternative to passive learning. *Academic Research International*, *1*(3), 193.

Markkanen, P., Välimäki, M., Anttila, M., & Kuuskorpi, M. (2020). A reflective cycle: Understanding challenging situations in a school setting. *Educational Research*, *62*(1), 46-62.

Marshall, T. (2019). The concept of reflection: a systematic review and thematic synthesis across professional contexts. *Reflective Practice*, *20*(3), 396-415.

Marshman, E. M., Kalender, Z. Y., Nokes-Malach, T., Schunn, C., & Singh, C. (2018). Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm? *Physical Review Physics Education Research*, *14*(2), 020123.

Martin, L. J. (2019). Introducing Components of Specifications Grading to a General Chemistry I Course. In *Enhancing Retention in Introductory Chemistry Courses: Teaching Practices and Assessments* (Vol. 1330, pp. 105-119). American Chemical Society. https://doi.org/10.1021/bk-2019-1330.ch007

Maruf, A. A., Khanam, F., Haque, M. M., Jiyad, Z. M., Mridha, M. F., & Aung, Z. (2024). Challenges and Opportunities of Text-Based Emotion Detection: A Survey. *IEEE Access*, *12*, 18416-18450. https://doi.org/10.1109/ACCESS.2024.3356357

Mattanah, J., Holt, L. J., Feinn, R. S., Katzenberg, C., Albert, E., Boarman, R., Bowley, O., Marszalek, K., Visalli, T., & Daramola, D. (2024). Attachment Representations and Emotions in Teaching as Antecedents to Teaching Styles in Higher Education. *International Journal of Higher Education*, *13*(1), 1-1.

Matz, R. L., Koester, B. P., Fiorini, S., Grom, G., Shepard, L., Stangor, C. G., Weiner, B., & McKay, T. A. (2017). Patterns of gendered performance differences in large introductory courses at five research universities. *Aera Open*, *3*(4), 2332858417743754.

McAlpin, J. D., Ziker, J. P., Skvoretz, J., Couch, B. A., Earl, B., Feola, S., Lane, A. K., Mertens, K., Prevost, L. B., Shadle, S. E., Stains, M., & Lewis, J. E. (2022). Development of the Cooperative Adoption Factors Instrument to measure factors associated with instructional

practice in the context of institutional change. *International Journal of STEM Education*, *9*(1), 48. https://doi.org/10.1186/s40594-022-00364-w

McAlpine, L., & Weston, C. (2002). Reflection: Issues related to improving professors' teaching and students' learning. *Teacher thinking, beliefs and knowledge in higher education*, 59-78.

McAlpine, L., Weston, C., Berthiaume, D., Fairbank-Roch, G., & Owen, M. (2004). Reflection on teaching: Types and goals of reflection. *Educational research and evaluation*, *10*(4-6), 337-363.

McConnell, M., Montplaisir, L., & Offerdahl, E. G. (2020). A model of peer effects on instructor innovation adoption. *International Journal of STEM Education*, *7*(1), 53.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*(1), 100-117. https://doi.org/10.1111/j.2044-8317.1981.tb00621.x

McDonald, R. P. (1999). *Test Theory: A Unifed Treatment*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410601087

McKay, T. (2019). Gendered performance in introductory STEM courses. APS April Meeting Abstracts,

McKnelly, K. J., Howitz, W. J., Thane, T. A., & Link, R. D. (2023). Specifications Grading at Scale: Improved Letter Grades and Grading-Related Interactions in a Course with over 1,000 Students. *Journal of Chemical Education*, *100*(9), 3179-3193. https://doi.org/10.1021/acs.jchemed.2c00740

McKnelly, K. J., Morris, M. A., & Mang, S. A. (2021). Redesigning a "Writing for Chemists" Course Using Specifications Grading. *Journal of Chemical Education*, *98*(4), 1201-1207. https://doi.org/10.1021/acs.jchemed.0c00859

Mendez, J. (2018a). *Standards-Based Specifications Grading in a Hybrid Course* 2018 ASEE Annual Conference & Exposition, Salt Lake City, UT. https://peer.asee.org/30982

Mendez, J. (2018b). *Standards-Based Specifications Grading in Thermodynamics* ASEE IL-IN Section Conference, West Lafayette, IN. https://doi.org/10.5703/1288284316862

Mendzheritskaya, J., & Hansen, M. (2013). Shall I show my anger? Display rules in lecturer-student interaction in Germany and Russia. 15th Biennal EARLI conference. Munich, Germany,

Mendzheritskaya, J., & Hansen, M. (2019). The role of emotions in higher education teaching and learning processes. *Studies in Higher education*, *44*(10), 1709-1711. https://doi.org/10.1080/03075079.2019.1665306

Michael, J. (2007). Faculty perceptions about barriers to active learning. *College teaching*, *55*(2), 42-47.

