

GRADING IN CHEMISTRY: VARIATIONS IN INSTRUCTORS' EVALUATION OF
STUDENT WRITTEN RESPONSES

by

Michelle Herridge

Copyright © Michelle Herridge 2021

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2021

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: Michelle Herridge

titled: Grading in Chemistry: Variations in Instructors' Evaluation of Student Written Responses

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.



Vicente Talanquer

Date: May 27, 2021



Michael Brown

Date: May 27, 2021



Andrei Sanov

Date: May 28, 2021



Kristin Gunckel

Date: May 28, 2021

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.



Vicente Talanquer
Chemistry and Biochemistry

Date: May 27, 2021

Signature: *Michelle Herridge*

Email: mdhh99@email.arizona.edu

ACKNOWLEDGEMENTS

This dissertation would not be possible without the support, advice, and guidance of several people. First and foremost, I need to thank Dr. Vicente Talanquer for supporting my growth, pushing me to reevaluate my perspectives, and helping me through this process. I am a better researcher and practitioner because of the experiences and opportunities you have provided. Thank you so much.

Thank you to my committee members who have provided such valuable guidance and feedback, Dr. Kristin Gunckel, Dr. Michael Brown, and Dr. Andrei Sanov. Thank you to my mentors and professors who have answered so many questions and provided immeasurable support, Dr. Karen Spear-Ellinwood, Dr. Sanlyn Buxner, and Dr. Erin Galyen.

A special thank you to the members of CBC who have offered me opportunities to try new things and get involved. A particular thanks is owed to Dr. Amy Graham, Dr. Tori Hidalgo, and Mark Yanagihashi, who have believed in me and supported me intellectually and financially. Education research does not happen without willing participants who will answer your questions and let you invade meetings and classrooms. Thank you to all my participants.

The support of my friends and family cannot be overstated, as they have put up with my moods and rants for the last few years and supported my work. Thank you to my friends, Jenna Tashiro, Bailey Anderegg, Ali Kane, and Sara Zachritz, who have kept me sane and talked me through emails, presentations, tears, and everything else. Many others deserve thanks, too many to list. Thank you to my husband, Robert Bailey, for feeding me, loving me, and making me take breaks. Last, but definitely not least, thank you to my parents, Vicki and Van Herridge, who never expected anything less from me than to be a doctor.

DEDICATION

I dedicate this work to two teachers who have done more than they realize

and have impacted my life in huge ways:

Ms. Carole Rathbun and Dr. Gautam Bhattacharyya

TABLE OF CONTENTS

LIST OF FIGURES	8
LIST OF TABLES	10
LIST OF BOXES	10
ABSTRACT.....	11
CHAPTER 1: INTRODUCTION	13
CHAPTER 2: REVIEW OF LITERATURE.....	15
Assessment.....	15
Evaluation	16
Grading	16
Noticing.....	19
Problem Statement	21
CHAPTER 3: DIMENSIONS OF VARIATION IN CHEMISTRY INSTRUCTORS' APPROACHES TO THE EVALUATION AND GRADING OF STUDENT RESPONSES	24
Methods.....	25
<i>Context and Participants</i>	25
<i>Data Collection</i>	25
<i>Data Analysis</i>	27
Major Findings.....	30
<i>Variation In Grading</i>	30
<i>Dimensions Of Variation</i>	31
Limitations	41

Discussion.....	42
Implications.....	45
Conclusions.....	47
CHAPTER 4: VARIATION IN CHEMISTRY INSTRUCTORS' EVALUATION OF	
STUDENT WRITTEN RESPONSES AND ITS IMPACT ON GRADING	48
Methods.....	49
<i>Setting and Participants</i>	49
<i>Exam Questions & Student Responses</i>	50
<i>Data Collection</i>	52
<i>Data Analysis</i>	52
Results.....	55
<i>Characterization of Variation in Evaluation Behaviour</i>	55
<i>Correlation of Dimensions with Assigned Grades</i>	73
Discussion.....	76
<i>Common Main Approaches</i>	77
<i>Consistency in Approach to Evaluation</i>	80
<i>Correlation of Dimensions with Assigned Grades</i>	83
Limitations	86
Conclusions and Implications.....	86
CHAPTER 5: CONCLUSIONS AND FUTURE DIRECTIONS	
Implications.....	91
Additional Data Collected and Possible Future Work.....	92
<i>Content and Expectations</i>	93

<i>Self-Assessment and Grading Workshop</i>	94
<i>Negotiation</i>	95
APPENDIX A – HUMAN SUBJECT APPROVAL.....	97
APPENDIX B - IRB CONSENT FORM	98
APPENDIX C - INTERVIEW PROTOCOL	100
APPENDIX D - INTERVIEW DOCUMENTS	104
APPENDIX E - INSTRUCTORS’ DOMINANT APPROACHES.....	116
APPENDIX F – PUBLICATION PERMISSIONS.....	118
COMPLETE LIST OF REFERENCES.....	119

LIST OF FIGURES

Figure 1: Graphical Abstract.....	12
Figure 2 (a) and (b): Student responses selected for evaluation, grading and feedback during interviews.....	28
Figure 3 (a) and (b): Distribution of grades for (a) Student Response 1 and (b) Student Response 2 assigned by graduate student instructors (GSIs) and faculty instructors (FIs).	31
Figure 4: Examples of explicit rubrics and expected answers built by interviewed instructors...	35
Figure 5: Different types of visual marks made by instructors while grading.....	40
Figure 6: Distribution of behaviors along dimensions of analysis in instructor's approaches to grading in our study. Frequencies are expressed in terms of number of instructors ($n = 28$) for the first five dimensions and in terms of instances of evaluation ($n_e = 56$) for the others.	42
Figure 7: Distribution of behaviors in the evaluation of all samples of student work along each dimension of analysis in our study. Frequencies are expressed in terms of the number of graded samples ($n = 268$).....	56
Figure 8: Dominant approaches in the evaluation of student work manifested by participating instructors in each dimension of analysis. Instructors who exhibited variable approaches are included in the middle bar in each dimension. Frequencies are expressed in terms of number of instructors ($n = 29$).	56
Figure 9: Examples of Rubrics and Expected Answers.....	61
Figure 10: Sample of instructor response to student work: GSI Dylan, Q2, Student Response 2, including expressed interaction, evaluative intention marking, and indirect and direct feedback.	73

Figure 11: Number of dimensions in which instructors are variable..... 81

LIST OF TABLES

Table 1: Major Dimensions of Variation in the Evaluation and Grading of Student Responses and Contrasting Approaches within Each Dimension	29
Table 2: Participant Information.....	49
Table 3: Average Scores Assigned	74
Table 4: Predicted grades for an average student response	76

LIST OF BOXES

Box 1: Question Prompt.....	27
Box 2: All Question Prompts.....	51

ABSTRACT

Evaluation of student written work during summative assessments is an important and critical task for instructors at all educational levels. Nevertheless, few research studies exist that provide insights into how different instructors approach this task. Chemistry faculty (FIs) and graduate student instructors (GSIs) regularly engage in the evaluation and grading of student written responses in formal formative and summative assessments. In this study, we characterized how different instructors teaching general chemistry at the University of Arizona differed in their approaches to the evaluation and grading of students' written answers in midterm exams. This work is critical to support the professional development of instructors to ensure a fair evaluation of all students.

In the first part of this project, we identified and characterized dimensions of variation in general chemistry instructors' approaches to the evaluation and grading of a conceptual question. Using qualitative methods of research, we conducted individual interviews in which participating chemistry FIs and GSIs were asked to evaluate and grade the same set of students' responses to a typical exam prompt and justify their decisions. Our results showed that observed variability in assigned grades emerged from the complex interaction of explicit and implicit decisions made and preferences manifested along various dimensions. Based on this characterization, we developed an analytical framework to characterize variation in the evaluation approaches of chemistry instructors when grading free response questions

In the second stage of this project, we applied the analytical framework to characterize variation in the evaluation and grading of a diverse set of open response questions and to analyze the impact of such variations on assigned grades. Our results revealed a wide variation in the approaches followed by instructors along different dimensions, although most instructors tended

to be consistent in the approaches they individually followed when analyzing different student responses to diverse questions. Instructors' experience seemed to impact assigned grades but relevant dimensions of variation in the evaluation differed for less and more experienced instructors.



Figure 1: Graphical Abstract

CHAPTER 1: INTRODUCTION

Formal formative and summative assessments are used in college chemistry courses to gather information about student learning. These types of assessments sometimes contain free-response or open-ended questions in which students are asked to summarize ideas, build explanations, or show their work when finding solutions to quantitative problems. These types of questions are encouraged by chemistry educators who seek to improve and diversify the nature of both formative and summative assessments used in college chemistry (Stowe & Cooper, 2019). In large-enrollment courses, students' answers to these questions are frequently graded with the assistance of graduate student instructors (Mutambuki & Schwartz, 2018) who are often provided with an evaluation rubric built by the chemistry faculty in charge of the course. Even if a rubric is used, instructors must make multiple decisions during the grading process about how and what to evaluate in the student work (Warren, 1971). Their decisions are influenced by diverse factors such as their prior knowledge and experience with the course material and their personal beliefs about learning the subject matter (Henderson et al., 2004; Marshman et al., 2017; Mutambuki & Fynewever, 2012; Petcovic et al., 2013). These factors are expected to introduce variability in the grading as they affect what the graders notice in student work and how they interpret it in relation to an evaluation rubric. These variations in the evaluation of student work have led to concerns about fairness (Holmes & Smith, 2003) that are particularly relevant when working with large numbers of diverse students who may express their understanding in quite different ways

Recent work on teacher noticing suggests that instructors' evaluation of student understanding may differ in various dimensions (Jacobs et al., 2010; Sherin et al., 2011; Talanquer et al., 2015; Van Es, 2011). Teachers might, for example, mostly describe what students are saying or writing (descriptive approach) versus building inferences based on the analysis of student work

(inferential approach). While building inferences of student understanding, they might focus on the level of correctness of students' answers (evaluative orientation) rather than trying to make sense of why students struggle with the targeted concepts and ideas (interpretive orientation). Guided by existing research in the area of noticing, in this study we sought to characterize different dimensions of variability in chemistry faculty and graduate student instructors' evaluation and grading of students' answers to free-response questions (we use the term "graduate student instructor," or GSI, instead of "graduate teaching assistant" as we believe it better represents the diverse instructional activities in which graduate students are asked to engage (Stacy, 2000).)

The central goal of our investigation was to characterize different dimensions of variability in the evaluation of student work and develop an analytical framework that could help increase reliability in this evaluation and support the training and professional development of college instructors. Assessment and evaluation practices in foundational science courses are known to have a major impact on students' progression and retention in STEM majors (Salehi et al., 2019); better understanding how these tasks are approached by different instructors is also critical to foster equity and broad inclusion in chemistry education.

CHAPTER 2: REVIEW OF LITERATURE

Assessment

Assessment is a broad term that is sometimes used interchangeably with evaluation and grading, but the three are distinct in several ways (Rea, 2010). Assessment is the gathering of information, evaluation is a judgement related to the collected information, and grading is an assignment of value based on the judgement. In education, there are two types of assessment: formative or summative. Formative assessment is often used as low-stakes assignments designed to gather information on a short-term scale about what students do and do not understand and how it might be best to proceed. A summary and interpretation of this information tells the instructor what strategies could be used to enhance student learning and the next steps in the learning cycle. Summative assessments provide information about whether students have met the learning goals over a longer time period; for example, the module, unit, or course. Both formative and summative assessments are used to gather information about students, and either may be used for evaluation and grading, though summative assessments are more likely to be high-stakes and therefore more impactful on the overall success of a student in a course (Glazer, 2014; Harlen & James, 2006; Harshman & Yeziarski, 2016; Sadler, Royce, 1989).

Assessment is a critical component of the learning cycle, as it informs instructors from the minute-level (with formative assessment through discussions or questions during a class) to the degree-level (with overall GPA and summative assessments like a dissertation) about how a student is performing, what is understood, and whether they are ready to proceed to the next stage (the next problem, course, or job). Evaluation, then, is the comparison of the collected information about a student's understanding and performance to the stated learning outcomes. Evaluation and

assessment are often used interchangeably to describe this act of judgement (Cullingford, 1997; Taras, 2005).

Evaluation

Evaluation is a judgement associated with the performance on an assessment task. While assessment concerns the collection of information, evaluation involves a comparison of the information with expectations and how closely the students' performance aligns with the expected learning outcomes. Evaluation of student work can take place through comparing student work with learning objectives, a rubric, other student answers, or expert judgement (Adie et al., 2012; Gibbons et al., 2018; Moskal, 2000; Sabel et al., 2016).

Evaluation is different from assessment and from grading in that evaluations can be more holistic in considering the student, be performed by someone other than the instructor, and may have different purposes. For example, students engaging in peer reviews with one another is a form of evaluation that assesses student work without providing a grade (Connor et al., 2019; Falchikov & Goldfinch, 2000). Ungraded assignments that receive feedback from the instructor may also fill this role.

Grading

Grading can be defined as something "earned" by students in return for academic achievement as judged or evaluated by an expert (Brookhart et al., 2016). Assigned grades serve alternative or additional purposes related to motivation, encouragement, or reward for student work (Stanley & Baines, 2001). Grading is the assignment of a ranking or score for a student based on their performance or understanding and is the main method of reporting evaluation and assessment, particularly in higher education. Various methods of ranking students have been used

at different institutions, including standards-based grading, letter grades, number grades, and categorical grades (Schwab et al., 2018). The use of grades has been long established and integrated into educational institutions. Therefore, grades can predict student success and persistence in schooling (Brookhart et al., 2016). With respect to systemic inequalities, research shows that privileged students are more likely to attend office hours (Guerrero & Rod, 2013) and/or ask for regrades (Li & Zafar, 2020), both of which correlate with higher grades. Concerns about grades as the outcome measure highlight the limited information that is portrayed by a single number or letter. Further, concerns about the equitable assignment and interpretation of traditional grades are emerging (Baines & Stanley, 2000; Salehi et al., 2019).

Studies on the reliability of teachers' grades have shown that great variations might be present in the grades assigned by the same or different graders to similar samples of student work. Records of these broad and inconsistent grading practices have existed since the early part of the twentieth century and indicate the subjectivity of assigning grades has been problematic in multiple disciplines (Starch & Elliott, 1912). Instructors' subjective inferences about student understanding can be expected to introduce variability in the evaluation of student work. Grading variability can thus stem from the evaluation stance adopted by graders, which might lead them to judge answers based merely on the degree to which they match a normative response or to engage on inferential analysis of expressed student thinking (Aydeniz & Dogan, 2016; Gotwals & Birmingham, 2016; Haagen, 1964). Meaningful engagement with students' ideas is frequently inhibited by the graders' limited understanding of the learning objectives of a course (Ebby et al., 2019), misinterpretation of a rubric or its components (Eckes, 2008), low grading proficiency (Wolfe et al., 1998), and personal misconceptions about how understanding is demonstrated (Marshman et al., 2017). Different authors have shown that some instructors approach grading as a black and white

judgement (Haagen, 1964), looking at answers as either right or wrong (Gotwals & Birmingham, 2016). Additional studies have shown variation in grades are related to differences in grading criteria (Marshman et al., 2017), the beliefs that the graders have about the purpose of the assessment (Petcovic et al., 2013), and the relative values graders place on different components of student work (Henderson et al., 2004; Mutambuki & Fynewever, 2012; Petcovic et al., 2013).

Research on the evaluation and grading practices of chemistry (Mutambuki & Fynewever, 2012; Petcovic et al., 2013) and physics (Henderson et al., 2004; Marshman et al., 2017) faculty and GSIs suggests that there is often a gap between the beliefs graders hold about what is desirable in students' answers and their practices when scoring actual student responses. Research in this area suggests that variations in the grading of student work are often linked to different conceptions on the "*burden of proof*" during the analysis of student work (Yerushalmi et al., 2016). Instructors often expect students to show their reasoning to evaluate their understanding. Nevertheless, they tend to project correct thought processes onto students' answers even when the expected reasoning is not explicitly included in the response (Henderson et al., 2004; Marshman et al., 2017; Mutambuki & Fynewever, 2012; Petcovic et al., 2013). For example, instructors may infer that a student knows how to solve a problem based on the presence of a correct numerical response even if mistakes are present in the procedure (Petcovic et al., 2013). Some faculty readily subtract points from explicitly incorrect answers but are hesitant to deduct points from more vague answers that could be interpreted as correct (Henderson et al., 2004).

Efforts to align grading practices and narrow the variability often rely on the use of rubrics, which "provide an effective, efficient, and equitable assessment method that can be understood and applied by both student learner and academic assessors" (Allen & Knight, 2009, p. 1). In addition to aligning grading practices, students' concerns related to the fairness of norm-referenced

grading or bell curves have been studied (Close, 2009; Kulick & Wright, 2008; Satria et al., 2020; Tan et al., 2020), and rubrics have been proposed as solutions to this complaint. Perceptions of fairness do increase with the use of these tools (Andrade, 2005; Holmes & Smith, 2003; Knight et al., 2012; Randall & Engelhard, 2010). Nevertheless, research suggests that grading inter- and intra-rater reliability marginally increases with rubric use (Howell, 2011; Kohn, 2006; Rezaei & Lovorn, 2010). Different graders are known to use rubrics in ways that depend on their prior knowledge and experience (Pula & Huot, 1993), which affect how they weigh different components of the scoring criteria (Wolfe et al., 1998). In chemistry education, the use of rubrics has been studied in the context of assessing and evaluating critical thinking (Reynders et al., 2020), laboratory skills (Harwood et al., 2020), and science writing (Logan & Mountain, 2018). Although ample resources exist on how to design and use rubrics in the classroom (Dawson, 2017; Mertler, 2001; Mohl et al., 2017; Reynders et al., 2020) and the laboratory (Fay et al., 2007; Harwood et al., 2020), most studies have focused on the use of rubrics to explicitly set expectations and provide feedback to students rather than on characterizing how rubrics are used by instructors to evaluate and grade student work.

Noticing

Teachers' prior knowledge and experiences, as well as personal beliefs about teaching and learning affect what teachers notice in student work and how they interpret it during the evaluation and grading processes (Abell & Siegel, 2011). Work on teacher noticing (Chan et al., 2020) has focused on the formative assessment practices of K-12 pre-service and in-service teachers (Barnhart & Van Es, 2015; Huang & Li, 2012; Jacobs et al., 2010; Luna, 2018; Ross & Gibson, 2010; Sherin et al., 2011; Talanquer et al., 2015; Van Es, 2011). Existing research in teacher thinking and practice suggests that what instructors notice in student work affects the inferences

and judgments that they make. Teacher noticing has thus become a productive “*line of research [which] investigates what teachers ‘see’ and how they make sense of what they see in classrooms*” (Chan et al., 2020). Much of the work in this area has focused on formative assessment practices in K12 classrooms (Amador et al., 2017; Huang & Li, 2012; Luna, 2018; Ross & Gibson, 2010), where the evaluation is expected to be continuous, iterative, and highly interpretive (Ainley & Luntley, 2007; Barnhart & Van Es, 2015; Ebby et al., 2019). But what teachers’ notice in student written work also influences their evaluation behavior (Herridge & Talanquer, 2020; Talanquer et al., 2013, 2015).

Although researchers have different views on what teacher noticing entails, two major elements are commonly identified: a) Attention to students' expressed ideas, knowledge, or behaviors, and b) interpretation of what is noticed with instructional purposes (Jacobs et al., 2010; Luna, 2018; Sherin et al., 2011; Van Es, 2011). Diverse personal and contextual factors affect what is noticed (object of noticing), the nature of the inferences that are built (noticing stance), and the specificity in what is attended to (Talanquer et al., 2013). Studies in this area indicate that teachers often a) struggle to pay attention to student thinking (Russ et al., 2009), b) adopt an evaluative rather than an interpretive stance in the analysis of student work (Aydeniz & Dogan, 2016; Crespo, 2000; Murray et al., 2020), and c) make general evaluations with little evidence (Talanquer et al., 2015).

Teachers' noticing and interpretation of student written work is known to vary in different dimensions (Jacobs et al., 2010; Sherin et al., 2011; Talanquer et al., 2015; Van Es, 2011). Some of these dimensions are domain-neutral and help characterize whether teachers' approach to the evaluation is descriptive versus inferential, whether their orientation in the analysis of students' responses is more evaluative (focus on correctness) or interpretive (focus on making sense), and

whether they attend to specific or general aspects of students' expressed ideas. Domain-dependent dimensions help characterize the extent to which teachers' inferences are supported with proper evidence, the types of knowledge or ways of reasoning the teachers notice, and the scope of the evaluation (narrowly focus on pieces of knowledge or on the integration of ideas).

Although formative and summative assessments have different goals and often involve different procedures, we believe that analytical frameworks used by researchers on teacher noticing can be helpful in the identification and characterization of factors that influence how college instructors approach the evaluation and grading of student work. Differences along various relevant dimensions can help to make sense of observed variability in evaluative judgments and grading decisions and provide insights into effective strategies to reduce it.

Problem Statement

There are over fifty journals related to educational assessment, evaluation, and grading, and subdisciplines of assessment include writing, higher education, and specific disciplines. Much of the research in assessment focuses on the development of instruments that assess student understanding, testing the validity of those instruments, and how to use the results (Brooker et al., 1998; Cooper & Sandi-Urena, 2009; Nielsen & Yeziarski, 2015). Additional work on assessment practices has considered classroom factors (Duncan & Noonan, 2007), effects on students (Braund & DeLuca, 2018; Dochy et al., 1999), and the types of assessment used (Panadero et al., 2019). Less work has been done on how instructors engage with student material and make decisions during the evaluation process. Some work in this area has been a result of perceived conflict between stated expectations and assigned valuations (Mutambuki & Fyneweever, 2012; Petcovic et al., 2013; Popova et al., 2020), while other work is related to the development and use of rubrics to align multiple instructors (Andrade, 1997; Brown et al., 2013; Dawson, 2017; Harwood et al.,

2020). Rubric use, while increasing interrater reliability and reducing some of the variation in grading (Howell, 2011; Stellmack et al., 2009), considers the expectations put forth by an instructor or writer, but does not consider the student work.

Considering the student work during assessment can be done through professional noticing. Research in teacher noticing started in K12 classrooms; primarily in mathematics (Amador et al., 2017; Barnhart & Van Es, 2015; Van Es, 2011). Noticing is a productive framework for considering the professional analysis of information during formative assessment (Cowie et al., 2018; Furtak & Thompson, 2016). Due to its productivity, noticing is now an area of research, particularly in relation to supporting preservice teachers (Amador et al., 2021; Aydeniz & Dogan, 2016), and has crossed disciplinary boundaries into science (Chan et al., 2020; Luna, 2018; Talanquer et al., 2013, 2015), though it has not yet been applied to summative assessments.

In this project, professional noticing is explored in conjunction with grading practices to define and characterize the implicit and explicit decisions that instructors make during the grading and evaluation process in order to better understand what occurs during grading, why instructors assign different grades to the same work, and the implications of those different decisions. The findings of this research may give insights that allow for professional development or teaching practices that align assessment within and between instructors and reduce the variation in assigned grades. Specifically, the following questions are addressed:

1. What are the major dimensions of chemistry instructors' approaches to the evaluation and grading of student work that give rise to variation?
2. How do approaches to evaluation and grading vary within each identified dimension?
3. What main approaches across different dimensions of variation in the evaluation and grading of student work do chemistry instructors more commonly follow?

4. How consistent are chemistry instructors in their approach to the evaluation and grading of student work across different dimensions of variation?
5. What individual variables and dimensions of variation correlate to assigned grades?

CHAPTER 3: DIMENSIONS OF VARIATION IN CHEMISTRY INSTRUCTORS' APPROACHES TO THE EVALUATION AND GRADING OF STUDENT RESPONSES¹

Given the importance of grades assigned to student work, we first explored dimensions of variation in the evaluation and grading of introductory chemistry free response exam questions by different types of instructors. This first part of our investigation was guided by the following research questions:

- What are the major dimensions of chemistry instructors' approaches to the evaluation and grading of student work that give rise to variation?
- How do approaches to evaluation and grading vary within each identified dimension?

The central goal of this part of the study was to develop an analytical framework that could be used to investigate and analyze instructors' behaviors while evaluating and grading students' written responses to conceptual questions. This framework can guide professional development of chemistry instructors in the evaluation of student understanding and learning.

¹ The main content of this chapter is reprinted (adapted) with permission from (Herridge & Talanquer, 2020). Copyright 2021. American Chemical Society.

