

**A VARIABILITY ANALYSIS OF GRADING OPEN-ENDED
TASKS WITH RUBRICS ACROSS MANY GRADERS**

by

Nathan M. Hicks

Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy



School of Engineering Education

West Lafayette, Indiana

August 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Kerrie A. Douglas, Co-Chair

School of Engineering Education

Dr. Heidi A. Diefes-Dux, Co-Chair

Department of Biological Systems Engineering, University of Nebraska-Lincoln

Dr. Edward J. Berger

School of Engineering Education

Dr. Charles M. Krousgrill

School of Mechanical Engineering

Dr. Brian D. Gane

Learning Sciences Research Institute, University of Illinois-Chicago

Approved by:

Dr. Donna Riley

This dissertation is dedicated to my friends and family who have always provided me with the love and support I have needed.

ACKNOWLEDGMENTS

This work would not have been possible without the many forms of support provided to me over the years by all my friends, family, and colleagues. It is said that it takes a village to raise a child but the same can be said for dissertation research, including the many student teaching assistants who provided the data that was analyzed for this research, my undergraduate research assistant, Mark Legalt, my fellow Engineering Education students and my research group.

I would like to express my deepest gratitude to my research advisers, Dr. Heidi Diefes-Dux and Dr. Kerrie Douglas. They always provided me with the directions I needed and trusted in my vision and judgments about my work. Their ability to push me to produce my best work in a timely fashion while being empathetic, kind, and nurturing will inform the way I handle my own students in the future. I would like to express my appreciation for all my committee members, Drs. Brian Gane, Ed Berger, and Chuck Krousgrill for their kind support along the way. Further, I would like to thank the Purdue Engineering Education community for creating such a wonderful and supportive environment. While the entire faculty and staff were always amazing and delightful, I would like to specific people with whom I had the most frequent and or meaningful interactions: Dr. Senay Purzer, Dr. Tamara Moore, Dr. Allison Godwin, Dr. Audeen Fentiman, Dr. Matt Ohland, Dr. Alice Pawley, Dr. Ruth Streveler, Dr. Michael Loui, Carol Brock, Teresa Walker, and Loretta McKinnis, all of whom contributed to my professional knowledge, experience, and my ability to navigate the program. I am also extremely thankful for the amazing teaching and course development experiences I had along the way. Working with James Whitford, Anne Delion, Dianne Bell, Jill Folkerts, Rick Womack, Nicole Towner, Dr. Isabel Jimenez-Useche, Dr. Matilde Sanchez-Pena, and Dr. Jeannete Aguilar were all great experiences. Finally, I want to thank Dr. Karl Smith for the invaluable experience of teaching with and learning from him and for his continued mentorship along the way.

I am also extremely grateful for the love and support from my friends and family. I will forever be grateful to Kaila Ames for her understanding and support throughout my entire graduate school career, for always looking out for me and all our wonderful pets, even when I did not deserve it. Further, I would like to thank all the wonderful pets for always providing the love and pick-me-ups that I needed, including Karma's contribution of the following snippet of text to this dissertation: "bh8lop66734ytkj-jklgtg."

TABLE OF CONTENTS

LIST OF TABLES	10
LIST OF FIGURES	13
LIST OF ABBREVIATIONS.....	16
ABSTRACT	17
1. INTRODUCTION	18
1.1 Validity	19
1.2 Fairness and Reliability as Primary Contextual Challenges to Grading Validity.....	20
1.2.1 Faculty understanding of assessment validity and goals.....	20
1.2.2 Trends toward the use of open-ended assessment items	21
1.2.3 Difficulties with increasing scale	22
1.3 Consequences of Invalid Grading	23
1.3.1 Consequences for students.....	24
1.3.2 Broader consequences	24
1.4 Means to Improve the Validity of Grading	25
1.5 Grading Open-Ended Problems in Large University Courses.....	27
1.6 Research Questions	28
1.7 Summary.....	30
2. CONCEPTUAL FRAMEWORK AND LITERATURE REVIEW	33
2.1 Assessment as a System	33
2.2 Aspects Associated with Grading Documents.....	35
2.2.1 Assignments and student work	36
2.2.2 Schemes and tools for grading.....	38
Rubrics	40
Rubric design	42
2.3 Aspects Associated with Graders.....	44
2.3.1 Grader training	46
2.3.2 Cognitive strategies in grading	47
2.4 Variability of Human Performance.....	49
2.4.1 Cognitive demand	51