Miller, E. R., & Fairweather, J. S. (2015). The Role of Cultural Change in Large-Scale STEM Reform: The Experience of the AAU Undergraduate STEM Education Initiative. *Transforming institutions: undergraduate STEM education for the 21st century*, *48*.

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. Routledge.

Minott, M. A. (2008). Valli's Typology Of Reflection And The Analysis Of Pre-Service Teachers' Reflective Journals. *Australian Journal of Teacher Education*, *33*(5), 55-65.

Mirsky, G. M. (2018). Effectiveness of specifications grading in teaching technical writing to computer science students. *Journal of Computing Sciences in Colleges*, *34*(1), 104-110. https://dl.acm.org/doi/abs/10.5555/3280489.3280505

Mohamed, M., Rashid, R. A., & Alqaryouti, M. H. (2022). Conceptualizing the complexity of reflective practice in education. *Frontiers in Psychology*, *13*, 1008234.

Mohammad, S. M., & Turney, P. D. (2013). CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON. *Computational Intelligence*, *29*(3), 436-465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Momsen, J., Offerdahl, E., Kryjevskaia, M., Montplaisir, L., Anderson, E., & Grosz, N. (2013). Using assessments to investigate and compare the nature of learning in undergraduate science courses. *CBE—Life Sciences Education*, *12*(2), 239-249.

Moog, R. S. (2019). Origins of POGIL: Process Oriented Guided Inquiry Learning. In *POGIL* (pp. 22-41). Routledge.

Moog, R. S., & Spencer, J. N. (2008). POGIL: An overview.

Mooring, S. R., Mitchell, C. E., & Burrows, N. L. (2016). Evaluation of a flipped, large-enrollment organic chemistry course on student attitude and achievement. *Journal of Chemical Education*, *93*(12), 1972-1983.

Morkowchuk, L. N. (2024). *Specifications Grading in General Chemistry Laboratory at UVA*.

Morris, M., Hensel, R., & Dygert, J. (2019). Why do students leave? An investigation into why well-supported students leave a first-year engineering program. ASEE annual conference & exposition proceedings,

Moster, C. A., & Zingales, S. K. (2024). Use of specifications-based grading in an online, asynchronous graduate organic chemistry course. Frontiers in Education,

Munby, H., & Russell, T. (1989). Educating the reflective teacher: An essay review of two books by Donald Schon. *Journal of Curriculum Studies*, *21*(1), 71-80.

Mutambuki, J., & Fynewever, H. (2012). Comparing chemistry faculty beliefs about grading with grading practices. *Journal of Chemical Education*, *89*(3), 326-334.

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*(1). https://doi.org/10.1007/s13278-021-00776-6

National Academies of Sciences, E., and Medicine. (2021). *Call to action for science education: Building opportunity for the future*.

Newman, S. (1999). Constructing and critiquing reflective practice. *Educational Action Research*, *7*(1), 145-163.

Newton, R. P. (2023). *Student Perceptions of Democratic Contract Grading, Specifications Grading, and Ungrading in Community College* University of Nebraska at Omaha].

Nilson, L. B. (2015). *Specifications grading: Restoring rigor, motivating students, and saving faculty time*. Stylus Publishing, LLC.

Noell, S. L., Rios Buza, M., Roth, E. B., Young, J. L., & Drummond, M. J. (2023). A bridge to specifications grading in second semester general chemistry. *Journal of Chemical Education*, *100*(6), 2159-2165.

O'Connell, T. S., & Dyment, J. E. (2004). Journals of post secondary outdoor recreation students: the results of a content analysis. *Journal of Adventure Education & Outdoor Learning*, *4*(2), 159-171.

O'Connell, T. S., & Dyment, J. E. (2011). The case of reflective journals: Is the jury still out? *Reflective Practice*, *12*(1), 47-59.