Methods

Context and Participants

This project was undertaken at a large public university in the southwest US. Study participants were faculty instructors (FIs) and graduate student instructors (GSIs) teaching 100-level general chemistry for science and engineering majors ($n = 29$; 22 GSIs, 7 FIs). All FIs teaching this course follow the same curriculum, share the same learning objectives, and build common midterm exams together. These midterm exams typically include a set of open response questions and students' answers that are graded by GSIs who use a rubric generated by the FIs to guide the process. This rubric is distributed, explained, and applied in a grading meeting following each exam. All participants in our study had taught at least one full semester of the course; FIs had taught the lecture section and the GSIs had taught at least one associated laboratory section, attended lecture, and graded exams. So, all study participants were familiar with the course structure, learning objectives, and assessment practices. Of the GSIs, 18 were in their first year of teaching at this university and the others had four or more semesters of experience teaching the general chemistry labs and assisting with other course components such as grading exams. Of the FIs, five had three or more years of experience teaching the general chemistry courses and two had been teaching them for less than two years. From now on, we refer to all study participants as instructors. This research project was approved by the IRB at our institution and all subjects consented to participate in the study.

Data Collection

The 29 instructors individually met with the researcher for a video-recorded, semi-structured interview which lasted from 30 to 180 minutes. In this interview, participants were

asked to read a sample exam question from a past exam in the general chemistry course, prepare for the evaluation of this question, and then evaluate and grade selected student responses. This process was repeated with six different exam questions and their associated student responses. After evaluation and grading, instructors were asked to provide written feedback to each of the student responses. Subsequently, the interviewer explored the instructors' rationale for their grade assignments and feedback. This interaction occurred after the evaluation of all questions had been completed to reduce the influence of the interviewer on the instructors' evaluation decisions. All participants had the option to skip the evaluation of any given question if they were unfamiliar with the subject matter.

The findings reported in this paper emerged from the analysis of instructors' evaluation and grading of one of the questions in the sample. The prompt for this question is presented in Box 1. This six-point question was selected because it was representative of both a) the types of free-response questions included in the exam for the general chemistry course, and b) the wide variation in assigned scores observed in the grading of students' responses during the interviews. The points assigned to the question were retained from the original exam structure. Exams in the general chemistry course for science and engineering majors at our institution commonly include 6 to 8 open response questions with a point value between 4 to 8 points, for a total of 40 points. We kept the same point value originally assigned to the questions due to the instructors' experience and familiarity with the typical point distribution for exams in the course. During the interview, instructors could divide and assign the six points based on their own evaluation decisions. The same two student responses were evaluated by each instructor (see Figure 2 (a) and (b)). These responses were photocopied and provided without alteration on the same page for grading and then separately for feedback. Only one instructor declined to evaluate this question, citing a lack of

confidence in understanding the subject matter. This reduced the number of analyzed interview transcripts and products to 28.

Box 1: Question Prompt

Pools constantly need to be replenished with water due to evaporation. One of the assumptions of the particulate model of matter is that particles exhibit a distribution of speed. Using the particulate model, explain how water evaporates and why the assumption about particles having a distribution of speed is essential to explaining evaporation.

Data Analysis

The grades assigned by participating instructors to the selected students' responses were analyzed using descriptive statistical methods to characterize grading variability in our sample. Subsequently, interview recordings were transcribed, and all qualitative data (e.g., observation notes, instructors' annotations) were analyzed using a standard iterative approach in the analysis of qualitative written data (Creswell, 2012; Miles & Huberman, 1994). The goal of this analysis was to identify dimensions of variability in instructors' approaches to the evaluation of student responses that affected their grading decisions. The analysis was guided by existing frameworks on teacher noticing (Jacobs et al., 2010; Sherin et al., 2011; Talanquer et al., 2015; Van Es, 2011), but new dimensions and categories of analysis emerged from the iterative analysis of the data (Maxwell & Miller, 2008). Initial coding helped characterize differences in, for example, whether graders based their evaluative judgments on what was explicitly stated in a student's response or inferred students' level of understanding from implicit cues (e.g., extent to which students' reasoning seemed coherent). This dimension of analysis was labeled "Noticing Stance" and the different approaches to the evaluation of student work within the dimension (e.g., Literal, Inferential) were thought as categories of behavior within that dimension. Other dimension of analysis included depth to which graders analyzed the question prompt, their focus of attention

Pools constantly need to be replenished with water due to evaporation. One of the assumptions of the particulate model of matter is that particles exhibit a distribution of speed. Using the particulate model, explain how water evaporates and why the assumption about particles having a distribution of speed is essential to explaining evaporation?

(a) Using the particulate model of water, water is able to evaporate due to the distribution of speed the particles exhibit. As the surface area increases the particles have an increased area to move around, eventually these particles are able to separate as the weak forces between them break. As temperature of the air increases, temperature of water increases allowing them to gain potential energy, as temp. increases, so does the average particle speed, as temp. increases the forces get weaker between the particles allowing them to go from liquid to gas and evaporate.

Pools constantly need to be replenished with water due to evaporation. One of the assumptions of the particulate model of matter is that particles exhibit a distribution of speed. Using the particulate model, explain how water evaporates and why the assumption about particles having a distribution of speed is essential to explaining evaporation?

(b) Particles that evaporate have greater speed than particles still in liquid phase. Most people assume that all particles have the same speed, this is false. Although all particles have different speeds, as a whole, they all have the same average speed. Since some particles have greater speeds than others, they also have more configurations which allows for some particles to escape their phase. Not all particles leave/enter a phase at the same time. Some particles might change phases faster or longer than others. This is why not all water evaporates at the same time & also why your water bottle doesn't freeze in one second. (sometimes half is frozen)

Figure 2 (a) and (b): Student responses selected for evaluation, grading and feedback during interviews.

while reading the prompt, the creation of a rubric to guide the evaluation, and the grading system they seemed to rely on when assigning scores.

Table 1 summarizes the different dimensions of variation in the evaluation of student work that emerged from our analysis and the contrasting approaches observed within each dimension. Major dimensions of variability were organized into four different groups as we recognized distinct behaviors in instructors' approaches to the evaluation and grading of student work at different stages of the process (e.g., while preparing for grading versus while responding to student work). The analytical framework presented in Table 1 resulted from the analysis carried out by two researchers through repeated cycles of independent coding followed by comparison of assigned

Table 1: Major Dimensions of Variation in the Evaluation and Grading of Student Responses and Contrasting Approaches within Each Dimension

Dimension	Approaches (Dichotomous Spectrum)	
Stage 1: Reading and Interpreting the Prompt		
Depth	Skimming: Grader quickly reads the prompt, looking for target ideas. Focuses on one part of the question	Close Reading: Grader carefully analyzes the prompt and makes notes or comments about different targets for assessment
Focus	Knowledge: Grader identifies pieces of knowledge to be demonstrated	Reasoning: Grader focuses on the type of reasoning that is expected
Stage 2: Preparing for and Engaging in the Evaluation of Student Work		
Rubric	Implicit: Grader does not build an explicit rubric or system for point distribution	Explicit: Grader builds an explicit rubric which often includes distribution of points for target areas of assessment
Expected Answer	Unexpressed: Grader does not write or verbalize an expected answer	Generated: Grader generates an expected answer in paragraph form or as a list of key points
Grading System	Norm-referenced: Grader assigns points to individual student responses in reference to answers from other students	Criterion-referenced: Grader assigns points to individual student responses based on some criterion
Stage 3: Noticing in Student Work		
Stance	Literal: Grader does not make assumptions about student understanding beyond what is expressed in writing	Inferential: Grader builds inferences about student understanding that go beyond what is explicitly written
Lens	Prescriptive: Grader looks for key concepts and ideas as prescribed in a rubric or expected answer	Adaptive: Grader evaluates student work contextually and recognizes alternative ways of expressing ideas
Scope	Piecemeal: Grader evaluates one idea or sentence at a time without much attention to connections across the answer	Wholistic: Grader analyzes an entire answer before making a final evaluation
Stage 4: Responding to Student Work		
Interaction	Unexpressed: Grader does not make any visual marks or annotations on students' responses	Expressed: Grader highlights or writes notes on elements of student work
Intention	Reactive: Grader reacts to students' ideas by making marks on elements of student work that catch their attention	Evaluative: Grader makes marks on student work to highlight correct or incorrect statements
Feedback	Direct: Grader provides feedback/guidance in the form of statements	Indirect: Grader provides feedback/guidance in the form of questions

codes and discussion of discrepancies until they were resolved. A third researcher participated in a second phase of the analysis where different dimensions and categories were re-labeled and classified to better represent and organize the findings. The grading of each student response was coded for all eleven dimensions in our analytical framework, assigning an instructor's behavior to one of the two contrasting approaches within each dimension. Although in some cases an instructor could exhibit a behavior that included aspects of both categories (e.g., focusing both on knowledge and reasoning), the code assigned corresponded to the dominant behavior in each grading instance. Thus, for coding purposes the different categories in Table 1 were treated as dichotomous.

Major Findings

The core results of our study are presented in the following subsections, where we first briefly describe the variation in grading of the two selected students' responses to the assessment prompt used during the interview, and then characterize the observed dimensions of variation in participating instructors' approaches to evaluation and grading of student work.

Variation In Grading

Descriptive statistical analysis of the grades assigned by instructors to each of the two student responses provided during the interview elicited a wide distribution of grades for both answers as shown in Figure 3 (a) (Student Response A; $M = 3.29$, $SD = 1.80$) and (b) (Student Response B; $M = 3.70$; $SD = 1.64$). Students' responses were graded on a 0 to 6 points scale as described in the Methods section and instructors decided how to assign or distribute those points based on their own evaluation decisions. Large variations were observed in the grades assigned by both FIs and GIs. These findings prompted us to qualitatively analyze the data seeking to identify sources of variation during the evaluation and grading of student answers.

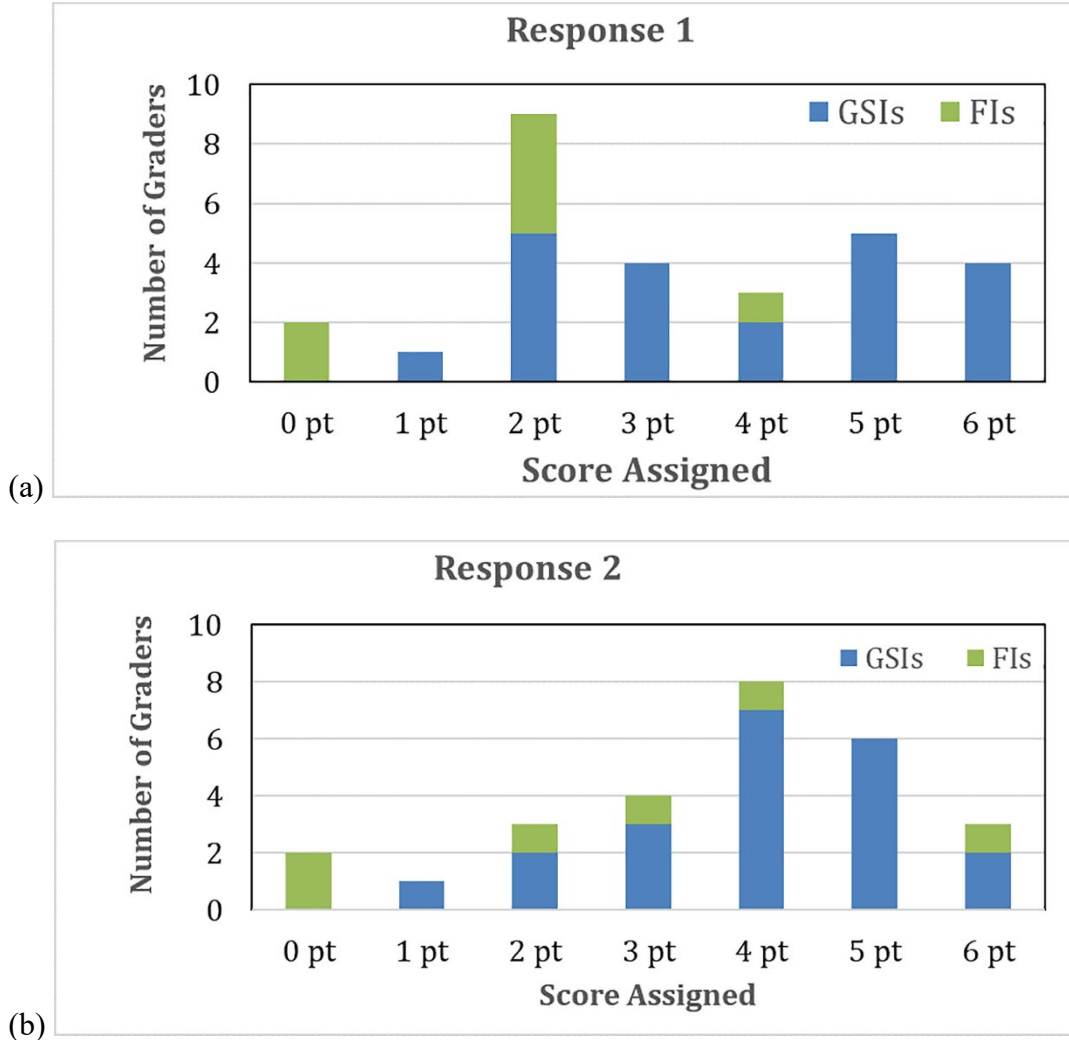


Figure 3 (a) and (b): Distribution of grades for (a) Student Response 1 and (b) Student Response 2 assigned by graduate student instructors (GSIs) and faculty instructors (FIs).

Dimensions Of Variation

During the qualitative analysis, we recognized differences in the approaches that participating instructors followed in various stages of the evaluation process that affected assigned grades. These stages included: a) Reading and Interpreting the Prompt, b) Preparing for and Engaging in the Evaluation of Student Work, c) Noticing in Student Work, and d) Responding to

Student Work. Within each of these stages we identified dimensions of analysis along which instructors' approaches to evaluation and grading differed (see Table 1). In each of these dimensions, we identified contrasting behaviors that affected the scoring of student work. We do not think of these contrasting categories as fixed and confined by hard boundaries. Nevertheless, we believe that their distinction is helpful in making sense of the observed variability in grading.

Approaches followed and decisions made by instructors at any given stage in the grading process often affected their behaviors and choices in subsequent stages. For example, the way the prompt was read and interpreted affected the selection of specific understandings to be evaluated, which informed decisions on preparing to assess (e.g., rubric design) and influenced what was noticed in students' answers. In the following paragraphs we describe the major dimensions of analysis, and associated contrasting categories, identified at each stage in the evaluation and grading process.

A) Reading and Interpreting The Prompt

This first stage of the evaluation and grading process refers to the instructor reading and interpreting the assessment prompt. Instructors in this study approached this task in different ways. The analysis of their behaviors and products during this stage led us to the identification of two dimensions along which differences were observed: Depth and Focus.

Depth: This dimension characterizes a grader's depth of analysis of the assessment prompt. Some instructors in our study ($n = 8$) read the prompt quickly, skimming the question for the target concepts of ideas to be evaluated. Instructors who took this approach typically focused on a single evaluation target in the prompt. For example, the question in Box 1 asks students to use the particulate model to explain how water evaporates and why the assumption about particles having a distribution of speed is essential to explaining evaporation. Instructors who skimmed the question

often paid attention and focused on only one portion of the evaluation, such as “explain how water evaporates.”

On the other hand, most graders in our sample ($n = 20$) closely read the prompt, identified different targets for evaluation, and made annotations to highlight these targets. These instructors often commented on the multifaceted nature of the prompt and identified two or three evaluation targets. For example, the instructor FI Addison identified three distinct parts in the question: “*what is distribution,*” “*explain how,*” and “*why is it essential.*”

Focus: This dimension characterizes what instructors express as their focus of attention while reading and interpreting the prompt. Some instructors ($n = 11$) seemed to focus more on the knowledge that students were expected to demonstrate in their answers. These instructors often listed key phrases in their notes that they expected to see in students' answers. For example, GSI Morgan listed: ‘*Avg kinetic energy, particles moving faster or slower, faster particles escaping*’. Nevertheless, a majority of instructors ($n = 17$) expressed interest in evaluating students' reasoning by including phrasing like ‘*why*’, ‘*how*’, ‘*explain*’, or ‘*connect*’ in their notes.

B) Preparing For and Engaging In The Evaluation Of Student Work

This second stage in the evaluation and grading process refers to the steps that instructors take before students' responses are analyzed. These steps affect how they engage in the evaluation of student work. During the interview, instructors were first asked to read the prompt and then were given time to prepare for the evaluation (i.e., they were asked to ‘do anything you would do to prepare to grade an exam’). Differences in approaches were observed along three major dimensions: Rubric, Expected Answer, and Grading Scheme.

Rubric: Of the 28 instructors that agreed to evaluate students' answers to the question in Box 1, fifteen of them generated an explicit rubric before analyzing students' answers. These

rubrics included explicit information describing what an instructor was looking for in a student response. The level of elaboration of different rubric components varied considerably among participating instructors as illustrated in Figure 4. Some of these rubrics included assigned point values for different elements (Figure 4a, 4b, 4c, and 4e), while others did not (Figure 4f). Some instructors did not write a rubric but expressed it verbally, as was the case with GSI Oakley, who said *"I thought that assigning two points to the particulate model and the four points of the distribution of speed was a good way to start with the question."* Instructors were judged to work with implicit rubrics ($n = 13$) when they assigned different point values to students' answers without ever explicitly stating the criteria they used in their evaluations.

Expected Answer: A few participating instructors ($n = 9$) generated a version of a correct or complete response in preparation for the evaluation of student work. As illustrated by Figure 4d, the answer could look like a student-constructed response in paragraph form, but more frequently was expressed as a bulleted list of ideas as shown in Figures 4a and 4b. Some of these instructors generated expected answers that were not completely correct. Most instructors in our sample ($n = 19$) did not write (see Figure 4c) or verbalize an expected answer; their behavior along this dimension was labelled as "unexpressed." The construction of a rubric and an expected answer were sometimes related, but it was possible to have a rubric without a clear explicit answer (Figures 4c, 4e, 4f), an answer without a rubric (Figure 4d), or an expected answer that served as a rubric (Figures 4a, 4b).

Grading System: Most instructors in our sample ($n = 21$) assigned grades using a criterion-referenced system, while a few ($n = 7$) applied a norm-referenced system. Instructors who followed a criterion-referenced approach often utilized rubrics and/or expected answers to guide their

a)

I20020904a

Question 1: Evp

Pools constantly need to be replenished with water due to evaporation. One of the assumptions of the particulate model of matter is that particles exhibit a distribution of speed. Using the particulate model, explain how water evaporates and why the assumption about particles having a distribution of speed is essential to explaining evaporation?

6 pts total

pts: 2) T is a measure of ~~PE~~ KE_{ave.}

2) if all particles had same KE then no particles could break IMF in liquid phase and go to the gas phase

3) distribution of speeds allows for fast part. to break IMFs (high KE) and go into gas phase without reaching b.p.

c)

A20030801a

Particle model = 2

How = 2

Why = 2 emphasis on avg KE

d)

A03030512a

Because molecules move at a distribution of speeds, some molecules, which are moving faster than others, have the opportunity to escape the liquid phase and enter the gas phase, thus causing evaporation to occur.

b)

I01040112a

Question 1: Evp

Pools constantly need to be replenished with water due to evaporation. One of the assumptions of the particulate model of matter is that particles exhibit a distribution of speed. Using the particulate model, explain how water evaporates and why the assumption about particles having a distribution of speed is essential to explaining evaporation?

6 pts

What distribution
Not all particles travel at THE SAME SPEED or have the same Kinetic Energy: They have Average speeds and Average Kinetic Energies.

2 pts Explain how
WATER is held together by potential energy and to escape the kinetic energy must exceed that.

Why is it essential

Some WATER particles have enough K.E to escape AT TEMP'S Lower than Boiling TEMP. This LEADS to Evaporation since AT Any temp, some particles will escape which is critical to explain Evaporation. Without this distribution, particles would have the same K.E. and would only leave AT the Boiling Temp.

e)

A03020918a

Avg Kinetic energy = 1

particles moving faster or slower = 1

faster particles escaping = 4

f)

A02182100a

def. of particulate model
distribution of speed

↳ connection to evaporation

↳ interaction of molecules

Figure 4: Examples of explicit rubrics and expected answers built by interviewed instructors

assignment of grades. Other instructors approached grading using a norm-referenced approach, assigning points after reviewing several student responses or *"making decisions more globally after looking at the entire body of questions"* (FI Shiloh). Instructors in our study evaluated only two student responses and they were presented in the same order every time. Graders that followed a norm-referenced grading system commonly justified their grade for a student's answer by comparing it with the response from a different student as illustrated by the following interview excerpts:

"Our first student, they either don't really – they have heard of it [the particulate model], but they don't understand how it actually works... [The second student] had a very good understanding of the particulate model." (GSI Finley)

"I felt that student 1 [had a better understanding], just because they talked more about all the factors of the particulate model, where this person just said – I don't know. They (the second student) didn't really talk about anything... I didn't think that they understood the particulate model." (GSI Nicky)

C) Noticing In Student Work

This stage refers to the instructor's attention to and interpretation of the student work. Differences in what participating instructors noticed in student work and what they inferred from it were detected in three major dimensions: Stance, Lens, and Scope. A few instructors' noticing approach varied from one student response to the other. Thus, in the following paragraphs we use number of evaluation instances ($n_e = 56$) rather than number of instructors ($n = 28$) to indicate relative frequencies of different behaviors. These variations in noticing approach in the evaluation of different responses by the same instructor were, however, infrequent.

Stance: In some instances ($n_e = 23$), instructors evaluated students' responses based literally on what was explicitly written, without relying on inferences about student understanding by 'reading between the lines' in making grading decisions. The following interview excerpt illustrates this noticing stance:

"That's an assumption that they're making, so they're not totally wrong, but it doesn't mean that they're actually right and should always provide that in their answer. So, I don't want to completely dock them being not wrong, but only want to give credit for things that are completely right." (GSI Dylan)

In this case, the instructor recognized that the answer provided by the student might have been indicative of additional understanding but decided to provide credit only for what was stated and was "completely right."

In a larger number of instances in our sample ($n_e = 33$), instructors engaged in interpretation of students' answers by building inferences that went beyond what was explicitly stated. For example, the second student response (Figure 2b) did not include any explicit reference to the word "energy." However, several instructors assumed this student demonstrated understanding of energy-related concepts as illustrated by this excerpt:

"They showed that they had an understanding about what the particulate model of matter is and how the kinetic energy and potential energy relates to evaporation." (GSI Georgie)

In general, instructors who adopted a literal stance tended to award fewer points for the same student response than instructors who took an inferential stance. This behavior is in line with the findings from Mutambuki and Fynewever,⁵ which indicate that chemistry faculty are often reluctant to score students poorly if some evidence points to the possibility that they actually understand.

Lens: In some instances ($n_e = 23$), instructors approached the evaluation of student work in rather prescriptive ways looking for specific ideas or words in a student's answer often guided by the instructor's rubric or expected correct response. Consider this example in the evaluation of student response 2 in Figure 2b:

"It felt like they really kind of just talked about what a distribution of speeds meant and didn't quite talk about energy or anything or how that related or why that mattered. So again, gave them points ... although they didn't bring energy into it." (GSI Pat)

In this case, the instructor expected students to refer to the concept of energy in their answers and deducted a significant number of points (4 points) for failing to do so.

Nevertheless, in a majority of instances ($n_e = 33$) instructors in our sample exhibited an adaptive approach in the evaluation of student responses, valuing different ways of building an explanation of expressing understanding. While many of the instructors who adopted an adaptive lens did not create a rubric or expected answer, generating these elements did not imply the adoption of a prescriptive approach. Some instructors modified their rubrics while grading student responses to incorporate alternative ways of expressing ideas.