2.4.2	Cognitive load	52
2.4.3	Decision making.....	53
2.5	Methods to Analyze Variability in Socio-Technical Systems.....	55
2.5.1	Measuring variability in grading	57
2.6	Synthesis and Summary	58
3.	METHODS	61
3.1	Research Design.....	61
3.2	Context	62
3.2.1	Course details.....	62
3.2.2	Grading	63
3.2.3	Rubrics.....	63
3.2.4	Training	64
3.3	Study Participants and Data Sources.....	68
3.4	Detailed Design.....	69
3.4.1	Stage 0: Think-aloud interviews	69
	Stage 0.1: Document design.....	71
	Stage 0.2: Pilot interviews.....	72
	Stage 0.3: Revised document design	73
	Stage 0.4: Data collection.....	73
3.4.2	Stage 1: Model Development.....	73
	Stage 1.1: Function identification and description (FRAM step 1).....	74
	Stage 1.1a: Interview analysis.....	76
	Stage 1.1b: Document analysis	77
	Function definitions and abstraction hierarchies.....	77
	Stage 1.2: Variability identification (FRAM step 2)	79
	Stage 1.3: Variability aggregation (FRAM step 3).....	80
3.4.3	Stage 2: Model Instantiations.....	81
	Stage 2.1: Work-as-imagined instantiations.....	81
	Stage 2.2: Work-as-completed instantiations.....	83
	Stages 2.3 and 2.4: Instantiation analysis and system variability analysis	83
3.4.4	Stage 3: Control mechanism identification (FRAM step 4)	84

3.5	Ecological Validity.....	85
4.	THE FRAM MODEL.....	89
4.1	Functional Purpose Level	90
4.1.1	IST's functional purpose	92
4.1.2	Teaching team's functional purpose	94
4.1.3	Students' functional purpose	95
4.1.4	Graders' functional purpose.....	96
4.2	Generalized Function Level.....	97
4.2.1	IST's generalized functions	100
4.2.2	Teaching team's generalized functions	103
4.2.3	Students' generalized functions	105
4.2.4	Graders' generalized functions	106
4.3	Cognitive Function Level	108
4.3.1	IST's cognitive functions.....	112
	Cognitive functions associated with developing course content	112
	Cognitive functions involved in setting schedules and deadlines	117
	Cognitive functions involved in developing class sessions	120
	Cognitive functions involved in designing assignment task	123
	Cognitive functions associated with designing grading guidelines.....	128
	Cognitive functions associated with designing grader training.....	132
4.3.2	Teaching team cognitive functions	137
	Cognitive functions associated with delivering course content	139
	Cognitive functions involved in guiding student practice	141
4.3.3	Student cognitive functions	144
	Cognitive functions associated with learning course content	145
	Cognitive functions involved in performing the assigned task	146
4.3.4	Graders' cognitive functions.....	148
	Cognitive functions associated with training to calibrate grading decisions	149
	Cognitive functions associated with preparing to evaluate task performance	156
	Cognitive functions associated with evaluating task performance.....	159
	Cognitive functions associated with scoring decisions.....	167

Additional cognitive functions observed during grading.....	170
5. MODEL INSTANTIATIONS	175
5.1 Guidance for Interpreting Findings	176
5.2 Problem Descriptions	177
5.3 IST Functions.....	178
5.3.1 Content creation	178
5.3.2 Assessment task design	187
5.3.3 Grading guideline design.....	190
5.3.4 Grader training design	195
5.4 Teaching Team and Student Functions	197
5.4.1 Task performance.....	197
5.5 Grader Functions.....	201
5.5.1 Training	201
5.5.2 Evaluation preparation.....	203
5.5.3 Evaluation of task performance	203
5.5.4 Scoring decisions	212
5.6 Function Variability and Aggregation.....	214
6. DISCUSSION.....	220
6.1 Extensions Beyond Previous Models.....	221
6.1.1 Grader cognition in context	221
6.1.2 Cognition outside of grading	222
6.1.3 Additional cognitive functions and nuance	223
6.1.4 Connections between functions	228
6.2 Impactful Variables.....	229
6.2.1 Learning objectives	230
6.2.2 Assigned tasks.....	232
6.2.3 Grading guidelines	234
6.2.4 Student work	235
6.3 Differences Between Imagined and Completed Grading.....	236
6.4 System Resilience	240
6.5 Dampening mechanisms.....	242

6.5.1	Course content.....	242
6.5.2	Assignments	243
6.5.3	Grading guidelines	244
6.5.4	Grader training	245
6.5.5	General organizational policies.....	247
6.6	Generalizability of Findings and Recommendations	249
7.	CONCLUSION.....	253
7.1	Overview	253
7.2	Major Contributions	256
7.3	Practical Implications and Recommendations	257
7.4	Future Work.....	258
	APPENDIX A. RECRUITMENT EMAIL	260
	APPENDIX B. OPEN CODING SAMPLE.....	261
	APPENDIX C. INITIAL FOCUS CODING.....	264
	APPENDIX D. FINAL FOCUSED CODES	266
	APPENDIX E. WORK-AS-IMAGINED CODING EXAMPLE	270
	APPENDIX F. WORK-AS-COMPLETED CODING EXAMPLE	271
	APPENDIX G. RUBRICS	273
	APPENDIX H. TASK PERFORMANCES	283
	REFERENCES	302