Offerdahl, E. G., Hodgson, A., & Krupke, C. (2016). Lowering the activation barrier, not the academic bar: The role of contract grading in decreasing DFW rates in introductory biochemistry. *The FASEB Journal*, *30*, 662.619-662.619.

Ohland, M. W., Sheppard, S. D., Lichtenstein, G., Eris, O., Chachra, D., & Layton, R. A. (2008). Persistence, engagement, and migration in engineering programs. *Journal of Engineering Education*, *97*(3), 259-278.

Oleson, A., & Hora, M. T. (2014). Teaching the way they were taught? Revisiting the sources of teaching knowledge and the role of prior experience in shaping faculty teaching practices. *Higher Education*, *68*, 29-45.

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

Onyenwe, I., & et al. (2020). The impact of political party/candidate on the election results from a sentiment analysis perspective using #AnambraDecides2017 tweets. *Social Network Analysis and Mining*, *10*(1).

Ost, B. (2010). The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review*, *29*(6), 923-934. https://doi.org/10.1016/j.econedurev.2010.06.011

Osterman, K. F., & Kottkamp, R. B. (2004). *Reflective practice for educators: Professional development to improve student learning*. Corwin Press.

Page, E. B. (1958). Teacher comments and student performance: A seventy-four classroom experiment in school motivation. *Journal of Educational Psychology*, *49*(4), 173.

Park, J. J., Kim, Y. K., Salazar, C., & Hayes, S. (2020). Student–faculty interaction and discrimination from faculty in STEM: The link with retention. *Research in Higher Education*, *61*, 330-356.

Parkman, A. (2016). The imposter phenomenon in higher education: Incidence and impact. *Journal of Higher Education Theory and Practice*, *16*(1), 51.

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., & Booth, R. (2007). The Development and Psychometric Properties of LIWC2007.

Petcovic, H. L., Fynewever, H., Henderson, C., Mutambuki, J. M., & Barney, J. A. (2013). Faculty grading of quantitative problems: A mismatch between values and practice. *Research in Science Education*, *43*(2), 437-455.

Peters, G.-J. Y., & Gruijters, S. (2023). *ufs: A collection of utilities*. In (Version R package version 0.5.10) https://ufs.opens.science/

Peters, M. L. (2013). Examining the relationships among classroom climate, self-efficacy, and achievement in undergraduate mathematics: A multi-level analysis. *International Journal of Science and Mathematics Education*, *11*(2), 459-480.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Sage.

Phillips, D. C. (1995). The good, the bad, and the ugly: The many faces of constructivism. *Educational Researcher*, *24*(7), 5-12.

Physics and Astronomy Faculty Teaching Institute. (2023). *Faculty Teaching Institute, June 2023*. Retrieved 01/04 from https://www.physport.org/fti/

Plack, M. M., Driscoll, M., Blissett, S., McKenna, R., & Plack, T. P. (2005). A method for assessing reflective journal writing. *Journal of allied health*, *34*(4), 199-208.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, *1*.

Pond, J. W., & Chini, J. J. (2017). Exploring student learning profiles in algebra-based studio physics: A person-centered approach. *Physical Review Physics Education Research*, *13*(1), 010119.

Popova, M., & Jones, T. (2021). Chemistry instructors' intentions toward developing, teaching, and assessing student representational competence skills. *Chemistry Education Research and Practice*, *22*(3), 733-748. https://doi.org/10.1039/d0rp00329h

Popova, M., Shi, L., Harshman, J., Kraft, A., & Stains, M. (2020). Untangling a complex relationship: Teaching beliefs and instructional practices of assistant chemistry faculty at research-intensive institutions. *Chemistry Education Research and Practice*, *21*(2), 513-527.

Postareff, L., & Lindblom-Ylänne, S. (2011). Emotions and confidence within teaching in higher education. *Studies in Higher education*, *36*(7), 799-813. https://doi.org/10.1080/03075079.2010.483279

Prescott, S. G. (2015). Will Instructors Save Time Using a Specifications Grading System? *Journal of Microbiology & Biology Education*, *16*(2), 298-298. https://doi.org/10.1128/jmbe.v16i2.1027

Prince, M. J., & Felder, R. M. (2006). Inductive teaching and learning methods: Definitions, comparisons, and research bases. *Journal of Engineering Education*, *95*(2), 123-138.