Scope: In some instances ($n_e = 27$), instructors analyzed students' responses in a piecemeal fashion, evaluating one idea or phrase at a time. These instructors often read a sentence, marked it correct or incorrect, and then evaluated the next sentence. This limited scope in the evaluation of student work is summarized in this interview excerpt:

"So once I find out some important things that's indicating the student understands what the question is asking for, then I just leave a checkmark or highlight that sentence. So based on those highlighted portions, I decide how many points are recorded for the student's answer." (GSI Ollie)

In a similar number of instances ($n_e = 29$), instructors adopted a more wholistic scope in the evaluation of students' responses, paying attention to the connectivity and integration of ideas across the entire answer. As described in the following interview excerpt, these instructors tended to completely read a full answer before focusing on key points:

"I usually read through their response first just to see if they have all the major points. And then I grade their major points based on what they actually wrote." (GSI Finley)

D) Responding To Student Work

During the evaluation and grading of student work, instructors interacted with students' answers in different ways. Expressed interactions were indicative of what instructors noticed and interpreted in students' answers. Instructors were not necessarily consistent in their interactions when evaluating different student responses, so we use number of instances again to provide information on relative frequencies for different types of behaviors. In this stage, we considered a single dimension of analysis, "Interaction," as indicative of whether the instructor explicitly and visually interacted with student work or not. When explicit interactions occurred, they differed along two subdimensions: Intention and Feedback.

Interaction: In some instances ($n_e = 18$), instructors read student responses, evaluated them, and assigned a grade without interacting in any visible manner with the written answer. In these situations (unexpressed interaction), instructors did not provide any visual cues for how or why they assigned a particular grade. When asked about this behavior, several instructors indicated that their past experiences having to grade many student responses at a time had conditioned them to just assign a grade and move on: *"I would have 500 more to do after this"* (FI Addison).

In most instances ($n_e = 38$), participating instructors visually interacted with student work by making different types of marks and annotations: underlining words or phrases, adding check

marks or crosses, making comments, and posing questions. Differences among these marks and annotations were characterized along two subdimensions: Intention and Feedback. An example of the different types of visual marks made by instructors while grading is shown in Figure 5.

Intention: In the cases where instructors visually made marks on student work, half of the instances ($n_e = 17$) involved the underlining of words and phrases that got the attention of the graders for a variety of reasons:

"it actually turned out to be a meaningless statement" (FI Blair)

"that's one of the things that they needed to address" (FI Jamie)

"it just isn't clear as to what they're talking about" (GSI Morgan)

"So, I can quickly go in [when students come to me asking about their grade] and say, oh, let's see. What did I underline? Let's talk about that" (GSI Micah)

"Those are things that are good" (GSI Taylor)

"Things that I feel support their answer" (GSI Jackie)

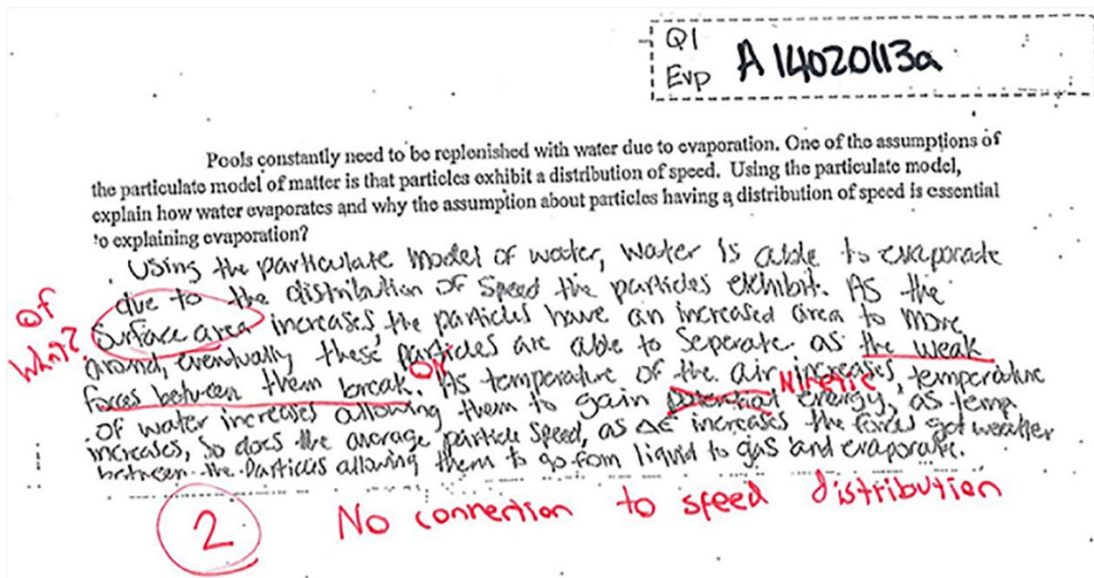


Figure 5: Different types of visual marks made by instructors while grading

As these interview excerpts illustrate, some instructors reacted to and underlined different elements of student work that they found indicative of comprehension or work completion.

In other instances ($n_e = 17$), instructors' markings were accompanied by some type of evaluative notation (e.g., check mark, cross mark, question mark) that indicated whether the marked element was correct, incorrect, or confusing. In all cases, marks made highlighted aspects of student work that seemed to influence an instructor's evaluation and grading of student work.

Feedback: In some cases, instructors included written feedback as part of their evaluation. When this occurred, the feedback could be direct ($n_e = 13$) in the form of statements that highlighted missing aspects of a response ("*need to better explain how it evaporates,*" GSI Dylan), or corrected or added to students' ideas. These comments were often added to justify the reduction of points. In fewer instances ($n_e = 7$), participating instructors provided indirect feedback in the form of questions that could also be used to highlight missing elements in an answer or prompt student reflection. Both direct and indirect feedback varied in length from one-word statements or questions to more extensive explanations or queries.

Limitations

This qualitative study involved a small number of faculty and graduate student instructors within a specialized area (general chemistry) and our analysis focused on their approaches to the evaluation and grading of a single exam question and two student answers. Consequently, one should be cautious in the generalization of any of our findings. For example, we did not observe major qualitative differences in the approaches to evaluation and grading followed by FIs and GSIs, or by instructors with fewer or more years of experience. This result could be due to our small sample size or to the type of question selected for the study. This question was representative of conceptual questions used in general chemistry exams at our institution, but different types of

questions may trigger different evaluation and grading behaviors. We cannot speak to the emotional or social experiences of the instructors during the interview that might have affected their observed decisions and preferences. Participating instructors were asked to grade a small number of students' responses and their behaviors may be different when confronted with many more answers to evaluate and grade as part of their typical duties.

Discussion

Our qualitative analysis revealed several dimensions along which instructors' approaches to the evaluation and grading of student work varied. These variations affected decisions about how many points to award or subtract from the maximum possible. Figure 6 summarizes the distribution of behaviors for the instructors in our study along each of these dimensions. In this figure, contrasting categories in each dimension are placed to the left or right in alignment with

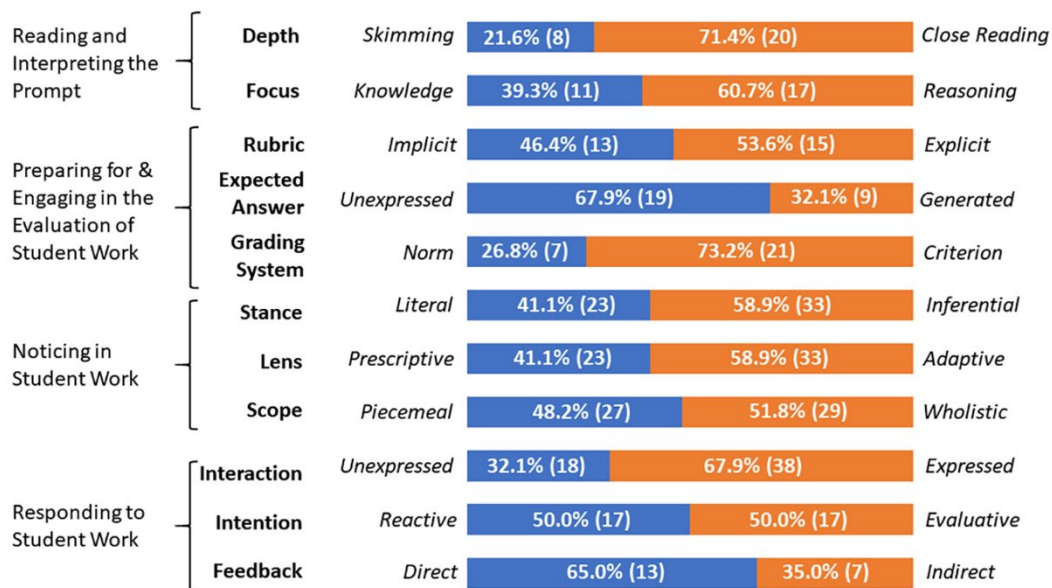


Figure 6: Distribution of behaviors along dimensions of analysis in instructor's approaches to grading in our study. Frequencies are expressed in terms of number of instructors ($n = 28$) for the first five dimensions and in terms of instances of evaluation ($n_e = 56$) for the others.

their description in Table 1. The placement of a category in either side is arbitrary. None of the instructors in our sample exhibited the same behaviors across all dimensions, and differences in approaches to grading resulted in variations in assigned grades for the same student response. Large variations in grading were observed in cases where grading approaches varied in a few or many dimensions. In some cases, instructors with quite dissimilar approaches to grading ended assigning similar grades to a student response but based on different reasons and decisions.

Consider the following example. Two instructors in our sample, GSI Oakley and GSI Riley, were aligned in their approaches to grading on most dimensions except for "rubric" and "grading system." They assigned 2/6 and a 6/6 points, respectively, to student response 1 in Figure 2a. Both instructors approached the evaluation of student work by skimming the prompt and without generating an expected answer. Each instructor focused on reasoning, took an inferential stance to noticing, and adopted an adaptive lens while evaluating the response holistically. However, instructor GSI Oakley used a criterion-referenced grading system and ended assigning 2/6 points, citing a verbal rubric which assigned "*two points to the particulate model and the four points of the distribution of speed [as] a good kind of rough way to start with the question.*" The second instructor, GSI Riley, gave 6/6 points to the same response, justifying their grade in reference to the second student response, which this grader judged of lesser quality than response 1 for having "*some mistakes*" and because "*they don't say quite as much as the one above.*"

The approach to evaluation and grading followed by the instructors in the example above was quite different from that of instructor FI Logan who closely read the prompt, focused on knowledge, wrote an explicit rubric with a generated answer, was literal and prescriptive while grading in a piecemeal fashion, and applied a criterion-referenced grading system. This instructor assigned a grade of 0/6 points as the approach followed in the evaluation of student work led the

instructor to find no value on the provided response: *"I felt like they weren't answering what the question was asking."* Despite major differences in approaches to evaluation and grading between this instructor and instructor GSI Oakley described above, the difference in assigned grades (2 points difference) is smaller than that between the latter instructor and GSI Riley (4 points difference) whose approaches differed in fewer dimensions.

These examples illustrate how assigned grades seemed to emerge from a complex interaction between a variety of explicit and implicit decisions made and preferences manifested along several dimensions. These decisions and preferences varied across instructors but also within the evaluation and grading practices of a single individual. Instructors in our sample frequently evaluated and graded the two student responses following the same approach, but occasionally their decisions varied from one student response to the other. These variations affected the assigned grades and their justification. For example, instructor FI Shiloh approached the evaluation of student work by skimming the prompt and without generating an explicit rubric or expected answer. In analyzing the provided student responses, this instructor took an inferential, adaptive, and wholistic approach to noticing. While evaluating student response 1, this instructor seemed to focus on student reasoning and use a criterion-referenced grading scheme as evidenced in this interview excerpt:

"This student really was not answering the question, was bringing in things that were not relevant... My two points come from recognizing that there is some understanding of the relationship between temperature and average speed, but that's it... There are fragments of knowledge... [but] this student has a lot of confusion." (FI Shiloh)

Based on this analysis, the instructor assigned 2 out of 6 points as a grade. On the other hand, while grading student response 2 the same instructor recognized several flaws in student

reasoning, but assigned 4 to 5 out of 6 based on what seemed to be a norm-referenced approach to grading:

"I was surprised by the connection of greater speed, more configuration... but overall, what the student was saying was mostly correct, although I never saw the student making a connection between distribution of speed and answering the question... This is one of the questions I would go back again... after looking at the entire body of questions and make my [grading] decisions more globally." (FI Shiloh)

Comparison between the two available answers seemed to have led this instructor to assign more points to the second response not necessarily based on its own merits, but on the extent to which one answer was perceived as better than the other.

Instructors in our study made a variety of in-the-moment decisions and manifested preferences in the analysis and evaluation of student work that affected the grades that were assigned. As suggested by previous work on the grading practices of college chemistry and physics,⁴⁻⁷ those decisions and preferences are likely influenced by personal beliefs about teaching and learning. Nevertheless, our results suggest that they are also affected by the nature of the question prompt, the decisions that are made before engaging in the evaluation, the collection of answers that are evaluated, and the specific aspects of students' answers that call the graders' attention. These findings indicate that the evaluation and grading of student responses is a complex task for which many instructors might not be well prepared and is likely to require sustained critical reflection to transform beliefs and practices.

Implications

The results of our study suggest that faculty and graduate student instructors would benefit from carefully reflecting on the different dimensions of variability in the evaluation and grading

of student work elicited in our investigation. Instructors are likely not aware of the many decisions they make and the implicit preferences that seem to guide their evaluation of student work. Professional development that engages instructors in metacognitive reflection about the potential sources of variation in the evaluation and grading of student responses to summative assessment prompts might help increase inter- and intra-rater reliability among graders. Conversations among instructors at our institution prior to and during the grading of student work typically focus on clarifications on the content of a provided rubric, and on discussions about how to interpret confusing student answers. These conversations may help align approaches to the evaluation and grading in some dimensions but not others and are not purposefully oriented to induce alignment in a reflective manner. The multidimensional nature of the evaluation as characterized in our study suggests that more structured and sustained opportunities for critical reflection might be needed for instructors to become aware of their approaches to evaluation and grading and be able to control and change them if needed.

Reviews on the pedagogical preparation of GSIs indicate that attention to issues related to assessment, evaluation, and grading is often minimal (Gardner & Jones, 2011), although these instructors value training that creates opportunities for them to discuss and reflect on issues of subjectivity and fairness in the evaluation of student work (Marbach-Ad et al., 2012). In regards to the professional development of college faculty, recent efforts have been mostly directed at changes in instructional strategies, somewhat on curricular decisions, but much less on assessment and evaluation beliefs and practices (Henderson et al., 2011). This lack of attention to assessment, evaluation, and grading practices is problematic as the questions we pose and the evaluations we make of student work make explicit the type of understandings we value and likely guide students' decision on what and how to study. Our findings clearly point to the need to engage in deeper

discussions and reflections on how we look at and judge student work not only to convey a clear stance on student learning but to ensure the fair and equitable treatment of all students.

Conclusions

Our study revealed eleven dimensions of variation in the approaches that instructors took during the evaluation and grading of students' answers to free response questions. These dimensions, summarized in Table 1, help characterize different instructors' behaviors during four major stages in the evaluation of student work. Within each dimension, we identified contrasting approaches that affect the decisions instructors make about the quality of a student response and have an impact on the points an answer gets assigned. None of the instructors who participated in our study exhibited the same behaviors in all dimensions when evaluating and grading the same samples of student work. Our findings suggest that professional development for instructors should help them reflect on the variety of factors that may affect their evaluation of student responses. Even if a common rubric is developed and discussed, our results indicate that instructors may approach evaluation differently in several other dimensions.

CHAPTER 4: VARIATION IN CHEMISTRY INSTRUCTORS’ EVALUATION OF STUDENT WRITTEN RESPONSES AND ITS IMPACT ON GRADING²

In the second part of our study, the analytical framework summarized in Table 1 in Chapter 3 was used to characterize chemistry instructors’ approaches to the evaluation of a diverse set of conceptual questions in general chemistry exams and the consistency of their behaviors across different questions and student responses. We were also interested in characterizing how individual instructor characteristics and approaches to the evaluation affected assigned grades. This part of our work was guided by the following research questions:

1. What main approaches across different dimensions of variation in the evaluation and grading of student work do chemistry instructors more commonly follow?
2. How consistent are chemistry instructors in their approach to the evaluation and grading of student work across different dimensions of variation?
3. What individual variables and dimensions of variation correlate to assigned grades?

² The main content of this chapter is reproduced (adapted) by permission of The Royal Society of Chemistry from (Herridge et al., n.d.).

Methods

Setting and Participants

This project was conducted at a large research-intensive public university in the southwest US. Study participants were faculty instructors (FIs) and graduate student instructors (GSIs) teaching a 100-level general chemistry curriculum that emphasizes the development of students' chemical thinking ($n = 29$; 7 FIs, 22 GSIs). Detailed demographic information for the participants is presented in Table 2. All participants had taught at least one full semester of the course; FIs had taught the lecture section and the GSIs had taught at least one associated laboratory section and attended lecture. In these courses, FIs designed the summative exams and GSIs graded them. All students enrolled in the course took common exams, and the FIs rotated exam writing responsibility. GSIs were expected to use a rubric (with a clear answer) generated by the FIs in grading the free response questions for each common exam. FIs and GSIs met after each exam to discuss how to use the given rubric and practice applying it to student responses. From now on,

Table 2: Participant Information

Role	<i>N</i>	Teaching Experience (Range in years)
Faculty Instructor (FI)	7	2-30
Graduate Student Instructor (GSI)	22	1-5
<hr/>		
Research Area	<i>N</i>	
Analytical	4	1-9
Biochemistry	8	1
Inorganic	4	1-19
Physical	9	1-6
Other	4	1-30
<hr/>		
Gender	<i>N</i>	
Male	17	1-30
Female	12	1-10

we refer to all study participants as instructors. This research project was approved by the IRB at our institution and all subjects consented to participate in the study.

Exam Questions & Student Responses

The findings reported in this paper emerged from the analysis of study participants' evaluation and grading of students' original answers to six free response questions from general chemistry exams given in 2016 or 2017 (see Box 2). Students in the general chemistry courses at our institution complete several midterm exams which include free response questions whose evaluation was the target of our investigation. Four of the selected questions corresponded to exams from the first semester of the course (GC I) and two questions came from exams from the second semester (GC II). These questions covered a wide range of topics in the curriculum. Five of the questions were worth six points, and the last question was worth eight points over a total of 100 per exam. These point values were retained when participants were asked to evaluate and grade student responses in our study.

Two original student responses per question were selected for evaluation and grading by our study participants. The student responses used in this study were real students' responses to actual exams given in prior years and were not artificially created. These responses were deidentified before use and each response was from a different student. The responses were unedited and retained handwriting, spacing, and notes exactly as written by the student during the exam. The researchers selected responses that were typical of those provided by students in the course and were expected to be judged as partially correct rather than completely right or wrong. Examples of the selected student responses can be found in Appendix D.

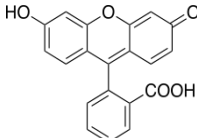
Box 2: All Question Prompts

Q1- Particulate model of matter (PMM) GC I Fall 2016

Pools constantly need to be replenished with water due to evaporation. One of the assumptions of the particulate model of matter is that particles exhibit a distribution of speed. Using the particulate model, explain how water evaporates and why the assumption about particles having a distribution of speed is essential to explaining evaporation.

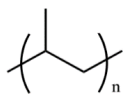
Q2 – Resonance Stabilization (RS) GC I Fall 2016

Fluorescein exhibits a lot of resonance stabilization. Build an argument for why the presence of resonance is stabilizing based on the potential and kinetic energy of the electrons.

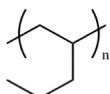


Q3 – Structure-Property Relationships (SPR) GC I Fall 2016

Other polymers commonly used in tattooing include the plastic cups used to hold water to rinse ink off the needles, and plastic wrap used to wrap the chair and table, as well as the skin after the tattoo is complete for healing and protection. Shown below are examples of each of these polymers. Justify why the polymeric unit corresponds to the specific type of plastic. Assume the chain length for both polymers is the same.



Plastic cup



Plastic wrap

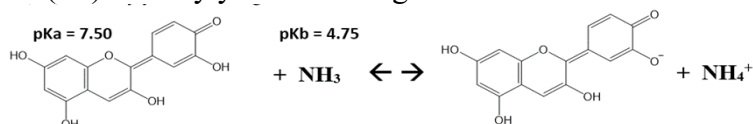
Q4 – Thermodynamic Stability (TS) GC I Fall 2017

Most proteins adopt a single unique structure when “folded” and many random structures when “unfolded”. Proteins can be unfolded (i.e. denatured) at high temperatures (ex: cooked eggs). Draw a PEC diagram [Potential Energy-Number of Configurations diagram] representing a “folded” protein (F) and an “unfolded” protein (UF). Explain how the two states of the protein differ (i.e. explain your logic in constructing this PEC diagram).



Q5 – Reaction Directionality (RD) GC II Spring 2017

Considering the reaction below, would you predict it to be reactant favoured (RF) or product favoured (PF)? Justify your reasoning.



Q6 – Perturbations to Chemical Equilibrium (PCE) GC II Spring 2017

This equilibrium is established in the buffer from question 18: $B + H_2O \leftrightarrow BH^+ + OH^-$. Build a.) thermodynamic and b.) kinetic arguments to explain how the equilibrium would shift if OH^- is added to the solution. Justify your reasoning. Note: include a Q vs. time plot in your answer. (This question provided two boxes, one labelled “Thermodynamic argument” and one labelled “Kinetic argument” for students to write their answers in.

Data Collection

The first author of this paper met with each of the 29 instructors for an individual, video-recorded, semi-structured interview which lasted from half an hour to three hours. During the interview, participants were asked to evaluate and grade the provided student responses to the selected questions as if it were an exam setting. For each question and its associated student responses, the instructor was first asked to read the question, prepare to evaluate, evaluate the students' answers, and then assign a score. All documents were provided as hard copies. The first page was the exam question written at the top of an otherwise blank sheet of paper. The instructor was asked to prepare as they would do in an exam grading setting. Then, the instructor was given a page with the student responses selected for that question. After completing the evaluation and grading of a question, both pages (preparation and evaluation pages) were collected from the instructor and the next question was provided following the same procedure. Once the evaluation and grading were completed, the interviewer explored the participants' rationale for each of their assigned grades. This interaction occurred after the evaluation of all questions had been completed to reduce the influence of the interviewer on the participants' evaluation decisions. All instructors were given the option to skip the evaluation of any question if they were unfamiliar with the subject matter. A total of 268 evaluations of individual students' responses were collected and analysed for this study.

Data Analysis

Participants' assigned grades to the selected students' responses were analysed using descriptive statistical methods to characterise grading variability in our sample. Subsequently, interview recordings were transcribed, and all data (e.g., observation notes, participants'

annotations) were analysed using a standard iterative approach in the analysis of qualitative written data (Creswell, 2012; Miles & Huberman, 1994).

Participants' approaches to evaluation and grading were analysed using the framework developed in previous work (Herridge & Talanquer, 2020) and summarised in a prior section. The unit of analysis was the evaluation of a single student's response coded in each of the eleven dimensions in the framework. Each of these codes was assigned based on the predominant approach followed in the evaluation of a student response. To ensure interrater reliability in the qualitative analysis, a second researcher independently analysed and coded 19% of the collected data. The two researchers then met and discussed differences until agreement was reached on all eleven categories of analysis for each student response in the sample. Then the remaining samples were analysed by one of the researchers. Any samples in the remaining set that presented difficulty in assigning a code were discussed and confirmed by both researchers.

During the data analysis, decisions about applying each code were made based on the diverse products collected during the interview. Preparatory work, notes, rubrics, marking, and assigned grades were the main source of codes for the Rubric, Expected Answer, Interaction, Marking/Intention, and Feedback dimensions. The video with timestamps was used to inform decisions about applying codes to the dimensions of Depth, Grading System, and Scope. The transcripts were the main source of information for applying codes to the dimensions of Focus, Stance, and Lens. All three data sources were used together to confirm the overall approaches to evaluation followed by each participant.

Once all samples were coded, descriptive statistics was used to identify patterns in the data. This quantitative analysis included the determination of frequencies of each code overall and by participant, average scores assigned by participant and by evaluation approach, frequencies of

shifts in evaluation approach from one question to another, and correlations between approaches across different dimensions of analysis.

Statistical modelling was used to evaluate the correlative relationships between the outcome variable of assigned grades and selected independent variables, including control and grader variables, as well as the eleven dimensions of the analytical framework described in this paper. The need for multilevel modelling was confirmed by the intraclass correlation coefficient ($\rho = 0.128$) and a likelihood ratio test (LRT) that showed a multilevel model with graders as the subunit had a better fit than a simpler, fixed-intercept model with $\chi^2(1) = 12.2, p = .005$ (Peugh, 2010; Raudenbush & Bryk, 2001). Assessment of assumptions of multilevel modelling indicated that the dependent variable was normally distributed with skew of -0.36 and kurtosis of -0.87, both within range of ± 1 to be considered as normal (Kim, 2013). Both grade residuals and grader zetas showed constant variance (homoscedasticity) and normality when assessed graphically by quantile-quantile (qq) plots and box plots.