Prosser, M., & Trigwell, K. (1997). Relations between perceptions of the teaching environment and approaches to teaching. *British Journal of Educational Psychology*, *67*(1), 25-35. https://doi.org/10.1111/j.2044-8279.1997.tb01224.x

Pulfrey, C., Buchs, C., & Butera, F. (2011). Why grades engender performance-avoidance goals: The mediating role of autonomous motivation. *Journal of Educational Psychology*, *103*(3), 683.

Pulfrey, C., Darnon, C., & Butera, F. (2013). Autonomy and task performance: Explaining the impact of grades on intrinsic motivation. *Journal of Educational Psychology*, *105*, 39-57. https://doi.org/10.1037/a0029376

Pultorak, E. G. (1996). Followling the Developmental Process of Reflection in Novice Teachers: Three Years of Investigation. *Journal of Teacher Education*, *47*(4), 283-291.

R: A language and environment for statistical computing. (2023). R Foundation for Statistical Computing.

Raker, J. R., Dood, A. J., Srinivasan, S., & Murphy, K. L. (2021). Pedagogies of engagement use in postsecondary chemistry education in the United States: results from a national survey. *Chemistry Education Research and Practice*, *22*(1), 30-42.

Rapchak, M., Hands, A. S., & Hensley, M. K. (2022). Moving Toward Equity: Experiences With Ungrading. *Journal of Education for Library and Information Science*, *64*(1), 89-98. https://doi.org/10.3138/jelis-2021-0062

Rask, K. (2010). Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review*, *29*(6), 892-900. https://doi.org/10.1016/j.econedurev.2010.06.013

Rebora, S. (2023). Sentiment Analysis in Literary Studies. A Critical Survey. *DHQ: Digital Humanities Quarterly*, *17*(3).

Rehmat, A. P., Diefes-Dux, H. A., & Panther, G. (2021, 2021). Engineering Instructors' Self-Reported Emotions During Emergency Remote Teaching.

Reinholz, D. L., & Andrews, T. C. (2020). Change theory and theory of change: what's the difference anyway? *International Journal of STEM Education*, *7*(2).

Revelle, W. (2024). *psych: Procedures for Psychological, Psychometric, and Personality Research*. In (Version R package version 2.4.3) https://CRAN.R-project.org/package=psych

Rice, K. G., Lopez, F. G., & Richardson, C. M. (2013). Perfectionism and performance among STEM students. *Journal of Vocational Behavior*, *82*(2), 124-134.

Richardson, G., & Maltby, H. (1995). Reflection-on-practice: Enhancing student learning. *Journal of advanced nursing*, *22*(2), 235-242.

Ring, J. (2017). ConfChem Conference on Select 2016 BCCE Presentations: Specifications Grading in the Flipped Organic Classroom. *Journal of Chemical Education*, *94*(12), 2005-2006. https://doi.org/10.1021/acs.jchemed.6b01000

Rittmayer, A. D., & Beier, M. E. (2008). Overview: Self-efficacy in STEM. *Swe-Awe Casee Overviews*, *1*(3), 12.

Roberson, C. (2018). Techniques for using specifications grading in computer science. *Journal of Computing Sciences in Colleges*, *33*(6), 192-193. https://dl.acm.org/doi/abs/10.5555/3205191.3205226

Rocabado, G. A., Komperda, R., Lewis, J. E., & Barbera, J. (2020). Addressing diversity and inclusion through group comparisons: a primer on measurement invariance testing. *Chemistry Education Research and Practice*, *21*(3), 969-988. https://doi.org/10.1039/D0RP00025F

Rodgers, C., & Laboskey, V. K. (2016). Reflective practice. In *International handbook of teacher education* (pp. 71-104). Springer.

Rogers, E. M. (2003). *Diffusion of Innovations* (5, Ed.). Free Press.

Rogers, E. M., & Shoemaker, F. F. (1971). *Communication of innovations: A cross-cultural approach* (2 ed.). Free Press.

Rosovsky, H., & Hartley, M. (2002). *Evaluation and the Academy: Are We Doing the Right Thing? Grade Inflation and Letters of Recommendation*. American Academy of Arts & Sciences.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Ruiz-Primo, M. A., Briggs, D., Iverson, H., Talbot, R., & Shepard, L. A. (2011). Impact of Undergraduate Science Course Innovations on Learning. *Science*, *331*(6022), 1269-1270. https://doi.org/doi:10.1126/science.1198976

Ryan, K., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving Survey Methods With Cognitive Interviews in Small- and Medium-Scale Evaluations. *American Journal of Evaluation*, *33*(3), 414-430. https://doi.org/10.1177/1098214012441499

Ryan, M. (2013). The pedagogical balancing act: Teaching reflection in higher education. *Teaching in Higher Education*, *18*(2), 144-155.

Saluga, S. J., Burns, A. M., Li, Y., Nguyen, M. M., & Edwards, K. D. (2023). A Specifications-Graded, Spice-Themed, General Chemistry Laboratory Course Using an Argument-Driven Inquiry Approach. *Journal of Chemical Education*, *100*(10), 3903-3915.

Sass, D. A. (2011). Testing Measurement Invariance and Comparing Latent Factor Means Within a Confirmatory Factor Analysis Framework. *Journal of Psychoeducational Assessment*, *29*(4), 347-363. https://doi.org/10.1177/0734282911406661

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695-729. https://doi.org/10.1177/0539018405058216 (Social Science Information)

Schinske, J., & Tanner, K. (2014). Teaching more by grading less (or differently). *CBE—Life Sciences Education*, *13*(2), 159-166.

Schneider, J., & Hutt, E. (2014). Making the grade: A history of the A–F marking scheme. *Journal of Curriculum Studies*, *46*(2), 201-224.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.

Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. Jossey-Bass.

Schön, D. A. (1991). *The reflective turn: Case studies in and on educational practice* (Vol. 131). Teachers College Press New York.

Schutz, P. A., Hong, J. Y., Cross, D. I., & Osbon, J. N. (2006). Reflections on Investigating Emotion in Educational Activity Settings. *Educational psychology review*, *18*(4), 343-360. https://doi.org/10.1007/s10648-006-9030-3

Schweingruber, H. A., Nielsen, N. R., & Singer, S. R. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. National Academies Press.

Servant-Miklos, V. F. C. (2019). Fifty Years on: A Retrospective on the World's First Problem-based Learning Programme at McMaster University Medical School. *Health Professions Education*, *5*(1), 3-12.

Seymour, E., & Hewitt, N. M. (1997). *Talking About Leaving: Why Undergraduates Leave the Sciences*. Westview Press.

Seymour, E., Hunter, A.-B., Thiry, H., Weston, T. J., Harper, R. P., Holland, D. G., Koch, A. K., & Drake, B. M. (2019). *Talking about Leaving Revisited : Persistence, Relocation, and Loss in Undergraduate STEM Education*. Springer International Publishing AG. http://ebookcentral.proquest.com/lib/uva/detail.action?docID=5995806

Shadle, S. E., Marker, A., & Earl, B. (2017). Faculty drivers and barriers: Laying the groundwork for undergraduate STEM education reform in academic departments. *International Journal of STEM Education*, *4*, 1-13.

Shapiro, S. (2010). Revisiting the teachers' lounge: Reflections on emotional experience and teacher identity. *Teaching and Teacher Education*, *26*(3), 616-621.

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information*, *22*(2), 63-75.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4-14.

Shields, K., Denlinger, K., & Webb, M. (2019). Not Missing the Point(s): Applying Specifications Grading to Credit-Bearing Information Literacy Classes. In M. Mallon, L. Hays, C. Bradley, R. Huisman, & J. Belanger (Eds.), *The Grounded Instruction Librarian: Participating in The Scholarship of Teaching and Learning* (pp. 87-97). Association of College and Research Libraries. http://hdl.handle.net/10339/94128

Simon, R. A., Aulls, M. W., Dedic, H., Hubbard, K., & Hall, N. C. (2015). Exploring student persistence in STEM programs: a motivational model. *Canadian Journal of Education*, *38*(1), n1.

Simonson, S. R., Earl, B., & Frary, M. (2022). Establishing a framework for assessing teaching effectiveness. *College teaching*, *70*(2), 164-180.

Smith, H., & Schneider, A. (2009). Critiquing models of emotions. *Sociological Methods & Research*, *37*(4), 560-589.

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE—Life Sciences Education*, *12*(4), 618-627.

Snyder, J. J., Sloane, J. D., Dunk, R. D., & Wiles, J. R. (2016). Peer-led team learning helps minority students succeed. *PLoS biology*, *14*(3), e1002398.