Model building was conducted using the methods and with the results outlined in Tashiro et al., (2021). Improvement in model fit, shown by LRT, was used for the determination of statistical significance and inclusion of independent variables and independent variable interactions. Although all possible random slope effects were assessed, only the random intercept effect was statistically significant and therefore all effects discussed in this paper can assumed to be fixed effects. R statistical software was used for the multilevel modelling (Wickham & Henry, 2018).

The proportion of variance in assigned grades accounted for by the variable “student response” was calculated by taking the difference in the proportion of variance explained by models with and without this variable. The effect size measure Cohen’s f^2 was given by the ratio

of the proportion of variance in assigned grades accounted for by a variable to the unaccounted variance observed in assigned grades (Selya et al., 2012). Details for these calculations for categorical variables with and without interaction can be found in Tashiro *et al.*, (2021). Cohen's f^2 indicates effect size as small above .02, medium above .15, and large above .35 (Cohen, 1977).

Results

The main results of our analysis are presented in this section where we first qualitatively characterise variations in instructors' approaches to the evaluation of student work by identifying the main approaches and how consistent instructors are in their approaches (research questions 1 and 2), and then present the results of a statistical analysis of assigned grades to determine how instructor-related variables account for observed assigned grade variability (research question 3). Our goal is to illustrate how the applied analytical framework can be used to characterise instructors' evaluation behaviour, consistency in evaluation approaches, and the impact of these approaches on assigned grades.

Characterization of Variation in Evaluation Behaviour

Participating instructors differed in their approaches to the evaluation of student work in terms of 1) how they read and interpreted the question prompts, 2) how they prepared for and engaged in the evaluation of student work, 3) what they noticed in students' answers, and 4) how they responded to what students wrote. In the following subsections we characterise major differences in each of the dimensions included in the analytical framework in Table 1. A summary of the frequencies with which contrasting approaches along each dimension of analysis were followed by study participants across all graded samples (n = 268) is presented in Figure 7.

Approach By Individual Response

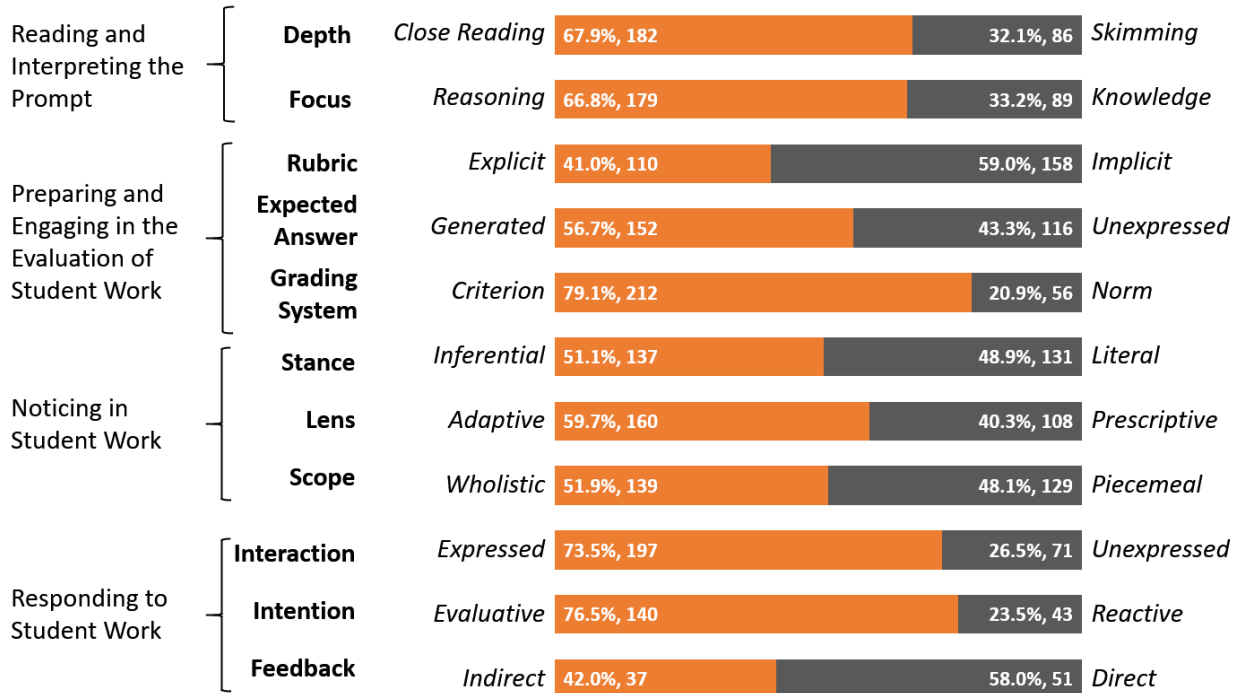


Figure 7: Distribution of behaviors in the evaluation of all samples of student work along each dimension of analysis in our study. Frequencies are expressed in terms of the number of graded samples ($n = 268$).

Instructors' Dominant Approaches

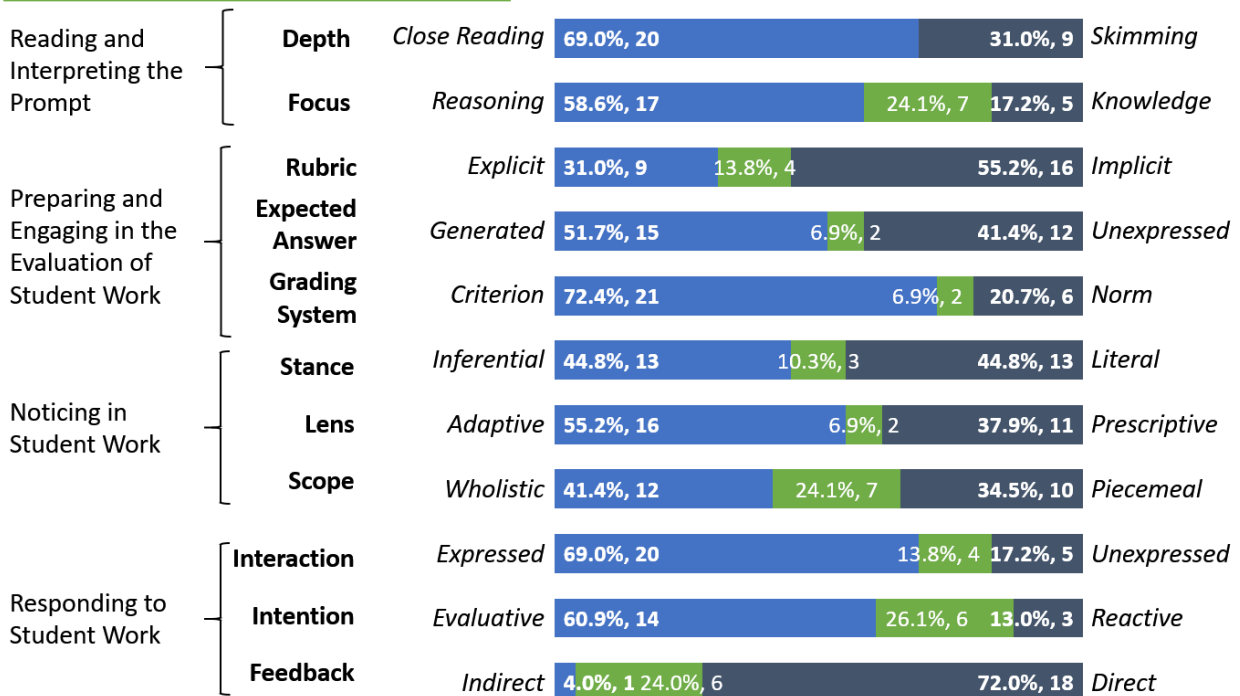


Figure 8: Dominant approaches in the evaluation of student work manifested by participating instructors in each dimension of analysis. Instructors who exhibited variable approaches are included in the middle bar in each dimension. Frequencies are expressed in terms of number of instructors ($n = 29$).

We further characterised the variation in evaluation behaviour through an analysis of consistency in instructors' approaches to the evaluation. This metric was built by analysing instructors' approaches across all samples graded. When instructors took a particular approach in at least 70% of the samples they evaluated, such an approach was considered as dominant in their behaviour. Any instructor who varied in their approach in over 30% of the instances was labelled "variable" in their evaluation approach in that dimension. This cut-off was chosen to only categorize those instructors as variable when they were oscillating between behaviours and not because a particular question caused a shift in the behavioural approach. For example, an experienced instructor recalled writing the first question prompt, and subsequently *skimmed* in the dimension of Depth. For all other questions, the instructor showed a dominant approach of *closely reading*. Analysis of which questions may cause these shifts and why is the subject of further research. A summary of the main findings related to variation in evaluation approach is presented in Figure 8.

Stage 1) Reading and interpreting the prompt. Differences in this area were observed along two dimensions: Depth and Focus. Depth refers to how closely an instructor reads the question being asked, while Focus characterises what instructors attend to while reading and interpreting the prompt.

Depth: This dimension includes two main approaches to the evaluation of student work, *close reading* and *skimming*. As shown in Figures 7 and 8, most instructors followed a close reading approach in the analysis of the question prompts (in 67.9% of all instances) and all of them were consistent in their behaviour. Of the 20 instructors who engaged in *close reading* with all prompts, 4 were FIs and 16 were GSIs. Eight instructors skimmed in all instances and one skimmed

in all but one question prompt. The 3 FIs who skimmed were the most experienced FIs, but the 5 of the 6 GSIs who skimmed were all in their first-year of teaching.

Instructors who engaged in *close reading* tended to spend more time analysing the prompt, identifying different parts of a question, highlighting words, or outlining expected requirements for a complete response. For example, GSI Georgie underlined portions of the prompt for every question to highlight different ideas or concepts thought to be targeted by the question. FI Jamie underlined and circled portions of each prompt in order to identify how to distribute total points among different parts of Expected Answers. When asked what their approach was for starting evaluations, GSI Robin stated, “*I was trying to see what the question wanted to ask specifically.*” *Close reading* was further evidenced by building Rubric items based on the prompt’s content and structure.

When *skimming* the prompt, instructors spent much less time looking at the question alone, often asking for the prompt and associated student responses at the same time. In these instances, instructors frequently indicated that they were already familiar with the question or remembered similar questions from prior grading experiences. In some cases, evidence of *skimming* the prompt was found in comments that revealed an instructor’s confusion during the evaluation of student work, such as FI Blair asking “*Why do they keep talking about wrapping the chair?*” while evaluating the responses to Q3, which specifically mentions chairs in relation to the flexibility properties of plastic wrap (see Box 2). Other instructors, such as FI Quinn, explicitly expressed a preference for *skimming* the prompt, looking at student answers, and more carefully reviewing the prompt only if they could not make sense of an answer independently. In these instances, instructors could be seen using their grading pen to mark a student response, and then going back to hover over the question prompt while the student response was analysed.

Focus: This dimension describes two categories of attention while analysing a question prompt and the associated students' answers, attention to *reasoning* and attention to *knowledge*. As shown in Figures 7 and 8, most instructors attended to student reasoning (in 66.8% of all graded samples) but some of them (24.7%) varied in their behaviour in this dimension. Behaviour in this dimension was characterised by analysing instructors' comments while reading the question prompt and as they evaluated students' answers.

Instructors who mostly focused on *reasoning* often referred to the importance of students' demonstrating understanding of an idea while reading a question prompt, and paid attention to the presence or lack of evidence related to sound reasoning or adequate understanding while evaluating students' answers. Some instructors, for example, used phrases that referred to the extent to which students integrated or connected ideas: "*a coherent train of thought*" (GSI Riley, Kendall), "*tied in*" (GSI Finley), "*integrated*" (FI Quinn) or "*a connection*" (GSIs Pat, Emerson, Alexis, Ollie, and FIs Quinn, and Shiloh), or identified areas in which students seemed to lack understanding: "*I was really looking for their understanding of resonance, which I didn't feel that they have based on their answer*" (GSI Oakley). Student understanding was often judged by the ability to use, apply, or connect different pieces of information to construct an argument or explanation, as illustrated by this interview excerpt from the analysis of student response 1 to question 5:

"OK, they recognize that that is weaker. So that's good. They can recognize that, but they cannot use it, or they didn't use it." (FI Shiloh)

In contrast, instructors who mostly attended to *knowledge* tended to rely on the presence of keywords or proper definitions to make evaluation judgments and grading decisions as illustrated by this justification for points awarded in grading student response 2 to question 2:

“They actually got the definition of resonance. They mentioned both potential energy and kinetic energy” (GSI Jackie)

Of the 17 instructors who focused on *reasoning*, 2 were FIs and 15 were GSIs. All 5 instructors who focused on *knowledge* predominantly were GSIs. We had 5 FIs and 2 GSIs who varied in their approach, with 2 of them having an almost even split on their Focus. For example, FI Addison stated that the student must *“link all that into a process of evaporation”* while evaluating question 1, but based evaluation judgments on *“looking for the word delocalization”* when grading question 2. Variability in the Focus of attention was also observed when grading different student responses for the same question. The presence of certain keywords or phrases seemed to trigger a positive response in some instructors who awarded points independently of whether that *knowledge* fragment was properly integrated in a student response. The following interview excerpt illustrates this behaviour:

“They used one sentence, one little phrase, that I like. I guarantee that their professor probably said to them. ‘Resonance increases the stability of a molecule.’ Great. They were paying attention for that one sentence. I went ahead and gave them a point for that.” (GSI Micah)

This type of variability seemed associated with some instructors’ disposition to *“find something to give points for”* as *explicitly* stated by 11 instructors, both GSIs and FIs, in our sample (GSIs Micah, Morgan, Oakley, Emerson, Ollie, Parker, and Sidney, and FIs Logan, Blair, Shiloh, and Elliot).

Stage 2) Preparing for and engaging in the evaluation of student work. Differences in this area were observed along three dimensions: Rubric, Expected Answer, and Grading System. The dimension of Rubric refers to whether the instructor *explicitly* stated the point value assigned

to different components of a question or an Expected Answer. Expected Answer refers to whether the instructor explicitly *expressed* an answer (either as a fully written response or as knowledge fragments) that they expected to see in students' answers. Grading System refers to whether the instructor evaluated each student response independently or in relation to others.

Rubric: Within this dimension, instructors could generate an *explicit* Rubric that outlined the assignment of points per section of a question prompt or answer or could apply an *implicit* Rubric that was not directly shared. Some instructors who generated and applied *explicit* rubrics built full expected responses, while others “*would try to... list the key words or key points that I’m looking for in the answer*” (GSI Lou). Examples of *explicit* rubrics for Q2 and Q3 built by the study participants are presented in Figure 9 - FI Logan and GSIs Parker and Georgie.

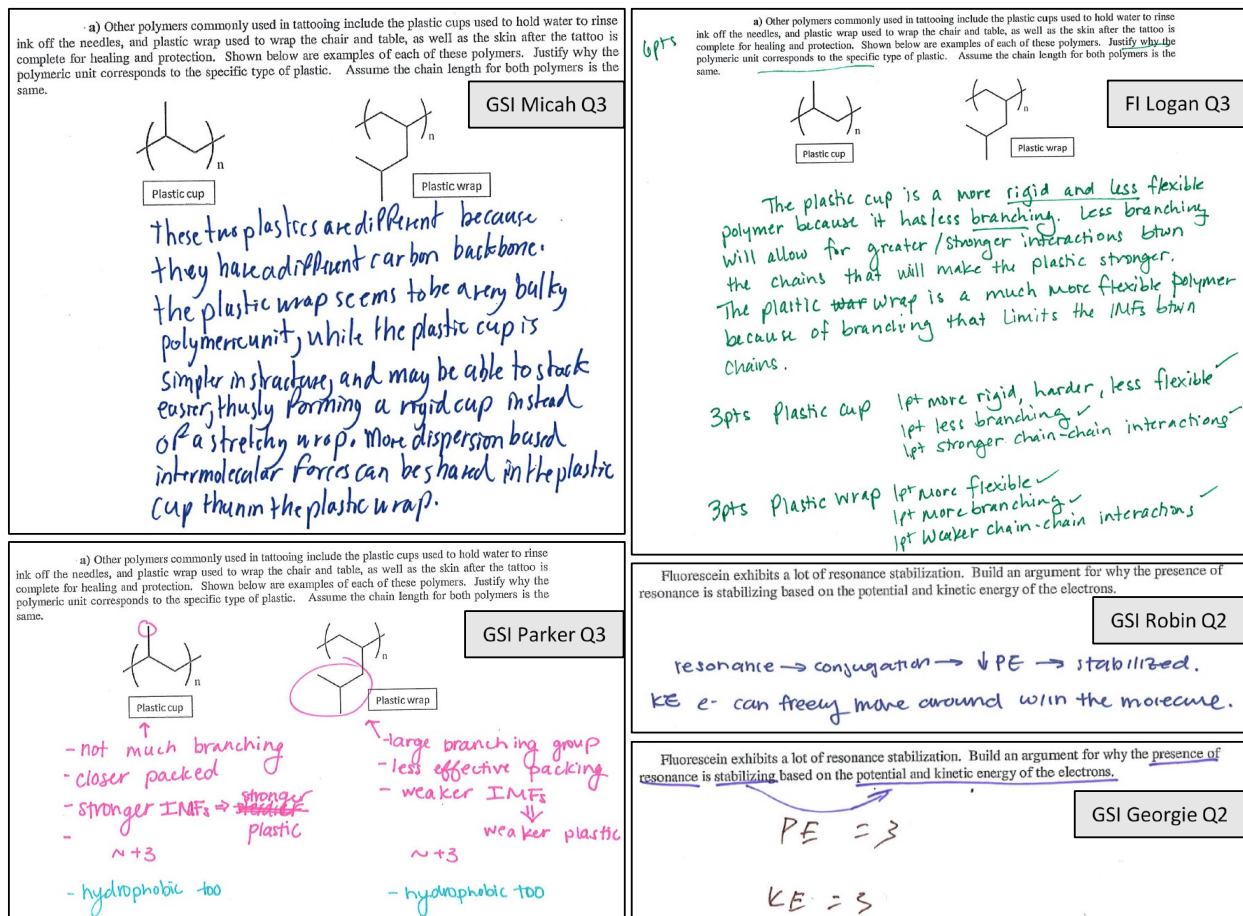


Figure 9: Examples of Rubrics and Expected Answers

Instructors in this category indicated that they built and used rubrics to ensure consistency in the evaluation (FI Elliot) or because they were used to having one while grading (GSI Riley).

In most instances (41%), instructors in our study relied on an *implicit* Rubric while assigning points (see examples for GSI Micah and GSI Robin in Figure 9). As illustrated by the following interview excerpt, several of these instructors could not justify their assignment of points when asked after the fact:

I: *“How did you decide that part of the answer was worth two points?”*

GSI Jackie: *“I don't know. I guess I felt a little bit arbitrary about assigning two points to that. But I didn't really feel like it was a fully zero-point answer. And I don't know. I think I might just be a bit of a generous grader to be perfectly honest.”*

Other instructors said they had a point breakdown in mind, but their assigned scores did not align with the assigned grades, such as with GSI Riley, who assigned 2/6 and 4/6 points to students' responses to Q4 (TS), and said this:

I: *Did you break down the points in your head in any particular way for this question?*

GSI: *Yeah. I kind of gave like 3 points to the... diagram just because it's probably what I've been habituated to do.*

I: *So, 3 points to the ... diagram and 3 points for the response or the explanation.... How did you get to four points?*

Other instructors chose not to write a Rubric before engaging in the evaluation because of their desire to see what the students would say (GSIs Frankie, Alexis, and Oakley).

Instructor behaviour along this dimension was quite consistent. Four FIs and 5 GSIs consistently generated an *explicit* Rubric before engaging in the evaluation of student work, while

3 FIs and 13 GSIs followed an *implicit* approach. Only four instructors (all GSIs) built *explicit* Rubrics for some questions but not others.

Expected Answer: In this dimension, instructors varied in whether they explicitly *expressed* or not (*unexpressed*) an Expected Answer for each question before engaging in the evaluation of student work. Most instructors were consistent in this dimension. There were 2 GSIs (one novice and one experienced) who varied in their approach and wrote answers for some questions but not others. Expected answers varied from full paragraphs (GSI Micah in Figure 9) to bulleted lists of main points (GSI Parker and GSI Robin in Figure 9). Some instructors first wrote an expert answer that they then annotated to highlight salient points (FI Logan in Figure 9). These instructors indicated that this allowed them to consider what the most important parts of the answer were for purposes of evaluation. Other instructors chose to simply write bullet points because they did not want to constrain themselves to a particular expression or construction of an answer, or because that allowed them to identify key words or phrases they expected to see in students' answers. Similar to the use of rubrics discussed above, several instructors, particularly FIs, indicated their decision to include an *expressed* answer was due to habit in writing rubrics for others to use for exam grading.

Some instructors did *not generate* an Expected Answer (*unexpressed*). The following excerpt illustrates the reasoning behind this behaviour:

"I just look at the question, try to understand what the question is about, and get a sense of what would be-- I don't necessarily think about the answer, but what I would like to see in the answer. So typically, I don't write an answer of my own. So, I just start looking at the answers." (FI Shiloh)

Some instructors in this category did have *explicit* rubrics while others did not. For example, GSI Georgie had an *explicit* Rubric for every question but decided not to include an *expressed* answer for any of them. These Rubrics described the allocation of points for different components of an answer without explicitly stating the expected response.

There was some correlation of instructors' behaviours in the dimensions in the first two stages. Primarily, engaging in *close reading* of a prompt correlated moderately with building a *generated* answer and with writing an *explicit* Rubric. There was also moderate correlation between generating an answer and building an *explicit* Rubric. This latter correlation is indicative of the tendency of some instructors to build an Expected Answer that was then used as a Rubric by assigning various points to different parts of the expected response. While this correlation between *closely reading*, having an *explicit* Rubric, and *generating* an Expected Answer existed across all student responses, we feel this is particularly important to acknowledge in the context of significant effects on grades for those first-year GSIs, discussed in a later section. The moderate, but not strong, correlation of these dimensions further describes the variable nature of the instructors' approaches and the need for multiple dimensions to describe the choices made in the first two stages.

Grading System: In this dimension, instructor behaviour was categorised as *criterion referenced* and *norm referenced*. Most instructors applied a criterion referenced approach (in 79.1% of all graded samples) and were quite consistent in their approach. Only two instructors (1 FI and 1 GSI) manifested a highly variable behaviour, often grading the first student response for a given question using a *criterion referenced* approach but using a *norm referenced* approach when analysing the response from a second student.

Instructors who used a *criterion referenced* system evaluated each student's work independently of others and referred to an *explicit* or *implicit* Rubric when explaining their grading choices. Instructors' approach in this dimension became apparent when they were pressed to justify their assigned grades. These instructors often referred to a Rubric or an Expected Answer to defend their assignment of points, citing specific information that was necessary to answer a question.

Instructors who graded with a *norm referenced* system frequently expressed the need to analyse several answers before making an *explicit* Rubric or *generated* an Expected Answer. The instructors who took this approach often *expressed* their preference explicitly, indicating their practices during exam grading were similar to those observed during the interview. The following interview excerpt illustrates this approach:

"I first looked at a relative thing. So, I didn't grade them one and then move on to the next. I looked at both answers and then made a relative grading kind of thing."

(GSI Lou)

Stage 3) Noticing in student work. Variations in the evaluation of student work in this area were observed along three major dimensions: Stance, Lens, and Scope. The dimension of Stance considers the extent to which instructors engaged in the interpretation of students' expressed ideas. The dimension of Lens characterises instructors' consideration of diverse factors that could have influenced a students' answers, while the dimension of Scope refers to the extent to which instructors analysed an answer in fragments or in a wholistic manner.