Sorby, S., Veurink, N., & Streiner, S. (2018). Does spatial skills instruction improve STEM outcomes? The answer is 'yes'. *Learning and Individual Differences*, *67*, 209-222.

Spalding, E., & Wilson, A. (2002). Demystifying reflection: A study of pedagogical strategies that encourage reflective journal writing. *Teachers college record*, *104*(7), 1393-1421.

Spurlock, S. (2023). *Improving Student Motivation by Ungrading* Proceedings of the 54th ACM Technical Symposium on Computing Science Education V.1 (SIGCSE 2023), Toronto, ON, Canada. https://doi.org/10.1145/3545945.3569747

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., Eagan Jr, M. K., Esson, J. M., Knight, J. K., & Laski, F. A. (2018). Anatomy of STEM teaching in North American universities. *Science*, *359*(6383), 1468-1470.

Stains, M., & Vickrey, T. (2017). Fidelity of implementation: An overlooked yet critical construct to establish effectiveness of evidence-based instructional practices. *CBE—Life Sciences Education*, *16*(1), rm1.

Steiger, J. H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research*, *25*(2), 173-180. https://doi.org/10.1207/s15327906mbr2502_4

Stewart, L. G., & White, M. A. (1976). Teacher comments, letter grades, and student performance: What do we really know? *Journal of Educational Psychology*, *68*(4), 488.

Stowe, R. L., Scharlott, L. J., Ralph, V. R., Becker, N. M., & Cooper, M. M. (2021). You are what you assess: The case for emphasizing chemistry on chemistry assessments. *Journal of Chemical Education*, *98*(8), 2490-2495.

Sturtevant, H., & Wheeler, L. (2019). The STEM faculty instructional barriers and identity survey (FIBIS): Development and exploratory results. *International Journal of STEM Education*, *6*(1), 1-22.

Sumsion, J., & Fleet, A. (1996). Reflection: can we assess it? Should we assess it? *Assessment & Evaluation in Higher Education*, *21*(2), 121-130.

Sutton, R. E., & Wheatley, K. F. (2003). Teachers' Emotions and Teaching: A Review of the Literature and Directions for Future Research. *Educational Psychology Review*, *15*(4), 327-358. https://doi.org/10.1023/a:1026131715856

Syed, M., Zurbriggen, E. L., Chemers, M. M., Goza, B. K., Bearman, S., Crosby, F. J., Shaw, J. M., Hunter, L., & Morgan, E. M. (2019). The role of self-efficacy and identity in mediating the effects of STEM support experiences. *Analyses of Social Issues and Public Policy*, *19*(1), 7-49.

Tajeddin, Z., & Aghababazadeh, Y. (2018). Blog-mediated reflection for professional development: Exploring themes and criticality of L2 teachers' reflective practice. *TESL Canada Journal*, *35*(2), 26-50.

Tan, C. (2008). Improving schools through reflection for teachers: Lessons from Singapore. *School effectiveness and school improvement*, *19*(2), 225-238.

Tea, A. (2024). A Model for Emotional Intelligence in Biology Education Research. *CBE life sciences education*, *23*(4).

Tharayil, S., Borrego, M., Prince, M., Nguyen, K. A., Shekhar, P., Finelli, C. J., & Waters, C. (2018). Strategies to mitigate student resistance to active learning. *International Journal of STEM Education*, *5*, 1-16.

The University of Edinburgh. (2021). *Reflecting on experience*. Retrieved 01/04 from https://www.ed.ac.uk/reflection/reflectors-toolkit/reflecting-on-experience

The University of Kansas Center for Teaching Excellence. (2024). *Benchmarks for teaching effectiveness*. Retrieved 01/03 from https://cte.ku.edu/sites/cte/files/documents/programs-initiatives/KU%20Benchmarks%20Framework%202020update.pdf

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintrón, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones, L., Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., . . . Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the national academy of sciences*, *117*(12), 6476-6483. https://doi.org/10.1073/pnas.1916903117

Thorpe, K. (2004). Reflective learning journals: From concept to practice. *Reflective Practice*, *5*(3), 327-343.