Stance: Within this dimension, an instructor could evaluate a student answer simply based on what was written (*literal*) or could construct inferences about student understanding that went beyond what was expressed (*inferential*). Instructors who took an *inferential* Stance often

expressed a desire to “*be on [the student’s] side*” (FI Addison), “[*give*] *the benefit of the doubt*” (GSI Parker) or expressed reluctance to take off points for “*tiny mistakes*” (GSI Riley). Some instructors who took an *inferential* approach did acknowledge that they were making assumptions (GSIs Jackie and Morgan) or inferring understanding (FIs Elliot, Jamie, and Shiloh and GSIs Frankie, Kendall, Morgan, and Riley), even if a student answers included mistakes. For example, GSI Oakley stated “*I think I would still give them full credit because I think this was just a slip up here.*”

In contrast, those instructors who took a *literal* Stance did not make assumptions or built inferences, like GSI Finley, who was aware of their Stance when they talk about the student response to Q4 (TS):

“From their language, right, how they wrote their question, you can get the impression that they’re trying to make us infer that if the unfolded is this way, then the folded has to be the other way. But that is not something that I think is OK with giving points for. If they don’t directly state what they’re talking about, I’m not supposed to be able to just assume because I already know what the answer is supposed to be. So, I can’t just assume that they know what the right answer is supposed to be, which is why they received most of the points, but not all of the points.” (GSI Finley)

Other instructors manifested their *literal* Stance through comments about only grading what was on the paper (GSI Morgan) or “*giving credit for things that are completely right*” (GSI Dylan). FI Shiloh explicitly stated “*I almost bet that if I ask this question to the student, they would be able to tell me. But I don’t have evidence [that they know the right answer]”* about student response 1 to Q5 (RD).

Of 29 instructors participating in our study, 2 FIs and 1 GSI were variable in their approach. These three instructors seemed to be aware of the variability in their behaviour along this dimension, such as FI Blair, who stated, “*This one was tough because this is someone who clearly gets it and completely went a different direction with the question. These are the ones that kill me*” when discussing a student response to Q4 (TS). This instructor assigned 1 out of 6 points to the response to be consistent in the evaluation but felt conflicted because the student “*grasped the idea of this [TS] diagram*”.

Lens: In this dimension, an instructor could approach the evaluation of student work looking for key ideas prescribed in an Expected Answer (*Prescriptive*) or could more flexibly recognize that students could express an idea in different ways (*Adaptive*). Participating instructors more frequently followed an adaptive approach (in 59.1% of all instances) and were quite consistent in their behaviour in this dimension.

Instructors who took an adaptive approach often revised their Rubric or Expected Answer (if they had been *explicit* and *generated*) in reaction to new student ideas (FIs Blair, Elliot, Logan, and GSIs Alexis, Dana, Emerson, Frankie, and Oakley), or even expressed a desire to look at student work before *generating* such a Rubric or Expected Answer (GSIs Alexis, Taylor, Pat, Sidney, and Micah and FI Logan). In Figure 9, a change to the Rubric can be seen by the different colours used by GSI Parker. A few explicitly mentioned a need to be flexible in their approach (FI Addison, GSI Emerson) because of previous experiences. Some instructors took an *adaptive* Lens when confronted with unusual or unexpected arguments that had some validity, as illustrated by this evaluation of student response 1 to Q2 (RS):

“The student was making a point that I thought was an interesting one. I had never ever thought in that way, but it was an idea. It's not the way I would think about

resonance, but you can think that one pair electrons for bonds and the formation of bonds typically lowers the energy of a system. So that was a clever argument.” (FI Shiloh)

In contrast, instructors who took a *prescriptive* Lens often claimed they could not give full credit because the student was missing a key term (GSIs Kendall, Frankie, Dana, Finley, Lou, Micah, and Riley and FI Blair) or expressed that the student answer was correct simply because it matched their key words (GSIs Alexis, Sidney, and Riley, and FI Addison). For example, FI Addison said this about Q2, *“I’m looking for the word delocalization, right?”*, while GSI Lou discussed whether or not the student *“mentioned”* a term for every response evaluated and said they were *“looking for”* a particular idea for each question. GSIs Alexis, Micah, and Morgan all explicitly used the term *“key word”* when talking about what they were looking for in the student response:

“Accelerates is a key word in that” – Morgan on Q6;

“I would mention randomness, because that’s a key word in the question for that.”

– Micah on Q4;

“They’re saying, when you’re adding more stuff, their action accelerates. That was the key word there.” – Alexis on Q6.

Most instructors in our sample were consistent in their behaviour along this dimension. Only two instructors (both FIs) were variable in their approaches in this dimension. This variation is exemplified by FI Elliot’s behaviour, which included point deductions for not following directions on some questions while considering it “fair” to give points for different types of arguments on other questions.

There were correlations between approaches in the dimension of Stance and the dimension of Lens. Taking a *literal* Stance in the evaluation of student responses correlated moderately with having a *prescriptive* Lens. Instructors who were looking for specific terms in a student answer were often *literal* in the interpretation of what students wrote and less likely to engage in interpretation, as illustrated by this excerpt:

“They do mention speed, which, OK. But again, we're really specific that if they're going to mention speed, they have to tie it into this average kinetic energy, which we measure through temperature. And so, using this model, those are key points. And I just didn't see them in here at all.” (FI Elliot)

Alternatively, those who engaged in interpretation were also more likely to value answers that deviated from their initial expectations for normative responses.

Scope: Along this dimension, instructors could evaluate student work by considering one expressed idea at a time (*piecemeal*) or analysing an entire answer (*wholistic*). As shown in Figures 7 and 8, these different approaches were observed almost equally in all graded instances and more instructors varied in their behaviour in this dimension. Those instructors who graded in a *wholistic* fashion were looking for “*the overall picture*” (GSI Kendall) and were aware of their preference to read the full response first, before providing an evaluation (GSIs Dylan, Finley, and Micah). These instructors often deducted points for internal inconsistencies in a student’s response, as illustrated by this evaluation of student response 1 to Q5 (RD):

“The reaction would be reactant favoured. This is because the conjugate base is weaker than NH₃. So, it is less likely to accept protons. Weaker, less likely. That is right, but it's not right... That sentence completely contradicts this sentence. That's

wrong. The two just completely contradict each other. That doesn't make any sense to me. That's absolutely incorrect.” (GSI Micah)

This instructor gave 0 out of 6 points to this student’s answer. The same response got 3 out of 6 points by another instructor who followed a *piecemeal* approach and identified the same pieces of correct information. A *piecemeal* approach was often linked to the use of *explicit* Rubrics that had clear point breakdowns and each idea or key phrase was awarded points regardless of what else was in a student’s answer. As GSI Ollie expressed, *“I just leave a checkmark or highlight that sentence. So based on those highlighted portions, I decide how many points are recorded for the student's answer.”* Several instructors (FI Addison, FI Logan, GSI Micah) explicitly noted their preference for assigning points to different parts of a response: *“key of two points for this, two points for this, two points for this”* (GSI Micah) and tended to highlight the specific portions of a response that earned those points: *“if they just say the shift, it's worth points”* (GSI Micah).

Compared to other dimensions, instructors were more variable in their evaluation Scope. Seven instructors (3 FIs and 4 GSIs) alternated between a *wholistic* and *piecemeal* approach when grading student work. In some cases, the shifting between approaches seemed to be linked to conflicting past experiences and personal preferences. For example, GSI Micah referred to the way rubrics were written in the department and how GSIs were expected to evaluate work in reference to those rubrics (which may favour a *piecemeal* approach) but this instructor did not use an *explicit* Rubric when grading independently. Several of the FIs mentioned having a “team of graders” (Addison, Logan, and Elliot) and their need for building rubrics that assigned points to specific components of an answer. Nevertheless, they were more likely to follow a *wholistic* approach when personally evaluating student work.

In line with these comments, having a Focus on *reasoning* correlated mildly with having a *wholistic* Scope. Instructors who looked at an answer in its entirety when evaluating student work were more likely to pay attention to student *reasoning*, while those who took a *piecemeal* approach more often focused on the presence of pieces of *knowledge*. Instructors who manifested a *wholistic* Scope and a *reasoning* Focus often referred to the need for reading the whole response, sometimes multiple times, in order to understand what students were trying to communicate.

Taking an *inferential* Stance correlated moderately with having a *wholistic* Scope. Instructors who evaluated a response as a whole were more likely to engage in making sense of student ideas and infer understanding. Consider, for example, the following excerpt:

“I think I would still give them full credit. Because I think this was just a slip up here. Because later, they talk about-- well, on the other hand, the plastic wrap has more branches that allow it to block the background and make it less rigid. So here, he's using less rigid in reference to the plastic cup as in my interpretation. So, it makes me to believe that this is just this word here is just a slip up. Because they even say that it makes it hard to move and is less flexible.” (GSI Oakley)

This instructor considered the entire answer to build an evaluation based on inferences about student understanding despite explicit inconsistencies in the written response.

Stage 4) Responding to Student Work. Variability in this area occurred along three dimensions: Interaction, Intention, and Feedback. The dimension of Interaction characterises the extent to which instructors makes *expressed* marks or feedback on students' responses. The Interaction dimension can be further subcategorized by type: Marking for underlining or circling student work and Feedback for written comments or notes. The dimension of Intention seeks to identify the intentionality of those evaluation marks, and the dimension of Feedback characterises

the types of guidance or comments provide by instructors in the evaluation of student work. An example of responding to student work can be seen in Figure 10. Every instructor was asked if the amount of Interaction expressed was typical of their practices during exam grading to identify variations caused by the unusual setting. No instructor indicated substantial differences.

Interaction: During the evaluation of student work, most instructors and in most instances made *explicit* marks or gave feedback on student work (*Expressed*) and fewer did not (*Unexpressed*). In terms of dominant approach, 4 FIs and 16 GSIs consistently marked students' answers, five instructors (all GSIs) consistently did not make any other marks beyond the final assigned grade, while 3 FIs and 1 GSI were variable in this dimension.

Intention: Some instructors made marks of student work to highlight elements that caught their attention (*reactive*) while others made marks to highlight correct or incorrect statements (*evaluative*). There were 183 student responses with *expressed* marks. The most common *evaluative* marks included cross outs and question marks to point to incorrect elements and check marks for correct ones. Instructors who were more *reactive* in their approach tended to use circles and underlines to indicate positive or negative reactions to students' answers. When asked what a circle or underline meant, these instructors' answers ranged from "*I made these marks more for me than for the students*" (FI Shiloh) to "*I'm like excited*" (FI Addison) to "*I'm not sure what they were saying*" (GSI Emerson). As shown in Figure 8, variability in this area was among the greatest in all dimensions.

Feedback: Instructors provided Feedback in the form of explicit statements (*direct*) or in the form of questions (*indirect*). There were 88 student responses with Feedback, *Direct* Feedback often mirrored statements or phrases that were found in an *explicit* Rubric or *generated* Expected Answer. Other Feedback pointed out specific portions of the answer that were incorrect or missing.

Fluorescein exhibits a lot of resonance stabilization. Build an argument for why the presence of resonance is stabilizing based on the potential and kinetic energy of the electrons.

Resonance will make a structure more stable because it shares more electrons which will delocalize them more than if they weren't shared. This delocalization will decrease the KE of the electrons due to quantum effect because they have more room. The decreased KE will make it so you have to add more energy to break the bond which is more stable because sharing the electrons. Likewise resonance is more stable because sharing the electrons creates a stronger force of attraction between the atoms. This makes the atoms closer which will lead to a decrease in PE. They are in order to separate it you will have to add extra energy which is more stable.

what shares more e-? →
-0.5
which electrons?
unclear writing

4.5/6

-no clear definition of resonance -1

Figure 10: Sample of instructor response to student work: GSI Dylan, Q2, Student Response 2, including expressed interaction, evaluative intention marking, and indirect and direct feedback.

Much of the Feedback was directive in nature, prompting specific actions that would make the answer more correct or complete. *Indirect* Feedback also addressed these holes in the student responses, and frequently took the form of “Why?” (7 instances) or “How?” (4 instances). Also frequent was a key word or two followed by a question mark, such as “IMF’s?” or “Only cup?” As shown in Figure 8, variability in this area was among the greatest in all dimensions.

Correlation of Dimensions with Assigned Grades

Overall, we collected 268 instructor evaluations of pairs of individual students’ responses to six different general chemistry questions with an average assigned score of $61\% \pm 30\%$. Variability in the quality of students’ answers can be expected to result in variation in the grades assigned by participating instructors. Our quantitative analysis of the collected data indicated that 46% of the variance observed in all assigned grades was uniquely accounted for by which student response was being graded (within the multilevel model). The effect size of the student response variables was large as measured by Cohen’s $f^2 = 0.88$. The variation in the grades assigned to different

student responses' is summarized in Table 3. As shown in this table, three of the student responses were assigned grades ranging from 0% to 100% by different instructors. Seven student responses were scored at a 0% by at least one instructor and eight were scored at 100% by at least one instructor. The average range of assigned grades was 70%.

Instructor differences accounted for 12.8% of the variance observed in assigned scores as determined by the interclass correlation coefficient ($\rho = 0.128$). Modelled instructor variables, however, were found to have negligible effect sizes (Cohen's $f^2 < 0.02$) indicating that these effects did not account for a substantial amount of the overall variance observed, leaving unexplained overall variance and between-grader variance. Instructor gender and research area did not have significant effects on grading outcome. Effects from the interaction between instructor role (FI or GSI) and teaching experience were, however, significant. In particular, being a first-year GSI had a significant effect on assigned grades compared to more experienced GSIs and FIs. Grades by the more novice GSIs were on average higher than those assigned by more experienced instructors.

Table 3: Average Scores Assigned

Student Response	<i>N</i>	Avg Assigned Score	Min Score Assigned	Max Score Assigned
PMM i	28	54.8%	0%	100%
PMM ii		61.6%	0%	100%
RS i	28	31.8%	0%	83%
RS ii		87.5%	33%	100%
SPR i	29	77.3%	33%	100%
SPR ii		68.1%	33%	100%
TS i	29	56.3%	0%	100%
TS ii		76.7%	25%	100%
RD i	10	21.7%	0%	50%
RD ii		90.8%	66%	100%
PCE i	10	33.8%	0%	63%
PCE ii		25.0%	0%	50%

Of the different dimensions of analysis in the framework applied in our study, four of them were found to be statistically significant effects on assigned grades but dependent on type of instructor. Differences in approach by first-year GSIs in the dimensions of Expected Answer (*generated* or *unexpressed*) and Grading System (*criterion referenced* or *norm referenced*) were significant effects.

For the more experienced instructors, the Stance (*inferential* or *literal*) dimension was a significant effect. Additionally, the dimension of the Marking Interaction (*expressed* or *unexpressed*) was a significant effect for experienced instructors, but not the Feedback Interaction (*expressed* or *unexpressed*) indicating that whether or not marks were made on the student work predicted a difference in assigned grades, but writing or not writing Feedback was not correlated to a difference in assigned grades. The Marking Intention (*evaluative* or *reactive*) dimension that further categorized markings was not a significant effect.

To illustrate the effects on assigned grades of variation in approaches to the evaluation along the relevant dimensions for different types of instructors, the results of the multilevel model were used to calculate predicted scores for an average student response (a response with the average score of 61%). The results are summarised in Table 4. All predicted grades can be assumed to be on average, across all graders, and controlling for all other variables. The average predicted grade was 68% for first-year GSIs and 47% for more experienced instructors. The difference between the highest predicted grade (78%) and the lowest predicted grade (35%) was 43%.

Predicted scores for first-year GSIs of an average student response ranged from 57% to 78%. Grades from instructors in this category who had an *unexpressed* Expected Answer were likely to be 9% higher on average than those who *generated* an Expected Answer, while grades

**Table 4: Predicted grades for an average student response
First-Year GSI's**

	Grading System	
	<u>Norm-Referenced</u>	<u>Criterion-Referenced</u>
Expected Answer		
Unexpressed	78%	67%
Generated	69%	57%
<hr/>		
All Other Instructors		
	Stance	
	<u>Inferential</u>	<u>Literal</u>
Interaction		
Expressed	57%	48%
Unexpressed	48%	35%

from those who used a *norm referenced* Grading System were likely to be 11% higher than those who applied a *criterion referenced* Grading System.

Predicted scores for the more experienced instructors ranged from 35% to 57%. The grades assigned by more experienced instructors who adopted an *inferential* Stance were likely to be 11% higher on average than those who took a *literal* Stance. The grades assigned by instructors in this category who interacted with student work were likely to be 11% higher than those assigned by instructors who did not make any Marking.

Discussion

Our study was guided by three research questions seeking to: 1) identify the main approaches to the evaluation and grading of student work that chemistry instructors in our sample more commonly followed, 2) characterise consistency in the application of those approaches across different evaluation instances, and 3) determine the potential impact of variations in approaches on assigned grades. In this section, we discuss our main results in these three areas.

Common Main Approaches

Our analysis revealed major differences in the approaches followed by different instructors in the evaluation of student work. These differences manifested in all of the dimensions of analysis used in the study as summarised in Figures 7 and 8. There were a few dimensions in which more than two thirds of our study participants followed a similar approach. These included Depth, with a majority of instructors engaging in *close reading*; Focus, with a majority of instructors paying attention to student *reasoning*; Grading System, with a majority of instructors using a *criterion referenced* approach; and Interaction, with a majority of instructors *expressing* interactions with student work by making some type of mark or feedback. In our small sample, no approach was found to correlate with specific instructor characteristics, such as gender, role, research area, or teaching experience.

The lack of common main approaches by instructors is intriguing due to the similar justifications that kept appearing among instructors. For instance, several instructors seemed to justify their deduction or award of points based on the presence (or lack thereof) of particular “key words”. This word-search method of evaluation has been utilized in attempts for automated essay scoring (Klobucar et al., 2012) and shows up in the dimensions (categories) of Focus (*knowledge*), Lens (*prescriptive*), and Scope (*piecemeal*). While each of these dimensions had this overlapping justification, the dimensions did not have strong correlations, meaning the same justification was used in different ways for different instructors.

The most common approaches to the evaluation of student work followed by our study participants may have been influenced by contextual factors. For example, participating instructors taught a general chemistry curriculum that emphasizes the development of students’ chemical thinking. At the level of discourse, the general chemistry teaching team often highlights the

importance of students building their own explanations and justifying their reasoning. Similarly, study participants were used to evaluating student work using a common rubric after holding a collective meeting in which expected answers, point assignments, and sample student responses are analysed and discussed. This practice may have favoured the application of a *criterion referenced* system when evaluating student work during the interview. Additionally, in grading meetings instructors were encouraged to justify their grades through *explicit* Marks and Feedback on student work, which may have also influenced their behaviour.

Reliance on a criterion-reference system when evaluating student work could be associated with instructors' beliefs about fairness. During the interviews, multiple instructors referred to the need “*to grade them all exactly the same*” so that “*it's fair*” (GSI Dana). References to fairness also appeared in comments by instructors regarding the adoption of a particular stance or lens. For example, whether or not it was “*fair*” to award points to alternative ways of writing or thinking when evaluating student work. Nevertheless, judgments about fairness led instructors to follow contrasting approaches in these dimensions: GSI Robin was inferential and prescriptive, and stated, “*Just grading, in general... I'm trying to be fair*” while GSI Morgan was literal and adaptive and regularly justified their grading as “*a fair way to do it.*”

Although local curriculum and evaluation practices may have influenced the approaches that study participants followed during the interviews, the impact of these factors was limited in other areas. Despite, the institutional practice of using a provided rubric based on expected answers to evaluate student responses during grading meetings, only 41% of the instructors generated an explicit rubric to guide their evaluations during the interview. Comments from participants suggest mixed ideas about the value or utility of rubrics as they were judged to be “*hit or miss*” (GSI Dana), a “*loose reference*” (GSI Oakley), or they were openly disliked (GSI Parker).

Another reason for the commonality in the *criterion referenced* Grading System may be a sense of fairness by the instructor. Fairness has been discussed by multiple researchers as a reason for using rubrics (Andrade, 2005; Holmes & Smith, 2003; Randall & Engelhard, 2010), but here we see instructors citing fairness in regards to their preferred Grading System. Multiple instructors cited their beliefs about the need “to grade them all exactly the same” so that “it’s fair” (GSI Dana), and this is easier to do with an external *criterion referenced* system. These instructors were able to apply this Grading System even when they did not have an *explicit* Rubric or *generated* an Expected Answer. This sense of fairness also appears in comments by instructors regarding the Stance and Lens; whether or not it is “fair” to allow alternative ways of writing or thinking when evaluating student work. These dimensions were more evenly split between their categorical options, indicating the decision on which end was “fair” was less consistent. In fact, different instructors did cite fairness as their reason for each category. GSI Robin was *inferential* and *prescriptive*, and stated, “*Just grading, in general... I’m trying to be fair.*” GSI Morgan was *literal* and *adaptive* and regularly justified how they graded as “*a fair way to do it.*”

Another surprising lack of commonality is for the dimension of Focus. The curriculum used for general chemistry at our institution has a heavy Focus on *reasoning* and critical thinking skills (Talanquer & Pollard, 2010). However, a third of the student responses were graded with a *knowledge* Focus. This approach may be a result of instructors seeking to award points to students for providing some kind of answer, even if the complete or ideal answer is not submitted. As noted in the results section related to Focus, parts of the answer could be identified as a good start or critical baseline information that earned points even if there was no connection or utilization of that information. This desire to find something to award point for is discussed further in the following section regarding consistency.

Consistency in Approach to Evaluation

While each instructor approached the evaluation of student work in different ways, each participant was generally consistent in the approach they followed during the evaluation of different student responses. These preferences may be indicative of strongly held beliefs regarding learning theories, purposes of assessment, and philosophies of teaching and education. Work in the area of beliefs about teaching and learning does indicate the strength of these preferences is likely to be high (Gess-Newsome et al., 2003; Gibbons et al., 2018). Evaluating consistency in the method presented in this paper also allows identification of trends even when actions are at a mismatch with stated values. While an instructor may state particular learning objectives and intentions but not actually grade in accordance with those values (Mutambuki & Fynewever, 2012; Petcovic et al., 2013), each instructor does tend to be consistent in their approach.

It is possible that the observed preferences in the evaluation of student work across different questions was influenced by the number of responses that participants were asked to analyse. Under real conditions instructors typically grade a far larger sample of responses which may influence their behaviour. For example, under such conditions, a *norm referenced* Grading System may become more prevalent or compete with a *criterion referenced* Grading System if graders are influenced by dominant students' responses.

As summarized in Figure 11, most instructors exhibited variation in at least one category, but no instructor was variable in more than four dimensions. In general, faculty instructors were more variable in their approaches across student responses than GSIs. Close to 41% (9 out of 22) GSIs were fully consistent in their approaches across all evaluations. Several instructors seemed to be aware of the variability in their behaviours, particularly when conflicted with a student answer.

Some instructors expressed acknowledgement of their own preferences and frustration with these approaches, such as FI Blair, who assigned a 1/6 to a student response and stated, “*This one was tough because this is someone who clearly gets it and completely went a different direction with the question. These are the ones that kill me.*” This instructor had a strong preference for a *literal* Stance and *prescriptive* Lens, yet recognized that there were alternative ideas that might be

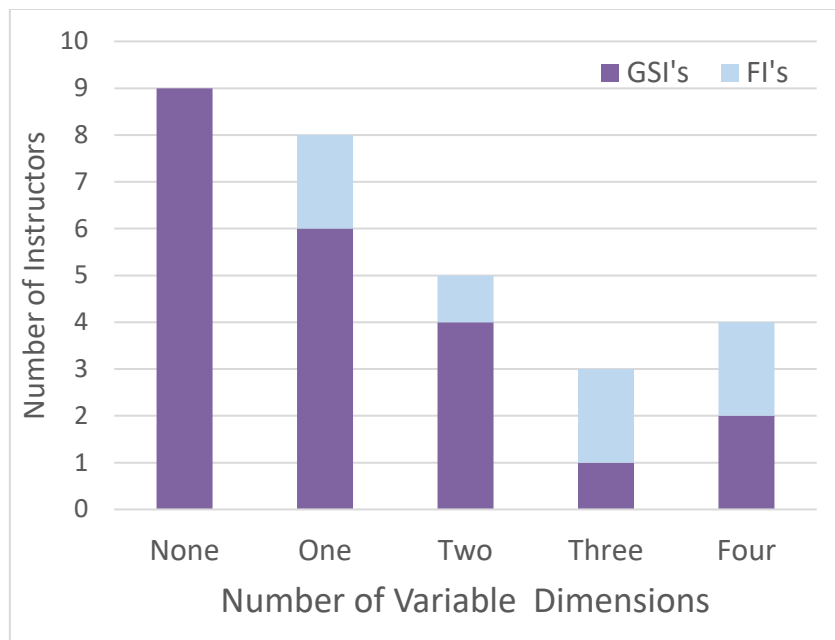


Figure 11: Number of dimensions in which instructors are variable

productive. FI Blair went on to say that they felt they “*had to be consistent with everyone*” in order to justify the low grade.