Toledo, S., & Dubas, J. M. (2017). A Learner-Centered Grading Method Focused on Reaching Proficiency with Course Learning Outcomes. *Journal of Chemical Education*, *94*(8), 1043-1050. https://doi.org/10.1021/acs.jchemed.6b00651

Trigwell, K. (2012). Relations between teachers' emotions in teaching and their approaches to teaching in higher education. *Instructional Science*, *40*(3), 607-621. https://doi.org/10.1007/s11251-011-9192-3

Tsoi, M. Y., Anzovino, M. E., Erickson, A. H. L., Forringer, E. R., Henary, E., Lively, A., Morton, M. S., Perell-Gerson, K., Perrine, S., Villanueva, O., Whitney, M., & Woodbridge, C. M. (2019). Variations in Implementation of Specifications Grading in STEM Courses. *Georgia Journal of Science*, *77*(2), 10.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1-10. https://doi.org/10.1007/BF02291170

Turpen, C., & Finkelstein, N. D. (2009). Not all interactive engagement is the same: variations in physics professors' implementation of peer instruction. *Physical Review Special Topics-Physics Education Research*, *5*(2), 020101.

Ulug, M., Ozden, M. S., & Eryilmaz, A. (2011). The effects of teachers' attitudes on students' personality and performance. *Procedia-Social and Behavioral Sciences*, *30*, 738-742.

Undersander, M. A., Lund, T. J., Langdon, L. S., & Stains, M. (2017). Probing the question order effect while developing a chemistry concept inventory. *Chemistry Education Research and Practice*, *18*(1), 45-54.

Unfried, A., Faber, M., Stanhope, D. S., & Wiebe, E. (2015). The development and validation of a measure of student attitudes toward science, technology, engineering, and math (S-STEM). *Journal of Psychoeducational Assessment*, *33*(7), 622-639.

Valli, L. (1997). Listening to other voices: A description of teacher reflection in the United States. *Peabody journal of Education*, *72*(1), 67-88.

van Manen, M. (1977). Linking Ways of Knowing with Ways of Being Practical. *Curriculum Inquiry*, *6*(3), 205-228.

Velicer, W. F., & Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*(2), 231-251. https://doi.org/10.1037/1082-989X.3.2.231

Vitale, S. E., & Concepción, D. W. (2021). Improving Student Learning with Aspects of Specifications Grading. *Teaching Philosophy*, *44*(1), 29-57.

von Renesse, C., & Wegner, S. A. (2023). Two Examples of Ungrading in Higher Education in the United States and Germany. *PRIMUS*, *33*(9), 1035-1054. https://doi.org/10.1080/10511970.2023.2229819

Walden, P. R. (2022). Student Motivation, Anxiety and Pass/Fail Grading: A SoTL Project. *Teaching and Learning in Communication Sciences & Disorders*, *6*(1), 13. https://doi.org/10.30707/TLCSD6.1.1649037808.651639

Wang, Y., Apkarian, N., Dancy, M. H., Henderson, C., Johnson, E., Raker, J. R., & Stains, M. (2024). A National Snapshot of Introductory Chemistry Instructors and Their Instructional Practices. *Journal of Chemical Education*, *101*(4), 1457-1468.

Watkins, M. W. (2018). Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology*, *44*(3), 219-246. https://doi.org/10.1177/0095798418771807

Weaver, G. C., Austin, A. E., Greenhoot, A. F., & Finkelstein, N. D. (2020). Establishing a better approach for evaluating teaching: The TEval Project. *Change: The Magazine of Higher Learning*, *52*(3), 25-31.

Whitcomb, K. M., & Singh, C. (2021). Underrepresented minority students receive lower grades and have higher rates of attrition across STEM disciplines: A sign of inequity? *International Journal of Science Education*, *43*(7), 1054-1089.

White, K. N., Vincent-Layton, K., & Villarreal, B. (2021). Equitable and Inclusive Practices Designed to Reduce Equity Gaps in Undergraduate Chemistry Courses. *Journal of Chemical Education*, *98*(2), 330-339. https://doi.org/10.1021/acs.jchemed.0c01094

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY. https://ggplot2.tidyverse.org

Williams, B. A. (2022). Teaching Students to Feel Pleasure and Pain at the Wrong Thing: The History of Grades and Grading. *Principia: A Journal of Classical Education*.