For other instructors, they noticed a shift in their own behaviour because they were attempting to “find a point” for the student (GSI Parker) or were responding to experiences that they, again, felt was fair or reasonable. These shifts often related to the dimensions of Focus or Scope, both of which had the highest levels of variability. With reference to the Focus and Scope, three instructors were flexible in both dimensions: GSI Micah and FIs Logan and Elliot. These instructors were all experienced and gravitated towards adjustments to benefit students, having

adaptive or flexible approaches to the dimension of Lens and utilizing *generated* Expected Answers and a *criterion referenced* Grading System. These instructors seemed to have some conflict between looking for well-reasoned answers and crediting students for “*little snippets*” of information that had “*merit*” (FI Logan).

FI Elliot was one of the most flexible instructors, with shifts in the dimensions of Focus, Lens, Scope, and Interaction. This instructor tended to make a claim and stick to it, but also adapted if there was sufficient volume. While talking about their approaches and experiences, this instructor stated,

“[If] there's a lot of misconceptions. Whether it's really how we teach or how they interpret it, it's hard to say. But I've always felt that if they're all my students and they all have the same misunderstanding, there's something I can do better and I'll adjust the Rubric because I do take some of the, quote unquote, "blame" for that.”

This instructor’s approaches were influenced by the content of students’ responses which led to changes in expectations and grading criteria. Further, this instructor’s perceptions on misconceptions and learning objectives shapes how they respond to alternative, incomplete, or incorrect ideas that are represented on assignments. Understanding how these perspectives shape approaches, and by extension the grades assigned, can highlight the importance of the act of noticing during the evaluation of summative assessment responses.

While each question may have been considered using different approaches, the shift between approaches did not occur regularly between student answers. Shifting approaches between the two student responses only happened in 10 cases out of 134 questions, and most commonly noted a shift in Scope ($n = 5$) or Focus ($n = 3$). With this in mind, the variability within an instructor does not seem as critical as the differences between instructors, and how different

approaches may impact the assigned grades. This highlights the productive use of noticing in understanding how instructors engage in evaluation of summative assessment and further demonstrates the contribution of the dimensions of variation framework in providing language to describe and discuss approaches.

Correlation of Dimensions with Assigned Grades

Although our sample was small, our quantitative analysis indicated that the role and teaching experience of the instructors had a significant correlation with the assigned grades. The less experienced first-year GSIs tended to assign higher grades than more experienced instructors. Unfortunately, in our study we only had GSIs in their first, third, and fifth years. Additional research with a larger sample of GSIs of differing experience levels would be needed to further analyse the observed relationships.

Interestingly, the dimensions of evaluation that were found to correlate to differences in assigned grades were not the same for these two groups of instructors. For first-year GSIs, Expected Answer and Grading System, both in the stage of “Preparing for and Engaging in the Evaluation of Student Work” had a significant effect on assigned grade, with instructors who *generated* an Expected Answer and/or adopted a *criterion referenced* approach assigning the lower grades. On the other hand, relevant dimensions for more experienced instructors included Stance and Marking Interaction dimensions in the stages of “Noticing in” and “Responding to Student Work”. In this case, instructors who adopted a *literal* Stance and/or did *not express* Marking on student work tended to assign lower grades.

For the least experienced GSIs, leaving the Expected Answer *unexpressed* was correlated with assigning higher grades to the student responses. The lack of significant effect of this dimension for more experienced instructors could be because they may have a more solid

understanding of the evaluated concepts and have been exposed to similar questions in the past. The lack of a *generated* Expected Answer by less experienced instructors may indicate they have a weaker understanding or lack of familiarity with a topic and therefore be less harsh in their evaluation of student responses.

For first-year GSIs, having a *norm referenced* Grading System was correlated with assigning higher grades to the student responses. This could be explained as their comparison may be limited to only the two samples provided during the interview, leading to assigning the best possible score to the better of the two responses. Using a *criterion referenced* Grading System would not result in such high scores due to the lack of a polarized high score. The lack of significant effect in this dimension for more experienced instructors could be due to their experience with a broader range of student responses. Having experience with better student responses may have led them to not feel obligated to assign the best possible score to one of student responses for each question given to them in this study.

These findings regarding first-year GSIs align with work that shows a need for explicit discussion regarding grading practices with teaching assistants (Marshman et al., 2018; Yerushalmi et al., 2016), and the need for more intensive training for new instructors (Mutambuki & Schwartz, 2018). The differences between first-year GSIs and all other instructors shows that experience has a significant effect on assigned grades and further suggests a need for continued professional development related to the stages of noticing and responding to student work to help close the gap between instructors' assigned grades that are not addressed by rubrics alone.

For more experienced instructors, having an *inferential* Stance was correlated with assigning higher grades to the student responses than having a *literal* Stance. It is possible that instructors who are *inferential* recognize and value productive reasoning in student responses even

when stated in non-normative, incomplete, or confusing ways. This interpretive behaviour could explain the assignment of higher grades. This could be further explained by where those instructors placed the burden of proof. Instructors who are *inferential* make assumptions, possibly about expert heuristics being present or mistakes being made due to carelessness, rushing, or absentmindedness. Instructors who are *literal* do not allow for this benefit of the doubt and stay strictly to what is represented on the page by the student. This echoes results found about mismatches between instructor intention and practice and the burden of proof (Mutambuki & Fynewever, 2012).

It is somewhat surprising that the Marking Interaction was a significant effect on the grades assigned by more experienced instructors. Having an *expressed* Marking Interaction was correlated with assigning higher grades to the student responses. One may hypothesize that Marking is indicative of the depth to which instructors analysed student answers. Instructors who *express* Marking may thus be more likely to identify ideas or pieces of knowledge that are judged as worthy of points, which is reflected in the higher assigned grades.

The negligible effect sizes of the instructor-related variables and the substantial unexplained variance in assigned grades suggest that other effects are at play. Variables such as the specific content of the students' responses, the value an instructor may place on specific information or reasoning displayed in such a response, and the instructor knowledge and understanding of the concepts and ideas under evaluation may be significant effects on assigned grades that could explain more of the observed variance. Further research is needed to determine the effects of these other variables.

Limitations

This study involved a small number of faculty and graduate student instructors within a specialized area (general chemistry) and our analysis focused on their approaches to the evaluation and grading of only six exam questions and two associated student answers. The questions were representative of conceptual questions used in general chemistry exams at our institution, but different types of questions or additional student responses may trigger different evaluation and grading behaviours. We cannot speak to the emotional or social experiences of the participants during the interview that may have affected their observed decisions and preferences.

This study's ability to generalize the findings from the statistical modelling in the section Correlation of dimensions with assigned grades is limited by several factors. The small number of participants yielded limited group sizes of categorical variables and therefore the consideration of overfitting the model should limit the generalization of the findings. Additionally, as independent variables were not randomly assigned, no claims of causal relationships can be made, but only correlative relationships.

Conclusions and Implications

Our work contributes to existing work in assessment and evaluation by demonstrating the utility of adopting a Noticing framework in the analysis of instructors' approaches to the evaluation of student written responses in summative assessments. Previous work on Noticing has predominately focus on formative assessment and taken place in the K12 classroom setting (Chan et al., 2020; Luna, 2018; Russ & Luna, 2013). In this study, we extend and connect ideas about how instructors approach formative assessment to how instructors approach summative evaluation and grading. While engaged in formative assessment, instructors must on-the-fly notice, interpret,

and respond to the ideas put forth by students during classroom discussion. In this study, we operationalized and applied these formative assessment techniques to characterize differences in instructors' approaches to the evaluation and grading of student written work and showed how these differences can greatly impact assigned grades. Particularly for experienced instructors, the choices made in the noticing and responding stages had significant effects on the grades assigned.

Although the instructors who participated in this study taught the same courses, used the same curriculum, and worked collaboratively in developing (FIs) and grading (GSIs) exams, their approach to the evaluation of student written responses was quite varied. Some of the dimensions had a significant effect on the grades assigned. These results suggest that chemistry instructors, both FIs and GI's, would benefit from more training in the evaluation of student work. This training should focus on the critical discussion and reflection of how implicit and explicit decisions made by instructors in multiple dimensions during the grading process can affect the outcome. Instructors must analyse their common practices, preferences, and biases when reading and interpreting question prompts, preparing to evaluate student work, engaging with, noticing, and interpreting students' ideas, as well as in responding to student written answers.

The analytical framework used in this study could be a useful tool in guiding the professional development of chemistry instructors by orienting discussions about implicit and explicit decisions made during evaluation and grading, how those decisions are affected by instructors' individual beliefs about the purpose of assessing, evaluating, and grading student work, and how those beliefs align with the learning goals of their department or institution. The work presented provides a reference in order to communicate more clearly about what is noticed and how to respond to student work. The introduction of noticing to summative assessment practices allows practitioners and researchers the opportunity to discuss and decide how to engage

in assessment and evaluate student work more consistently. Eliciting and characterising instructors' beliefs about the purpose of assessment, evaluation, and grading may help understand the approaches they chose to follow and the reasons for their variation in different contexts. Opening opportunities for instructors to align their goals and develop common beliefs in this area could be critical to ensure the fair evaluation of student learning.

Our findings also indicate that further work is necessary to untangle the relationships between instructors' approaches, philosophies, experience in evaluating student work, and the assigned grades. Although our investigation revealed important dimensions of variation, it also suggested that there are other relevant variables not considered in our study. Identifying the philosophical opinions may help to understand why instructors demonstrate a dominant approach and could speak to the strength of their preferences. Knowing how strongly a preference for an approach is held as well as how differences between instructors evaluating the same work are negotiated could further identify the best practices for implementing training and discussions related to this framework and help reduce variability. Beliefs about the purpose of assessment or the goals could further elucidate how best to reduce unwanted variation in the assigned grades. With this framework established, we can extend research to further identify how best to communicate assessment practices to instructors and better serve the students completing those assessments. Our results further suggest that instructor training may need to be tailored to teaching experience to focus on those areas that are more impactful in the grading behaviour of novice versus more experienced instructors.

CHAPTER 5: CONCLUSIONS AND FUTURE DIRECTIONS

In asking chemistry instructors to evaluate samples of student work submitted as part of exams, we discovered clear stages that each participant moved through during the process. Within each stage, there were various dimensions that had to be considered, with dichotomous categories describing the options in approach for each dimension. A total of eleven dimensions in four stages were identified and described as a framework for analysis. A summary of the framework can be found in Chapter 2, Table 1.

Once the framework was established, instructors' behaviors were coded for each student response, with a total of 268 student responses across 29 instructors and six exam questions. There were no necessary adjustments to the framework and all student responses were able to be coded with the dichotomous method. Trends in approach, dominant preferences, and correlations were then analyzed. Despite all instructors working with the same curriculum and many receiving the same training, there were no predominant trends or dominant approaches that were universally undertaken. The dimensions of grading system and interaction had the highest levels of uniformity across instructors, with approximately 75% engaging in a criterion referenced grading system and expressing interaction. Instructors taking different approaches to evaluating student work could be problematic if the grade depends on the instructor and not the student work.

We then considered how consistent an instructor was in relation to their own habits and preferences. Instructors using the same approach 70% of the time or more were considered consistent. The evaluation of consistency shows whether an individual instructor changes their approach to a student response between questions or between answers, which could be problematic if the shift results in unequal evaluation and assignment of grades, similar to how differences between graders taking different approaches could be problematic. The dimensions of depth,

expected answer, grading system, and lens were the most consistent, while focus, scope, and marking intention were the most variable or flexible.

Next, we considered the relationships between various approaches for two reasons: to confirm all dimensions were necessary for characterization and to further identify patterns in approach. No categories were correlated so strongly as to diminish the need for the detail. The strongest correlations in pattern were between the dimensions of: depth, rubric, and answer; lens and stance; focus and scope; and stance and scope. The correlation of depth with rubric and answer, and between rubric and answer, are somewhat expected, as they pertain to reading and interpreting the prompt and preparing to engage in the evaluation of student work – all of which contribute to setting expectations for what will have value in the student response. The other correlations are less expected and may speak to a desire to find partial credit for poor answers.

Finally, we used HLM to determine relationships between the score assigned and the various dimensions and instructor characteristics. From this, we determined that first year graduate student instructors are likely to assign higher scores than more experienced instructors of any kind, and that different dimensions are relevant to these two groups. First year GSI's have significant effects from dimensions in the stage of preparing and engaging in the evaluation of student work, while experienced instructors have significant effects from dimensions in noticing and responding to student work. Having a generated expected answer and using a criterion-referenced grading system correlated with lower scores for first year GSI's, with an average student response scoring 57%. Comparatively, a first year GSI taking the opposite approach in those dimensions would likely score the same average student response 78%. For experienced instructors, taking an inferential stance and marking the student work correlated with higher scores, with an average student response scoring 57%. This highest combination score for experienced instructors matches

the lowest predicted score by a first year GSI. An experienced instructor taking a literal stance and having an unexpressed interaction would likely assign a score of 35%.

Implications

There are over two thousand combinations to describe the preferences in approach by an individual grader when considering only a dichotomous approach to each dimension. As demonstrated by the variability within instructors, it is likely that instructors exist on a spectrum, meaning the possibilities are nearly endless. With this in mind, and given the impacts of certain dimensions on the assigned grade, it is important to set clear expectations for instructors working in the same course on how to approach grading and interpretation of student work. In particular, it seems necessary to establish teaching and assessment practices related to generating an expected answer and grading with a criterion-referenced system for first year graduate students and to encourage communication between experienced instructors about the stance to take when reading student work. This may reduce differences between graders. Using reflective teaching practices may allow instructors to identify where they fall on the spectrum for each dimension, which can minimize their personal variability.

It is important that different instructors in the same course are assigning the same or very similar grades to the same samples of student work for a variety of reasons, including setting clear expectations, ensuring learning objectives are being met, and minimizing inequities. If different instructors are assigning different grades to the same work, that implies a student's score is not indicative of their understanding. Since the purpose of a grade is to assign a value to the level of understanding, grades that do not align with the objectives are not meaningful. Similarly, if the grades are dependent on the person grading, then meeting expectations becomes difficult, as the person grading may be different for different assignments, as is the case in our department.

Further, the disparities in assigned grade by different instructors to the same work can lead to student complaints about fairness and perpetuate systemic inequities. Grades are predictors of academic success and participation, even if that is because a student simply can not progress to the second semester of general chemistry without passing the first semester course. Asking for a regrade often results in a higher score, which could perhaps be attributed to the differences in the approach between the initial grader and the regrader. Additional research to explore this possibility is needed.

There is much work still to be done in the areas of assessment and grading of student written work. The data presented in this project begins to unravel the complex nature of how various instructors approach grading, and with this information new practices can be established to orient instructors of the same course for better alignment of expectations, evaluation, and assignment of grades. The framework for describing the various approaches that instructors take in the process of evaluating student written work will allow further work to be completed that can assist instructors in grading consistently with themselves and with others.

Additional Data Collected and Possible Future Work

As part of this project, we collected additional data that would further extend the investigation of how instructors evaluate student work and how they respond to the work. I now discuss the additional data and questions that could potentially be addressed with analysis. These additional analyses are beyond the scope of the presented project and, while extending the line of research, are independent projects.

Content and Expectations

As noted, setting clear expectations for students is important for them to be able to succeed. Setting expectations or clarifying expectations can be communicated through rubrics and feedback, which are two areas that participants shared their thoughts during the interview process but have not been fully explored.

Rubrics

During the interview process, the first step was a request for instructors to prepare to assess student work. They were provided with the question prompt and given time to write any notes or thoughts pertaining to the question. For approximately 40% of the questions, the instructor wrote an explicit rubric. Using resource graphs, analyzing the expectations and valuations associated with specific portions of the answer can give additional insight about what chemistry topics instructors expected to see and how they were identified or valued in the student response. Understanding the variety of information, point distributions, and relationships in explicit rubrics can further inform researchers about what content instructors value, as well as clarify how rubrics are used. In a few instances, an instructor generated an explicit rubric with point breakdowns that did not match the assigned scores. (For example, splitting the answer into three two-point sections and then assigning a score of four.) This discrepancy might lead to further understanding of why instructors did not write any rubric, though they are expected for all assignments in our department. Specifically, research probing the content and expectations of rubrics would investigate the commonalities and differences between explicit rubrics generated by instructors and how rubrics relate to the scores assigned.

Feedback

After all student samples were graded during the interview, instructors were asked to shift gears and provide feedback for each student response. This feedback ranged in content and format, and analysis could further explore the expectations that instructors had for the student response. In general, the feedback that was given to students seemed detailed and was mostly phrased as questions. Exploring the feedback that was given could highlight how exam questions are interpreted and how the interpretation relates to the expectations. Additionally, evaluating the feedback in conjunction with the student response could identify specific phrases or concepts that trigger similar responses from instructors. This could be useful in the translation to online or automatic assessment technologies, where feedback has historically been limited. Being able to predict feedback for a student response may also benefit exam writers, who can include information that better captures a student's understanding or be more attuned to separating those who understand from those who don't. Specifically, research probing the feedback would investigate the common style and content of feedback from different instructors and what portions of the student response most commonly result in feedback.

Self-Assessment and Grading Workshop

In addition to the interviews, data was collected from an intervention performed with a different set of first year GSI's. During this intervention, the GSI's were first asked to grade some samples of student work, then completed a self-assessment where they scored themselves on a variation of the framework. The variation included adaptations to several of the first dimensions because they were provided with a rubric. After grading and completing the self-evaluation, the GSI's participated in a workshop/discussion focused on how to use the rubric, what noticing is and the department's preferred approaches in that stage, and how to respond to the student work. Then,

the GSI's regraded the student work and completed the self-assessment a second time. Evaluating this data could indicate the strength of an instructor's preference or their resistance to change, as well as further confirm the findings presented regarding the significant effects of the different dimensions on first year graduate students. The population for this intervention is too small for statistical power, so further data would need to be collected.

Negotiation

The project presented contains information regarding grades assigned by individual instructors and the variations between and within them. However, no one individual is responsible for general chemistry, nor the grading or evaluation. Further research could probe how instructors negotiate and come to an agreement (or fail to) about a grade on student work. Data related to this was collected in two settings: grading meetings and faculty discussions.

Grading Meetings

Each semester there are typically three exams with a free response section, and these exams are graded immediately following the common exam by the graduate student instructors with some leadership by the faculty instructors. Primarily, faculty instructors present a rubric for the GSI's to use while grading and answer questions about student samples. Conversations that took place during these meetings were recorded to identify how differences in interpretations and valuations were resolved. Additional data was collected in the form of the provided rubrics, notes about the discussions, and sample work that was universally graded. Grade variations were found even with a discussed rubric, and negotiation seemed to happen on the small scale rather than wide acceptance. Further research and analysis would investigate who the contributing members in a

negotiation are, who refrains, how an agreement is reached, and whether a higher or lower score is favored in a conflict resolution.

Faculty Discussions

In addition to the grading meetings, a few faculty members participated in discussions about grading for a semester. During these meetings, a handful of student responses from the exam that had just been given were graded by the faculty, who then shared the grades assigned and discussed what the “right” or “appropriate” grade would be. In these meetings, the variations in assigned grade were just as prevalent as in the individual interviews, and instructors were confident in their assessments and resistant to change. While some instructors tended to align with one another, more often each stuck with their initial assignment, even when pushed. Specifically, research in this area would require more data to explore the group dynamics, philosophies, and expectations that give rise to the persistence of variation and resistance to negotiation.

Together, the project presented and the additional lines of research could inform researchers and practitioners alike on how to assess student work consistently, equitably, and meaningfully. Grades are an integral part of higher education and therefore assigning them to student work should be indicative of the student’s performance and not a result of which instructor is reading it. Understanding more about instructors’ approaches can lead to better professional development and more equitable practices.

APPENDIX A – HUMAN SUBJECT APPROVAL



THE UNIVERSITY OF ARIZONA
Research, Discovery
& Innovation

Human Subjects
Protection Program

1618 E. Helen St.
P.O. Box 245137
Tucson, AZ 85724-5137
Tel: (520) 626-6721
<http://rgw.arizona.edu/compliance/home>

Date: November 27, 2018
Principal Investigator: Michelle Denyse Herridge
Protocol Number: 1811096432
Protocol Title: Instructor Evaluation of Written Responses to Assessment in Chemical Thinking

Determination: Approved
Expiration Date: November 24, 2023

Documents Reviewed Concurrently:

Data Collection Tools: *Herridge IRB Attachment A.doc*
Data Collection Tools: *Herridge IRB Attachment C.docx*
HSPP Forms/Correspondence: *Herridge F107.doc*
HSPP Forms/Correspondence: *Herridge IRB Application 2.0.pdf*
HSPP Forms/Correspondence: *list_of_research_personnel_Herridge.pdf*
HSPP Forms/Correspondence: *Required Signatures.pdf*
HSPP Forms/Correspondence: *Required Signatures.pdf*
Informed Consent/PHI Forms: *Herridge IRB Attachment B.doc*
Informed Consent/PHI Forms: *Herridge IRB Attachment B.pdf*
Other Approvals and Authorizations: *COI Certification Complete for 1811096432.msg*
Participant Material: *Herridge IRB Attachment D.pdf*

Regulatory Determinations/Comments:

- The project is not federally funded or supported and has been deemed to be no more than minimal risk.
- The project listed is required to update the HSPP on the status of the research in 5 years. A reminder notice will be sent 60 days prior to the expiration noted to submit a 'Project Update' form.

This project has been reviewed and approved by an IRB Chair or designee.

- The University of Arizona maintains a Federalwide Assurance with the Office for Human Research Protections (FWA #00004218).
- All research procedures should be conducted according to the approved protocol and the policies and guidance of the IRB.
- The Principal Investigator should notify the IRB immediately of any proposed changes that affect the protocol and report any unanticipated problems involving risks to participants or others. Please refer to Guidance Investigators Responsibility after IRB Approval, Reporting Local Information and Minimal Risk or Exempt Research.
- All documents referenced in this submission have been reviewed and approved. Documents are filed with the HSPP Office.

APPENDIX B - IRB CONSENT FORM

University of Arizona

Consent to Participate in Research

Study Title: Instructor Evaluation of Written Responses to Assessment in Chemical Thinking

Principal Investigator: Michelle Herridge

You are being asked to participate in a research study. Your participation in this research study is voluntary and you do not have to participate. This document contains important information about this study and what to expect if you decide to participate. Please consider the information carefully. Feel free to ask questions before making your decision whether or not to participate.

This study will explore how instructors, including graduate assistant instructors, make decisions about what is important and relevant in a student's written answer on formative and summative assessments. Instructors will be interviewed about their exam writing techniques, grading practices, feedback practices, and student relationships, and what is used to evaluate student reasoning. Part of the interview process will include evaluating samples of student work and discussing expectations for course content. This study will also explore what happens when groups of instructors grade together and the negotiations that take place to grade consistently over multiple people.

If you chose to participate, you will participate in individual, video- or audio-taped interviews where you will be asked questions about your exam writing and grading experiences. This individual interview will coincide with the exam you are writing and should last no longer than 60 minutes. Rubric distribution, discussion, and grading decisions made electronically via email will also be collected. The one-hour grading meetings with faculty and graduate instructor group grading experiences may be video-taped or observed. Faculty instructor group interviews will occur following this exam grading during the regularly scheduled faculty meeting. Decisions to continue involvement with the on-going study will be discussed at the beginning of each semester.

There are no expected risks to you as a result of participating in this study. Although the researchers have tried to avoid risks, you may feel that some questions that are asked of you will be stressful or upsetting. You do not have to answer anything you do not want to. You may benefit from participating through developing a deeper understanding of the assessment process and improve their abilities to grade effectively and consistently. Questions posed in our study will require participants to objectively evaluate de-identified student chemistry work not associated with their professional duties. Participants will not be asked to share any personal information, beliefs, or ideas as part of this research. Participants' responses will be recorded in the form of electronic files and stored in a secured computer, using a coding system devoid of

personal information to label them. If computer security were to be breached, all participants will be informed. Data breach would pose minimal risks to the participants as the information contained in the files would not pertain to their personal or professional lives.

If you are interviewed, I would like to audiotape it so that I can make an accurate transcript. If you are observed in a group setting, I would like to videotape the interactions during the one-hour meeting after each exam and the first hour of the faculty meeting following the exam. The video will be transcribed, making sure that no individual not participating in the study can be identified in the tapes by angling the camera away as much as possible, and by blurring or removing faces, should they be captured. The data will be transcribed within a few days of conducting the observations and the video recordings will not be destroyed until 6 years after the completion of the research in case the recordings need to be reviewed at a later date. Your name will not be in the transcript or my notes. The information that you give in the study will be anonymous. Your name will not be collected or linked to your answers. Your responses will be assigned a code number. The list connecting your name to this code will be kept in an encrypted and password protected file. Only the research team will have access to the file. When the study is completed, and the data have been analyzed, the list will be destroyed. Even though we will tell all participants in the study that the comments made during the group grading should be kept confidential, it is possible that participants may repeat comments outside the group. 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.

Information collected about you will not be used or shared for future research studies. The information that you provide in the study will be handled confidentially. However, there may be circumstances where this information must be released or shared as required by law. The University of Arizona Institutional Review Board may review the research records for monitoring purposes.

For questions, concerns, or complaints about the study you may contact **Michelle Herridge**, mdhh99@email.arizona.edu.

For questions about your rights as a participant in this study or to discuss other study-related concerns or complaints with someone who is not part of the research team, you may contact the Human Subjects Protection Program at 520-626-6721 or online at <http://rgw.arizona.edu/compliance/human-subjects-protection-program>.

Signing the consent form

I have read (or someone has read to me) this form, and I am aware that I am being asked to participate in a research study. I have had the opportunity to ask questions and have had them answered to my satisfaction. I voluntarily agree to participate in this study.

I am not giving up any legal rights by signing this form. I will be given a copy of this form.

Printed name of subject

Signature of subject

Date

APPENDIX C - INTERVIEW PROTOCOL

Instructor Evaluation of Written Responses to Assessment in Chemical Thinking

Attachment A: Interview protocol

Two protocols are outlined here for the purposes of validation. The participant may either be asked to give feedback on student responses or to grade student responses first. In either order, the materials and student answers are the same.

FEEDBACK FIRST

Introduction

We are trying to better understand how instructors and TAs approach the assessment of student work. For that purpose, I am going to ask you to look at some samples of student work and evaluate them as you would do for any CHEM 151/152 class. Do you have any questions?

Part 1

Let's start. Imagine that students in a CHEM 151 class were given this question to answer as homework (hand empty sheet with the question (Attachment C)). Your first task will be to assess what [two/three] different students did and provide feedback to them. I am going to give you some time to read the questions.

What do you normally do to prepare for assessing student work? Could you do it now, please? (It should be up to them to decide if they want to answer the questions, create a rubric, etc. Give time for them to prepare as they would normally do.)

Are you ready to see the student work? (Hand student work (Attachment D))

Please evaluate these samples of student work and provide feedback in writing to the students. I will give you some time to complete the evaluation. (Give them time to evaluate the work in silence)

(Once they are done.) Could you walk me through your evaluation of each of the answers? I want to know what you were thinking as you read the answers, what caught your attention, what feedback you decided to provide, and why. (Ask clarification and exploratory questions as needed as they talk about how they approached the evaluation.)

Part 2

Now, imagine that instead of having this question as homework, it was one of the free response questions in a midterm exam. The question was assigned [8] points and you have to grade it. You must decide how to assign the points and then grade each student's answers. I will give you some time to complete the grading.

(Once they are done.) Could you walk me through your assignment of points? I want to know what you were thinking as you assigned the points and made your grading decisions. (Ask clarification and exploratory questions as needed as they talk about how they approached the evaluation.)

GRADING FIRST

Introduction

We are trying to better understand how instructors and TAs approach the assessment of student work. For that purpose, I am going to ask you to look at some samples of student work and evaluate them as you would do for any CHEM 151/152 class. Do you have any questions?

Part 1

Let's start. Imagine that students in a CHEM 151 class were given this question to answer as one of the free response questions in a midterm exam (hand empty sheet with the question (Attachment C)). The question was assigned [8] points. I will give you time to read the question.

What do you normally do to prepare for assessing student work? Could you do it now, please? (It should be up to them to decide if they want to answer the questions, create a rubric, etc. Give time for them to prepare as they would normally do.)

Are you ready to see the student work? (Hand student work (Attachment D)) I will give you some time to complete the grading.

(Once they are done.) Could you walk me through your assignment of points? I want to know what you were thinking as you assigned the points and made your grading decisions. (Ask clarification and exploratory questions as needed as they talk about how they approached the evaluation.)

Part 2

Now, imagine that instead of having this question as part of an exam, it was homework. Your next task will be to assess what [two/three] different students did and provide feedback to them. I will give you some time to complete the feedback.

Please evaluate these samples of student work and provide feedback in writing to the students. I will give you some time to complete the evaluation. (Give them time to evaluate the work in silence)

(Once they are done.) Could you walk me through your evaluation of each of the answers? I want to know what you were thinking as you read the answers, what caught your attention, what feedback you

decided to provide, and why. (Ask clarification and exploratory questions as needed as they talk about how they approached the evaluation.)

The participants may be asked the following questions:

Can you elaborate on that?

Can you tell me more about that?

Can you give me an example?

What do you mean by that?

What makes you think that?

What information led you to that conclusion?

You would say that [rephrase response] ?

Is there anything else that you would like to add?

You said [rephrase student response] can you tell me more about that?

APPENDIX D - INTERVIEW DOCUMENTS



Question 1: Evp

Pools constantly need to be replenished with water due to evaporation. One of the assumptions of the particulate model of matter is that particles exhibit a distribution of speed. Using the particulate model, explain how water evaporates and why the assumption about particles having a distribution of speed is essential to explaining evaporation?

Q1

EvP

Pools constantly need to be replenished with water due to evaporation. One of the assumptions of the particulate model of matter is that particles exhibit a distribution of speed. Using the particulate model, explain how water evaporates and why the assumption about particles having a distribution of speed is essential to explaining evaporation?

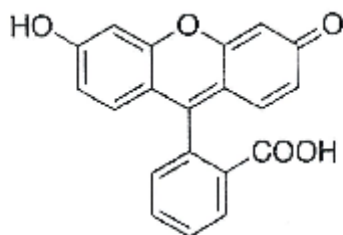
Using the particulate model of matter, water is able to evaporate due to the distribution of speeds the particles exhibit. As the surface area increases, the particles have an increased area to move around, eventually these particles are able to separate, as the weak forces between them break. As temperature of the air increases, temperature of water increases allowing them to gain potential energy, as temp increases, so does the average particle speed, as we increase the kinetic energy between the particles allowing them to go from liquid to gas and evaporate.

Particles that evaporate have greater speed than particles still in liquid phase.

Pools constantly need to be replenished with water due to evaporation. One of the assumptions of the particulate model of matter is that particles exhibit a distribution of speed. Using the particulate model, explain how water evaporates and why the assumption about particles having a distribution of speed is essential to explaining evaporation? Most people assume that all particles have the same speed, this is false. Although all particles have different speeds, as a whole, they all have the same average speed. Since some particles have greater speeds than others, they also have more configurations which allows for some particles to escape their phase. Not all particles leave/enter a phase at the same time. Some particles might change phases faster or longer than others. This is why not all water evaporates at the same time & also why your water bottle doesn't freeze in one second. (sometimes half is frozen)



Question 2: Flr



Fluorescein exhibits a lot of resonance stabilization. Build an argument for why the presence of resonance is stabilizing based on the potential and kinetic energy of the electrons.

Q2

Flr

Fluorescein exhibits a lot of resonance stabilization. Build an argument for why the presence of resonance is stabilizing based on the potential and kinetic energy of the electrons.

Resonance increases the number of bonds and decreases the number of unpaired electrons. The increase in bonds lowers the total energy of the molecule because when bonds form, energy is released. This decrease in energy increases the stability. The more unpaired electrons in a molecule, the greater the potential and kinetic energy. Since resonance decreases the number of unpaired electrons, the average kinetic and potential energy decreases, increasing the stability of the molecule.

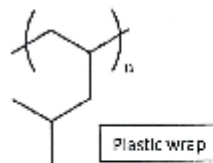
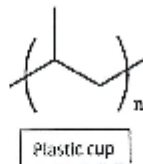
Fluorescein exhibits a lot of resonance stabilization. Build an argument for why the presence of resonance is stabilizing based on the potential and kinetic energy of the electrons.

Resonance will make a structure more stable because it shares more electrons which will delocalize them more than if they weren't shared. This delocalization will decrease the KE of the electrons due to quantum effect because they have more room. The decreased KE will make it so you have to add more energy to break the bond which is more stable. Likewise resonance is more stable because sharing the electrons makes a stronger force of attraction between the atoms. This makes the atoms closer which will lead to a decrease in PE. That way in order to separate it you will have to add extra energy which is more stable.



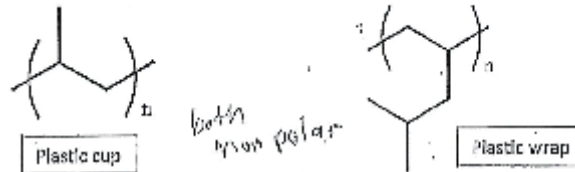
Question 3: Pol

a) Other polymers commonly used in tattooing include the plastic cups used to hold water to rinse ink off the needles, and plastic wrap used to wrap the chair and table, as well as the skin after the tattoo is complete for healing and protection. Shown below are examples of each of these polymers. Justify why the polymeric unit corresponds to the specific type of plastic. Assume the chain length for both polymers is the same.



Q3
Pb1

n) Other polymers commonly used in tattooing include the plastic cups used to hold water to rinse ink off the needles, and plastic wrap used to wrap the chair and table, as well as the skin after the tattoo is complete for healing and protection. Shown below are examples of each of these polymers. Justify why the polymeric unit corresponds to the specific type of plastic. Assume the chain length for both polymers is the same.

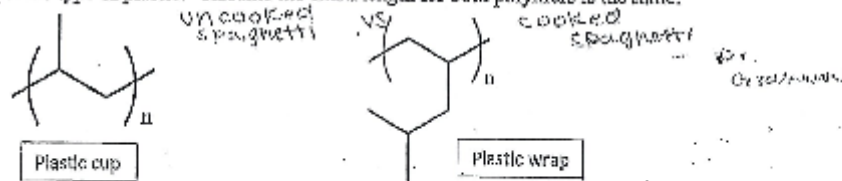


The polymer for the plastic cup only has one branching which means it's more dense than the plastic wrap. Plastic wrap has more branching which makes it less dense.

Since the plastic wrap is less dense it's more flexible and it allows us to wrap the chair and table.

Since they are both nonpolar they cannot form H-bonds so they cannot absorb water. That's why the cup can hold water.

n) Other polymers commonly used in tattooing include the plastic cups used to hold water to rinse ink off the needles, and plastic wrap used to wrap the chair and table, as well as the skin after the tattoo is complete for healing and protection. Shown below are examples of each of these polymers. Justify why the polymeric unit corresponds to the specific type of plastic. Assume the chain length for both polymers is the same.

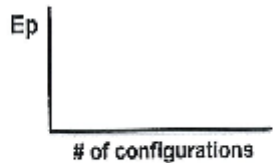


The plastic cup polymer has less branching which means it is less rigid. This makes it hard to move and is less flexible, this is good for a plastic cup because you want that polymer to be strong to hold the water. While on the other hand the plastic wrap has more branches that allow it to block the back bone and make it less rigid, thus causing it to be more flexible. This is good because it is going to be on things like skin which is in motion while the person is moving so it needs to be made up of the more branched, more flexible polymer. Also it is on the chair because again the less rigid polymer allows the flexibility of the wrap to go with the curvature of the person sitting on it.



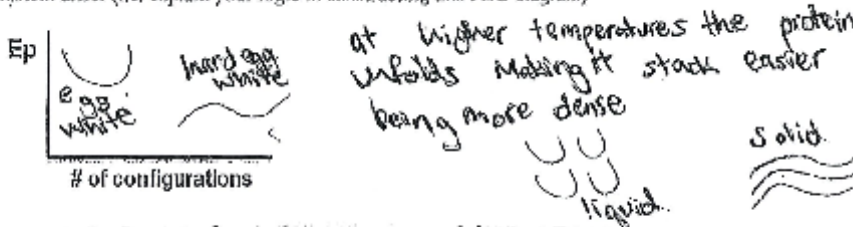
Question 4: PEC

Most proteins adopt a single unique structure when “folded” and many random structures when “unfolded”. Proteins can be unfolded (i.e. denatured) at high temperatures (ex: cooked eggs). Draw a **PEC diagram** representing a “folded” protein (F) and an “unfolded” protein (UF). Explain how the two states of the protein differ (i.e. explain your logic in constructing this PEC diagram)

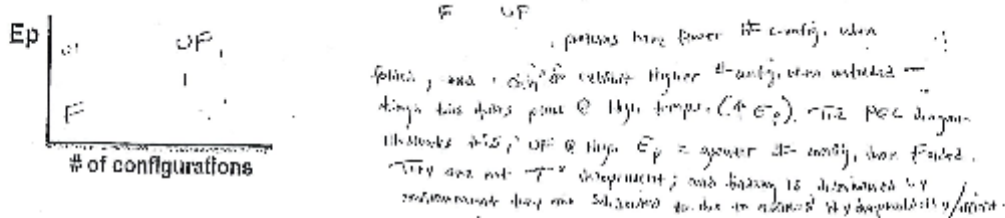


Q4
PEC

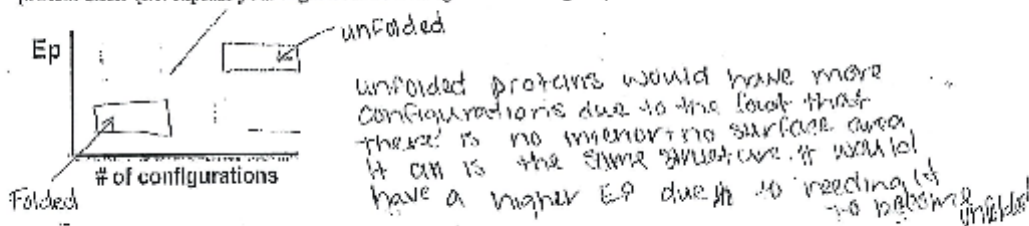
Most proteins adopt a single unique structure when "folded" and many random structures when "unfolded". Proteins can be unfolded (i.e. denatured) at high temperatures (ex: cooked eggs). Draw a PEC diagram representing a "folded" protein (F) and an "unfolded" protein (UF). Explain how the two states of the protein differ (i.e. explain your logic in constructing this PEC diagram)



Most proteins adopt a single unique structure when "folded" and many random structures when "unfolded". Proteins can be unfolded (i.e. denatured) at high temperatures (ex: cooked eggs). Draw a PEC diagram representing a "folded" protein (F) and an "unfolded" protein (UF). Explain how the two states of the protein differ (i.e. explain your logic in constructing this PEC diagram)



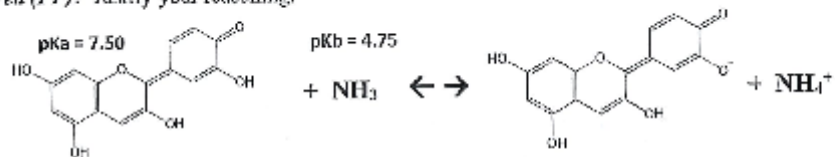
Most proteins adopt a single unique structure when "folded" and many random structures when "unfolded". Proteins can be unfolded (i.e. denatured) at high temperatures (ex: cooked eggs). Draw a PEC diagram representing a "folded" protein (F) and an "unfolded" protein (UF). Explain how the two states of the protein differ (i.e. explain your logic in constructing this PEC diagram)





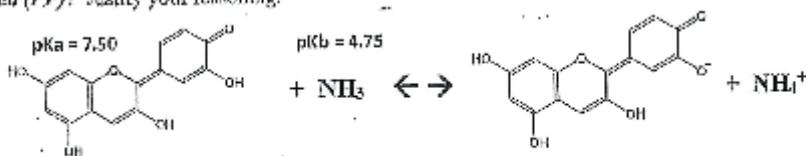
Question 5: Fav

Considering the reaction below, would you predict it to be *reactant favored (RF)* or *product favored (PF)*? Justify your reasoning.



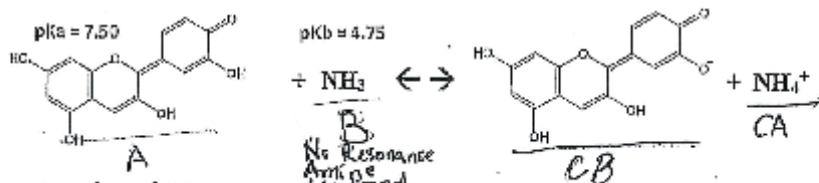
Q5
Fov

Considering the reaction below, would you predict it to be reactant favored (RF) or product favored (PF)? Justify your reasoning.



The reaction would be reactant favored. This is because the conjugate base $\text{C}_6\text{H}_5\text{O}_6^-$ is weaker than NH_3 so it is less likely to accept protons.

Considering the reaction below, would you predict it to be reactant favored (RF) or product favored (PF)? Justify your reasoning.



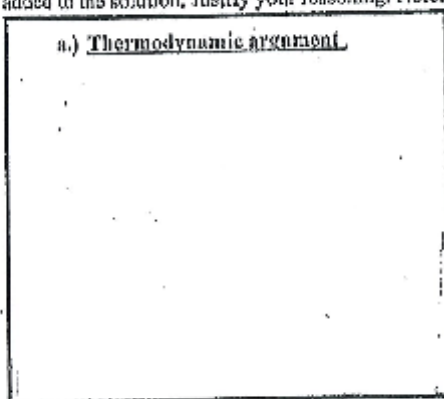
* The Base (NH_3) on the reactant side is stronger than the CB on the product side. This is because the charge density is greater for NH_3 . The charge will be more localized on the N. It also has an amine group. The charge in the CB is more delocalized and distributed. Having the weaker base in the products means that it will be PF.



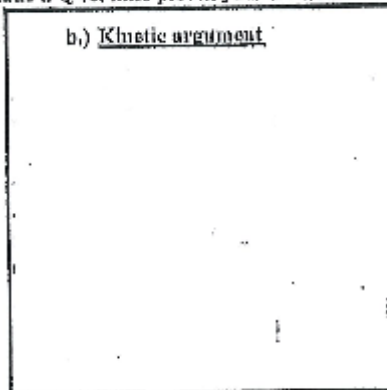
Question 6: Buf

This equilibrium is established in the buffer from question 1B: $B + H_2O \leftrightarrow BH^+ + OH^-$
Build c.) *thermodynamic* and b.) *kinetic arguments* to explain how the equilibrium would shift if OH^- is added in the solution. Justify your reasoning. Note: **Include a Q vs. time plot in your answer.**

a.) Thermodynamic argument



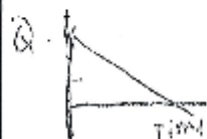
b.) Kinetic argument



Q6
Buf

This equilibrium is established in the buffer from question 18: $B + H_2O \leftrightarrow BH^+ + OH^-$
Build a.) *thermodynamic* and b.) *kinetic arguments* to explain how the equilibrium would shift if OH^- is added to the solution. Justify your reasoning. Note: Include a Q vs. time plot in your answer.

a.) Thermodynamic argument (4 pts)



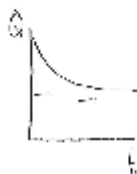
A smaller decrease but stabilize over time as more products are added and OH^- added to make more reactants

b.) Kinetic argument (4 pts)

As more products are added to the solution, the reaction accelerates to make more reactants in order to go back to equilibrium. Ex, the reaction shifts to reactants and is an endothermic reaction.

This equilibrium is established in the buffer from question 18: $B + H_2O \leftrightarrow BH^+ + OH^-$
Build a.) *thermodynamic* and b.) *kinetic arguments* to explain how the equilibrium would shift if OH^- is added to the solution. Justify your reasoning. Note: Include a Q vs. time plot in your answer.

a.) Thermodynamic argument (4 pts)



The reaction will become more thermodynamically stable on the reactant side if OH^- is added to the solution.

b.) Kinetic argument (4 pts)

OH^- is a very strong base, therefore the reaction will move to the reactant side to move away from the strong base.

APPENDIX E - INSTRUCTORS' DOMINANT APPROACHES

The Table on the next page describes the dominant approach in each dimension for each instructor, based on a 70% or higher engagement with the listed category. Those instructors who engaged with the non-dominant approach more than 30% of the time are listed as “flexible”.

Instructor	Avg Score	Avg Score (%)	#Q	Depth	Focus	Rubric	Expected Answer	Grading System	Stance	Lens	Scope	Interaction	Marking	Feedback
GSI Riley	5.31	89%	8	Skimming	Reasoning	Implicit	Unexpressed	Norm	Inferential	Adaptive	Wholistic	Expressed	Reactive	Direct
GSI Robin	5.13	85%	8	Close Reading	Reasoning	Implicit	Generated	Norm	Inferential	Prescriptive	Wholistic	Expressed	Evaluative	Direct
GSI Jackie	5.13	85%	8	Close Reading	Reasoning	Implicit	Generated	Norm	Inferential	Adaptive	Wholistic	Expressed	Reactive	Direct
GSI Dana	5.06	84%	8	Close Reading	Knowledge	Implicit	Unexpressed	Criterion	Literal	Prescriptive	Piecemeal	Unexpressed	Reactive	Direct
GSI Lou	4.50	75%	8	Close Reading	Reasoning	Implicit	Unexpressed	Norm	Literal	Prescriptive	Piecemeal	Expressed	Flexible	Direct
GSI Oakley	4.50	75%	8	Skimming	Reasoning	Implicit	Unexpressed	Criterion	Inferential	Adaptive	Wholistic	Expressed	Flexible	Flexible
GSI Nicky	4.38	73%	8	Close Reading	Knowledge	Implicit	Generated	Criterion	Literal	Prescriptive	Piecemeal	Expressed	Flexible	Direct
GSI Hollis	4.25	71%	8	Skimming	Knowledge	Implicit	Unexpressed	Criterion	Literal	Prescriptive	Flexible	Unexpressed	(none)	(none)
GSI Georgie	4.25	71%	8	Close Reading	Reasoning	Explicit	Flexible	Criterion	Flexible	Adaptive	Flexible	Expressed	Flexible	Direct
GSI Kendall	4.13	69%	8	Skimming	Reasoning	Implicit	Unexpressed	Norm	Inferential	Adaptive	Wholistic	Unexpressed	(none)	(none)
GSI Dylan	4.13	69%	8	Close Reading	Reasoning	Explicit	Unexpressed	Criterion	Literal	Prescriptive	Piecemeal	Expressed	Evaluative	Indirect
GSI Sidney	4.00	67%	8	Close Reading	Reasoning	Flexible	Generated	Flexible	Inferential	Adaptive	Wholistic	Unexpressed	(none)	(none)
GSI Alexis	3.94	66%	12	Close Reading	Reasoning	Explicit	Generated	Criterion	Inferential	Prescriptive	Wholistic	Expressed	Evaluative	Flexible
GSI Finley	3.88	65%	8	Skimming	Reasoning	Implicit	Unexpressed	Criterion	Inferential	Adaptive	Wholistic	Flexible	Reactive	Direct
FI Jamie	3.88	65%	12	Close Reading	Flexible	Implicit	Generated	Criterion	Flexible	Adaptive	Piecemeal	Expressed	Evaluative	Flexible
GSI Parker	3.83	64%	6	Close Reading	Knowledge	Flexible	Generated	Criterion	Literal	Prescriptive	Piecemeal	Expressed	Evaluative	Flexible
GSI Taylor	3.63	60%	8	Close Reading	Reasoning	Explicit	Generated	Criterion	Literal	Prescriptive	Piecemeal	Expressed	Evaluative	Direct
GSI Pat	3.50	58%	8	Close Reading	Reasoning	Flexible	Generated	Criterion	Literal	Adaptive	Wholistic	Expressed	Reactive	Flexible
GSI Ollie	3.50	58%	8	Close Reading	Knowledge	Implicit	Unexpressed	Criterion	Inferential	Adaptive	Piecemeal	Expressed	Evaluative	Direct
GSI Frankie	3.42	57%	6	Close Reading	Reasoning	Implicit	Generated	Criterion	Inferential	Prescriptive	Wholistic	Unexpressed	(none)	(none)
FI Addison	3.35	56%	12	Close Reading	Flexible	Explicit	Generated	Criterion	Inferential	Adaptive	Piecemeal	Expressed	Evaluative	Direct
FI Logan	3.23	54%	12	Close Reading	Flexible	Explicit	Generated	Criterion	Flexible	Adaptive	Flexible	Expressed	Flexible	Direct
GSI Morgan	2.98	50%	12	Skimming	Reasoning	Implicit	Flexible	Criterion	Literal	Adaptive	Flexible	Expressed	Evaluative	Flexible
FI Shiloh	2.94	49%	12	Skimming	Reasoning	Implicit	Unexpressed	Flexible	Inferential	Adaptive	Flexible	Flexible	Evaluative	Direct
FI Elliot	2.85	48%	12	Skimming	Flexible	Explicit	Generated	Criterion	Literal	Flexible	Flexible	Flexible	Evaluative	Direct
GSI Emerson	2.56	43%	8	Close Reading	Flexible	Explicit	Generated	Criterion	Inferential	Adaptive	Wholistic	Expressed	Evaluative	Direct
GSI Micah	2.52	42%	12	Close Reading	Flexible	Flexible	Generated	Criterion	Literal	Adaptive	Flexible	Expressed	Flexible	Direct
FI Quinn	2.42	40%	12	Skimming	Reasoning	Implicit	Unexpressed	Norm	Literal	Flexible	Wholistic	Flexible	Evaluative	Direct
FI Blair	2.33	39%	12	Close Reading	Flexible	Explicit	Generated	Criterion	Literal	Prescriptive	Piecemeal	Expressed	Evaluative	Direct