Williams, K. (2018). Specifications-Based Grading in an Introduction to Proofs Course. *PRIMUS*, *28*(2), 128-142. https://doi.org/10.1080/10511970.2017.1344337

Williams, L., Arribas-Ayllon, M., Artemiou, A., & Spasić, I. (2019). Comparing the Utility of Different Classification Schemes for Emotive Language Analysis. *Journal of Classification*, *36*(3), 619-648. https://doi.org/10.1007/s00357-019-9307-0

Wilson, D., Bates, R., Scott, E. P., Painter, S. M., & Shaffer, J. (2015). Differences in self-efficacy among women and minorities in STEM. *Journal of Women and Minorities in Science and Engineering*, *21*(1).

Winchester, T. M., & Winchester, M. (2011). Exploring the impact of faculty reflection on weekly student evaluations of teaching. *International Journal for Academic Development*, *16*(2), 119-131.

Winget, M., & Persky, A. M. (2022). A Practical Review of Mastery Learning. *American Journal of Pharmaceutical Education*.

Witherspoon, E. B., Vincent-Ruz, P., & Schunn, C. D. (2019). When Making the Grade Isn't Enough: The Gendered Nature of Premed Science Course Attrition. *Educational Researcher*, *48*(4), 193-204. https://doi.org/10.3102/0013189X19840331

Witteveen, D., & Attewell, P. (2020). The STEM grading penalty: An alternative to the "leaky pipeline" hypothesis. *Science Education*, *104*(4), 714-735. https://doi.org/10.1002/sce.21580

Wong, F. K., Kember, D., Chung, L. Y., & CertEd, L. Y. (1995). Assessing the level of student reflection from reflective journals. *Journal of advanced nursing*, *22*(1), 48-57.

Wood, D. F. (2003). Problem based learning. *Clinical Review*, *326-330*, 328-.

Woodbury, S., & Gess-Newsome, J. (2002). Overcoming the paradox of change without difference: A model of change in the arena of fundamental school reform. *Educational Policy*, *16*(5), 763-782.

Wu, S. P., & Rau, M. A. (2019). How students learn content in science, technology, engineering, and mathematics (STEM) through drawing activities. *Educational psychology review*, *31*, 87-120.

Wu, X., Deshler, J., & Fuller, E. (2018). The effects of different versions of a gateway STEM course on student attitudes and beliefs. *International Journal of STEM Education*, *5*, 1-12.

Yik, B. J., Machost, H., Streifer, A. C., Palmer, M. S., Morkowchuk, L., & Stains, M. (2024). Students' Perceptions of Specifications Grading: Development and Evaluation of the Perceptions of Grading Schemes (PGS) Instrument. *Journal of Chemical Education*.

Yik, B. J., Raker, J. R., Apkarian, N., Stains, M., Henderson, C., Dancy, M. H., & Johnson, E. (2022a). Association of malleable factors with adoption of research-based instructional strategies in introductory chemistry, mathematics, and physics. Frontiers in Education,

Yik, B. J., Raker, J. R., APkarian, N., Stains, M., Henderson, C., Dancy, M. H., & Johnson, E. (2022b). Evaluating the impact of malleable factors on percent time lecturing in gateway chemistry, mathematics, and physics courses. *International Journal of STEM Education*, *9*(15). https://doi.org/10.1186/s40594-022-00333-3

Yin, R. K. (2018). *Case Study Research and Applications: Design and Methods* (6th ed.). Sage Publications.

Ying, J., Qiu, J., Wu, Y., Zhao, L., & Bai, Y. (2023). An Investigation on the Application of a Competency Assessment System in a Blended Learning Course "Organic Chemistry Laboratory". *Journal of Chemical Education*, *100*(10), 3916-3924.

York, S., & Orgill, M. (2023). Experienced tertiary instructors' perceptions of the benefits and challenges of systems thinking in chemistry education. *Journal of Chemical Education*, *101*(1), 10-23.

Young, A. M., Wendel, P. J., Esson, J. M., & Plank, K. M. (2018). Motivational decline and recovery in higher education STEM courses. *International Journal of Science Education*, *40*(9), 1016-1033.

Zahid, M., & Khanam, A. (2019). Effect of Reflective Teaching Practices on the Performance of Prospective Teachers. *Turkish Online Journal of Educational Technology-TOJET*, *18*(1), 32-43.

Zeichner, K., & Liston, D. (1987). Teaching student teachers to reflect. *Harvard educational review*, *57*(1), 23-49.