APPENDIX F – PUBLICATION PERMISSIONS



RightsLink®



Dimensions of Variation in Chemistry Instructors' Approaches to the Evaluation and Grading of Student Responses



Author: Michelle Herridge, Vicente Talanquer

Publication: Journal of Chemical Education

Publisher: American Chemical Society

Date: Feb 1, 2021

Copyright © 2021, American Chemical Society

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

[BACK](#)

[CLOSE WINDOW](#)

© 2021 Copyright - All Rights Reserved | [Copyright Clearance Center, Inc.](#) | [Privacy statement](#) | [Terms and Conditions](#)
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

[PUBLICATION PERMISSION FROM CERP PENDING ACCEPTANCE/PUBLICATION]

COMPLETE LIST OF REFERENCES

- Abell, S. K., & Siegel, M. A. (2011). Assessment literacy: What science teachers need to know and be able to do. In *The Professional Knowledge Base of Science Teaching*. https://doi.org/10.1007/978-90-481-3927-9_12
- Adie, L. E., Klenowski, V., & Wyatt-Smith, C. (2012). Towards an understanding of teacher judgement in the context of social moderation. *Educational Review*, 64(2), 223–240. <https://doi.org/10.1080/00131911.2011.598919>
- Ainley, J., & Luntley, M. (2007). The role of attention in expert classroom practice. *Journal of Mathematics Teacher Education*, 10(1), 3–22. <https://doi.org/10.1007/s10857-007-9026-z>
- Allen, S., & Knight, J. (2009). A Method for Collaboratively Developing and Validating a Rubric A Method for Collaboratively Developing and Validating a Rubric. *International Journal for the Scholarship of Teaching and Learning*, 3(2). <https://doi.org/10.20429/ijstl.2009.030210>
- Amador, J. M., Bragelman, J., & Castro Superfine, A. (2021). Prospective teachers ' noticing : A literature review of methodological approaches to support and analyze noticing. *Teaching and Teacher Education*, 99. <https://doi.org/10.1016/j.tate.2020.103256>
- Amador, J. M., Estapa, A., Araujo, Z. de, Kosko, K. W., & Weston, T. L. (2017). Eliciting and Analyzing Preservice Teachers' Mathematical Noticing. *Mathematics Teacher Educator*, 5(2), 158–177. <https://doi.org/10.5951/mathteaceduc.5.2.0158>
- Andrade, H. G. (1997). Understanding rubrics. *Educational Leadership*, 54(4), 14–17. <http://learnweb.harvard.edu/ALPS/Thinking/docs/rubricar.pdf>
- Andrade, H. G. (2005). Teaching With Rubrics: The Good, the Bad, and the Ugly. *College Teaching*, 53(1), 27–31. <https://doi.org/10.3200/CTCH.53.1.27-31>
- Aydeniz, M., & Dogan, A. (2016). Exploring pre-service science teachers' pedagogical capacity

- for formative assessment through analyses of student answers. *Research in Science and Technological Education*, 34(2), 125–141. <https://doi.org/10.1080/02635143.2015.1092954>
- Baines, L. A., & Stanley, G. (2000). “We want to see the teacher”. Constructivism and the rage against expertise. *Phi Delta Kappan*, 82(4), 327–330. <https://doi.org/10.1177/003172170008200422>
- Barnhart, T., & Van Es, E. A. (2015). Studying teacher noticing: EXAMINING the relationship among pre-service science teachers’ ability to attend, analyze and respond to student thinking. *Teaching and Teacher Education*, 45, 83–93. <https://doi.org/10.1016/j.tate.2014.09.005>
- Braund, H., & DeLuca, C. (2018). Elementary students as active agents in their learning : an empirical study of the connections between assessment practices and student. *The Australian Educational Researcher*, 45(1), 65–85. <https://doi.org/10.1007/s13384-018-0265-z>
- Brooker, R., Muller, R., Mylonas, A., & Hansford, B. (1998). Improving the assessment of practice teaching: A criteria and standards framework. *Assessment and Evaluation in Higher Education*, 23(1), 5–24. <https://doi.org/10.1080/0260293980230101>
- 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>
- Brown, B., Eaton, S., Jacobsen, M., Roy, S., & Friesen, S. (2013). Instructional Design Collaboration: A Professional Learning and Growth Experience. *Journal of Online Learning and Teaching*, 9(3), 439. <https://doi.org/10.11575/PRISM/34910>
- Chan, K. K. H., Xu, L., Cooper, R., Berry, A., & van Driel, J. H. (2020). Teacher noticing in science education: do you see what I see? *Studies in Science Education*, 00(00), 1–44.

<https://doi.org/10.1080/03057267.2020.1755803>

Close, D. (2009). Fair grades. *Teaching Philosophy*, 32(4), 361–398.

<https://doi.org/10.5840/teachphil200932439>

Cohen, J. (1977). F-Tests of Variance Proportions in Multiple Regression/Correlation Analysis. In *Statistical Power Analysis for the Behavioral Sciences* (pp. 407–453).

Connor, M., Gere, A. R., & Shultz, G. V. (2019). Characterizing Peer Review Comments and Revision from a Writing- to-Learn Assignment Focused on Lewis Structures. *Journal of Chemical Education*, 96, 227–237. <https://doi.org/10.1021/acs.jchemed.8b00711>

Cooper, M. M., & Sandi-Urena, S. (2009). Design and validation of an instrument to assess metacognitive skillfulness in chemistry problem solving. *Journal of Chemical Education*, 86(2), 240–245. <https://doi.org/10.1021/ed086p240>

Cowie, B., Harrison, C., & Willis, J. (2018). Supporting teacher responsiveness in assessment for learning through disciplined noticing. *Curriculum Journal*, 29(4), 464–478. <https://doi.org/10.1080/09585176.2018.1481442>

Crespo, S. (2000). Seeing More Than Right and Wrong Answers: Prospective Teachers' Interpretations of Students' Mathematical Work. *Journal of Mathematics Teacher Education*. <https://doi.org/10.1023/A:1009999016764>

Creswell, J. W. (2012). Educational research: Planning, conducting, and evaluating quantitative and qualitative research. In *Educational Research*.

Cullingford, C. (Ed.). (1997). *Assessment Versus Evaluation*. Cassell.

Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment and Evaluation in Higher Education*, 42(3), 347–360. <https://doi.org/10.1080/02602938.2015.1111294>

- Dochy, F., Segers, M., & Buehl, M. M. (1999). The Relation Between Assessment Practices and Outcomes of Studies : The Case of Research on Prior Knowledge. *Review of Educational Research, 69*(2), 145–186.
- Duncan, C. R., & Noonan, B. (2007). Factors Affecting Teachers ' Grading and Assessment Practices. *The Alberta Journal of Educational Research, 53*(1), 1–21.
- Ebby, C. B., Remillard, J., & D'Olier, J. (2019). *Pathways for Analyzing and Responding to Student Work for Formative Assessment : The Role of Teachers ' Goals for Student Learning Pathways for Analyzing and Responding to Student Work for Formative.*
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. In *Language Testing* (Vol. 25, Issue 2). <https://doi.org/10.1177/0265532207086780>
- Falchikov, N., & Goldfinch, J. (2000). Student Peer Assessment in Higher Education : A Meta-Analysis Comparing Peer and Teacher Marks. *Review of Educational Research, 70*(3), 287–322.
- Fay, M. E., Grove, N. P., Towns, M. H., Bretz, S. L., & Lowery Bretz, S. (2007). A rubric to characterize inquiry in the undergraduate chemistry laboratory. *Chemistry Education Research and Practice, 8*(2), 212–219. <https://doi.org/10.1039/B6RP90031C>
- Furtak, E. M., & Thompson, J. (2016). Formative Assessment and Noticing: Toward a Synthesized Framework for Attending and Responding During Instruction. *Annual Meeting of the American Educational Research Association, April, 1–15.* <https://doi.org/10.13140/RG.2.1.3471.1443>
- Gardner, G. E., & Jones, M. G. (2011). Pedagogical Preparation of the Science Graduate Teaching Assistant: Challenges and Implications. *Science Educator.*

- Gess-Newsome, J., Southerland, S. A., Johnston, A., & Woodbury, S. (2003). Educational Reform, Personal Practical Theories, and Dissatisfaction: The Anatomy of Change in College Science Teaching. *American Educational Research Journal*, 40(3), 731–767. <https://doi.org/10.3102/00028312040003731>
- Gibbons, R. E., Villafañe, S. M., Stains, M., Murphy, K. L., & Raker, J. R. (2018). Beliefs about learning and enacted instructional practices: An investigation in postsecondary chemistry education. *Journal of Research in Science Teaching*, 55(8), 1111–1133. <https://doi.org/10.1002/tea.21444>
- Glazer, N. (2014). Formative Plus Summative Assessment in Large Undergraduate Courses: Why Both? *International Journal of Teaching and Learning in Higher Education*, 26(2), 276–286. <http://www.isetl.org/ijtlhe/>
- Gotwals, A. W., & Birmingham, D. (2016). Eliciting, Identifying, Interpreting, and Responding to Students' Ideas: Teacher Candidates' Growth in Formative Assessment Practices. *Research in Science Education*, 46(3), 365–388. <https://doi.org/10.1007/s11165-015-9461-2>
- Guerrero, M., & Rod, A. B. (2013). Engaging in Office Hours : A Study of Student- Faculty Interaction and Academic Performance. *Journal of Political Science Education*, 9(4), 403–416. <https://doi.org/10.1080/15512169.2013.835554>
- Haagen, C. H. (1964). The Origins of a Grade. *The Journal of Higher Education*, 35(2), 89–91. <https://doi.org/10.2307/1979782>
- Harlen, W., & James, M. (2006). Assessment and Learning : differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy and Practice*, 4(3), 365–379. <https://doi.org/10.1080/0969594970040304>
- Harshman, J., & Yeziarski, E. J. (2016). Characterizing high school chemistry teachers' use of

- assessment data: Via latent class analysis. *Chemistry Education Research and Practice*, 17(2), 296–308. <https://doi.org/10.1039/c5rp00215j>
- Harwood, C. J., Hewett, S., & Towns, M. H. (2020). Rubrics for Assessing Hands-On Laboratory Skills. *Journal of Chemical Education*. <https://doi.org/10.1021/acs.jchemed.0c00200>
- 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*. <https://doi.org/10.1002/tea.20439>
- Henderson, C., Yerushalmi, E., Kuo, V. H., Heller, P., & Heller, K. (2004). Grading student problem solutions: The challenge of sending a consistent message. *American Journal of Physics*, 72(2), 164–169. <https://doi.org/10.1119/1.1634963>
- 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. (n.d.). Variation in Chemistry Instructors' Approaches to the Evaluation of Student Written Responses and Its Impact on Grading. *Chemistry Education Research and Practice*.
- Holmes, L. E., & Smith, L. J. (2003). Student Evaluations of Faculty Grading Methods. *Journal of Education for Business*, 78(6), 318–323. <https://doi.org/10.1080/08832320309598620>
- Howell, R. J. (2011). Exploring the impact of grading rubrics on academic performance. *Journal of Excellence in College Teaching*, 22(2), 31–49.
- Huang, R., & Li, Y. (2012). What Matters Most: A Comparison of Expert and Novice Teachers' Noticing of Mathematics Classroom Events. *School Science and Mathematics*, 112(7), 420–432. <https://doi.org/10.1111/j.1949-8594.2012.00161.x>

- Jacobs, V. R., Lamb, L. L. C., & Philipp, R. A. (2010). Professional noticing of children's mathematical thinking. *Journal for Research in Mathematics Education*. <https://doi.org/10.5951/jresematheduc.41.2.0169>
- Kim, H.-Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, 38(1), 52–54. <https://doi.org/10.5395/rde.2013.38.1.52>
- Klobucar, A., Deane, P., Elliot, N., Ramineni, C., Deess, P., & Rudniy, A. (2012). Automated essay scoring and the search for valid writing assessment. *International Advances in Writing Research: Cultures, Places, Measures*.
- Knight, J., Allen, S., & Mitchell, A. M. (2012). Establishing Consistency Measurements of Grading for Multiple Section Courses. *Journal of the Academy of Business Education*, 13, 28–47. <https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=85607939&site=ehost-live>
- Kohn, A. (2006). Speaking My Mind: The Trouble with Rubrics. *English Journal*, 95(4), 12. <https://doi.org/10.2307/30047080>
- Kulick, G., & Wright, R. (2008). The Impact of Grading on the Curve: A Simulation Analysis. *International Journal for the Scholarship of Teaching and Learning*, 2(2). <https://doi.org/10.20429/ijstol.2008.020205>
- Li, C. H., & Zafar, B. (2020). Ask and You Shall Receive? Gender Differences in Regrades in College. *Working Paper*, 1–45. <http://www.nber.org/papers/w26703>
- Logan, K., & Mountain, L. (2018). Writing Instruction in Chemistry Classes: Developing Prompts and Rubrics. *Journal of Chemical Education*. <https://doi.org/10.1021/acs.jchemed.8b00294>

- Luna, M. J. (2018). What Does it Mean to Notice my Students' Ideas in Science Today?: An Investigation of Elementary Teachers' Practice of Noticing their Students' Thinking in Science. *Cognition and Instruction*, 36(4), 297–329. <https://doi.org/10.1080/07370008.2018.1496919>
- Marbach-Ad, G., Schaefer, K. L., Kumi, B. C., Friedman, L. A., Thompson, K. V., & Doyle, M. P. (2012). Development and evaluation of a prep course for chemistry graduate teaching assistants at a research university. *Journal of Chemical Education*. <https://doi.org/10.1021/ed200563b>
- Marshman, E., Sayer, R., Henderson, C., & Singh, C. (2017). Contrasting grading approaches in introductory physics and quantum mechanics: The case of graduate teaching assistants. *Physical Review Physics Education Research*, 13(1), 1–16. <https://doi.org/10.1103/PhysRevPhysEducRes.13.010120>
- Marshman, E., Sayer, R., Henderson, C., Yerushalmi, E., & Singh, C. (2018). The challenges of changing teaching assistants' grading practices: Requiring students to show evidence of understanding. *Canadian Journal of Physics*, 96(4), 420–437. <https://doi.org/10.1139/cjp-2017-0030>
- Maxwell, J., & Miller, B. (2008). Categorizing and Connecting Strategies in Qualitative Data Analysis. *Handbook of Emergent Methods*.
- Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research and Evaluation*, 7(25), 2000–2001.
- Miles, M. B., & Huberman, A. . (1994). An expanded sourcebook: Qualitative data analysis (2nd Edition). In *Sage Publications*.
- Mohl, E., Fifield, C., Lafond, N., Mickman, S., Saxton, R., & Smith, B. (2017). Using Rubrics to

- Integrate Crosscutting Concepts. *Science Scope*, 40(5), 84.
- Moskal, B. M. (2000). Scoring rubrics: what, when and how?. Moskal, Barbara M. *Practical Assessment, Research & Evaluation*, 7(3), 3–7. <http://pareonline.net/getvn.asp?v=7&n=3>
- Murray, S. A., Huie, R., Lewis, R., Balicki, S., Clinchot, M., Banks, G., Talanquer, V., & Sevian, H. (2020). Teachers' Noticing, Interpreting, and Acting on Students' Chemical Ideas in Written Work. *Journal of Chemical Education*, 97(10), 3478–3489. <https://doi.org/10.1021/acs.jchemed.9b01198>
- Mutambuki, J. M., & Fynewever, H. (2012). Comparing chemistry faculty beliefs about grading with grading practices. *Journal of Chemical Education*, 89(3), 326–334. <https://doi.org/10.1021/ed1000284>
- Mutambuki, J. M., & Schwartz, R. (2018). We don't get any training: the impact of a professional development model on teaching practices of chemistry and biology graduate teaching assistants. *Chemistry Education Research and Practice*, 19(1), 106–121. <https://doi.org/10.1039/C7RP00133A>
- Nielsen, S. E., & Yeziarski, E. J. (2015). Exploring the Structure and Function of the Chemistry Self-Concept Inventory with High School Chemistry Students. *Journal of Chemical Education*, 92(11), 1782–1789. <https://doi.org/10.1021/acs.jchemed.5b00302>
- Panadero, E., Fraile, J., Ruiz, J. F., Castilla-estévez, D., & Ruiz, M. A. (2019). Spanish university assessment practices: examination tradition with diversity by faculty. *Assessment & Evaluation in Higher Education*, 44(3), 379–397. <https://doi.org/10.1080/02602938.2018.1512553>
- 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. <https://doi.org/10.1007/s11165-011-9268-8>
- Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, 48(1), 85–112. <https://doi.org/10.1016/j.jsp.2009.09.002>
- 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. <https://doi.org/10.1039/c9rp00217k>
- Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*.
- Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education*, 26(7), 1372–1380. <https://doi.org/10.1016/j.tate.2010.03.008>
- Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Sage.
- Rea, J. (2010). *You Say Ee-ther and I Say Eye-ther : Clarifying Assessment and Evaluation*.
- Reynders, G., Lantz, J., Ruder, S. M., Stanford, C. L., & Cole, R. S. (2020). Rubrics to assess critical thinking and information processing in undergraduate STEM courses. *International Journal of STEM Education*. <https://doi.org/10.1186/s40594-020-00208-5>
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18–39. <https://doi.org/10.1016/j.asw.2010.01.003>
- Ross, P., & Gibson, S. A. (2010). Exploring a conceptual framework for expert noticing during literacy instruction. *Literacy Research and Instruction*, 49(2), 175–193. <https://doi.org/10.1080/19388070902923221>
- Russ, R. S., Coffey, J. E., Hammer, D., & Hutchison, P. (2009). Making classroom assessment

- more accountable to scientific reasoning: A case for attending to mechanistic thinking. *Science Education*. <https://doi.org/10.1002/sce.20320>
- Russ, R. S., & Luna, M. J. (2013). Inferring teacher epistemological framing from local patterns in teacher noticing. *Journal of Research in Science Teaching*, *50*(3), 284–314. <https://doi.org/10.1002/tea.21063>
- Sabel, J. L., Forbes, C. T., & Flynn, L. (2016). Elementary teachers' use of content knowledge to evaluate students' thinking in the life sciences. *International Journal of Science Education*, *38*(7), 1077–1099. <https://doi.org/10.1080/09500693.2016.1183179>
- Sadler, Royce, D. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144. [http://michiganassessmentconsortium.org/sites/default/files/Formative Assessment and Design of Instructional Systems.pdf](http://michiganassessmentconsortium.org/sites/default/files/Formative%20Assessment%20and%20Design%20of%20Instructional%20Systems.pdf)
- Salehi, S., Cotner, S., Azarin, S. M., Carlson, E. E., Driessen, M., Ferry, V. E., Harcombe, W., McGaugh, S., Wassenberg, D., Yonas, A., & Ballen, C. J. (2019). Gender Performance Gaps Across Different Assessment Methods and the Underlying Mechanisms: The Case of Incoming Preparation and Test Anxiety. *Frontiers in Education*. <https://doi.org/10.3389/educ.2019.00107>
- Satria, I., Nursyirwan, Wardana, & Mustamin, A. A. Bin. (2020). The Sociology of Education and Bell Curve Controversy: How Should Grades Be Calculated? *Palarch's Journal Of Archaeology Of Egypt/Egyptology*, *17*(7), 12431–12447.
- Schwab, K., Moseley, B., & Dustin, D. (2018). Grading Grades as a Measure of Student Learning. *SCHOLE: A Journal of Leisure Studies and Recreation Education*, *33*(2), 87–95. <https://doi.org/10.1080/1937156X.2018.1513276>

- Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A practical guide to calculating Cohen's f^2 , a measure of local effect size, from PROC MIXED. *Frontiers in Psychology*, 3. <https://doi.org/10.3389/fpsyg.2012.00111>
- Sherin, M. G., Jacobs, V. R., & Philipp, R. A. (2011). Mathematics teacher noticing seeing through teachers' eyes. In *Mathematics Teacher Noticing Seeing Through Teachers' Eyes*. <https://doi.org/10.4324/9780203832714>
- Stacy, A. (2000). The Graduate Student in the Dual Role of Student and Teacher. In *Graduate Education in the Chemical Sciences: Issues for the 21st Century: Report of a Workshop*. <https://www.ncbi.nlm.nih.gov/books/NBK44904/>
- Stanley, G., & Baines, L. A. (2001). No More Shopping for Grades at B-mart: Re-establishing Grades as Indicators of Academic Performance. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 74(4), 227–230. <https://doi.org/10.1080/00098650109599197>
- Starch, D., & Elliott, E. C. (1912). *Reliability of the Grading of High-School Work in English*. 20(7), 442–457. <http://www.jstor.org/stable/1076706>
- Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P. (2009). An Assessment of Reliability and Validity of a Rubric for Grading APA-Style Introductions. *Teaching of Psychology*, 36(2), 102–107. <https://doi.org/10.1080/00986280902739776>
- Stowe, R. L., & Cooper, M. M. (2019). Assessment in Chemistry Education. *Israel Journal of Chemistry*, 59(6), 598–607. <https://doi.org/10.1002/ijch.201900024>
- Talanquer, V., Bolger, M., & Tomanek, D. (2015). Exploring prospective teachers' assessment practices: Noticing and interpreting student understanding in the assessment of written work.

- Journal of Research in Science Teaching*, 52(5), 585–609. <https://doi.org/10.1002/tea.21209>
- Talanquer, V., & Pollard, J. (2010). Let's teach how we think instead of what we know. *Chemistry Education Research and Practice*, 11(2), 74–83. <https://doi.org/10.1039/c005349j>
- Talanquer, V., Tomanek, D., & Novodvorsky, I. (2013). Assessing students' understanding of inquiry: What do prospective science teachers notice? *Journal of Research in Science Teaching*, 50(2), 189–208. <https://doi.org/10.1002/tea.21074>
- Tan, L., Ling, Y., Yuen, B., Lam, P., Loo, W. L., Prinsloo, C., & Gan, M. (2020). Students' Conceptions of Bell Curve Grading Fairness in Relation to Goal Orientation and Motivation. *International Journal for the Scholarship of Teaching and Learning*, 14(1), 1–9.
- Taras, M. (2005). ASSESSMENT – SUMMATIVE AND FORMATIVE – SOME THEORETICAL REFLECTIONS. *British Journal of Educational Studies*, 53(4), 466–478. <https://doi.org/10.1111/j.1467-8527.2005.00307.x>
- Tashiro, J., Parga, D., Pollard, J., & Talanquer, V. (2021). Investigating factors that affect change in students' self-assessment of understanding when engaged in in-class tasks. *Chemistry Education Research and Practice*.
- Van Es, E. A. (2011). A framework for learning to notice student thinking. In *Mathematics Teacher Noticing Seeing Through Teachers' Eyes*. <https://doi.org/10.4324/9780203832714>
- Warren, J. (1971). College Grading Practices: An Overview. *ETS Research Bulletin Series, March 1971*, 0–3. <https://doi.org/10.1002/j.2333-8504.1971.tb00423.x>
- Wickham, H., & Henry, L. (2018). tidy: Easily Tidy Data with “spread()” and “gather()” Functions. *R Package Version 0.8.0*. <https://CRAN.R-Project.Org/Package=tidy>.
- Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, 15(4), 465–492.

<https://doi.org/10.1177/0741088398015004002>

Yerushalmi, E., Sayer, R., Marshman, E., Henderson, C., & Singh, C. (2016). *Physics graduate teaching assistants' beliefs about a grading rubric: Lessons learned*. 408–411.

<https://doi.org/10.1119/perc.2016.pr.097>