LIST OF TABLES

Table 3.1. Summary of processes and purposes aligned with research questions and data sources.	70
Table 3.2. An empty FRAM frame	75
Table 4.1. Overview of functional purposes.....	92
Table 4.2. Functional purpose of the IST and course curators	93
Table 4.3. Functional purpose of each teaching team.	95
Table 4.4. Functional purpose of the students	96
Table 4.5. Functional purpose of the graders	97
Table 4.6. Overview of generalized functions.....	98
Table 4.7. The IST generalized functions associated with content development.....	101
Table 4.8. The IST generalized functions associated with assignments and grading.....	102
Table 4.9. The teaching team’s generalized functions	104
Table 4.10. Generalized student functions	105
Table 4.11. Grader generalized functions	107
Table 4.12. Cognitive level IST functions involved in developing course content.....	113
Table 4.13. Potential variability of course content development functions	114
Table 4.14. Cognitive level IST functions involved in setting schedules	118
Table 4.15. Potential variability of schedule setting functions.....	120
Table 4.16. Cognitive level IST functions involved in developing class sessions	121
Table 4.17. Potential variability of lesson planning functions	123
Table 4.18. Cognitive level IST functions involved in designing assignment tasks	124
Table 4.19. Potential variability of assessment task design functions	126
Table 4.20. Cognitive level IST functions associated with designing grading guidelines.....	129
Table 4.21. Potential variability of grading guideline design functions	131
Table 4.22. Cognitive level IST functions associated with designing grader training	133
Table 4.23. Potential variability of training design functions	135
Table 4.24. Cognitive teaching team functions associated with delivering course content.....	139
Table 4.25. Potential variability of course content delivery functions	140

Table 4.26. Cognitive teaching team functions involved in guiding student practice	142
Table 4.27. Potential variability of in-class activity guidance functions	143
Table 4.28. Cognitive student functions associated with learning course content	145
Table 4.29. Potential variability of content learning functions	146
Table 4.30. Cognitive student functions associated with performing the assigned task	147
Table 4.31. Potential variability of task performance functions.....	148
Table 4.32. Cognitive grader functions associated with training to calibrate grading decisions	150
Table 4.33. Potential variability of grader training functions.....	153
Table 4.34. Cognitive grader functions associated with preparing to evaluate task performance	156
Table 4.35. Potential variability of preparing to evaluate task performance functions	158
Table 4.36. System 1 cognitive grader functions associated with evaluating task performance	161
Table 4.37. System 2 cognitive grader functions associated with evaluating task performance	162
Table 4.38. Potential variability of evaluation and scoring functions	164
Table 4.39. Cognitive grader functions associated with scoring decisions.....	168
Table 4.40. Potential variability of evaluation and scoring functions	169
Table 4.41. Cognitive grader functions associated with working memory.....	171
Table 4.42. Cognitive grader functions associated with the interpretation process	172
Table 4.43. Potential variability of extra evaluation functions.....	173
Table 4.44. Cognitive grader functions associated with doubt.....	174
Table 4.45. Potential variability of extra evaluation functions.....	174
Table 5.1. Observed variabilities of content creation function outputs	179
Table 5.2. Observed variability of evidence item articulation for problem 1 learning objectives	181
Table 5.3. Observed variability of evidence item articulation for problem 2 learning objectives	182
Table 5.4. Observed variability of evidence item articulation for problem 3 learning objectives	184
Table 5.5. Observed variability of assessment task design function outputs	187
Table 5.6. Observed variabilities of grading guideline design function outputs	191
Table 5.7. Observed variabilities of grader training design function outputs	196

Table 5.8. Observed variabilities in task performance function outputs.....	199
Table 5.9. Observed variabilities of training function outputs	202
Table 5.10. Observed variability of criterion judgment outputs.....	210
Table 5.11. Observed variability of scoring output	213
Table 5.12. Observed variability and impact on the system for IST functions	215
Table 5.13. Observed variability and impact of student task performance.....	216
Table 5.14. Observed variability and impact of grader cognitive functions	217
Table B.1. Sample of initial open coding.....	261
Table C.1. Sample of focus coding.....	264
Table D.1. High-level focused codes	267
Table D.2. Detailed focused codes	268
Table E.1. Example of work-as-imagined instantiation coding in Excel.....	270
Table F.1. Example of work-as-completed coding in Excel	272

LIST OF FIGURES

Figure 1.1. Summary of research questions.	32
Figure 3.1. High-level overview of research design aligned with research questions.	61
Figure 3.2. Example of rubric used for grading in the course.	64
Figure 3.3. Learning objective as shown in training.	65
Figure 3.4. The learning objective description and rubric item from training.	66
Figure 3.5. What is not assessed by the learning objective and common student mistakes.	66
Figure 3.6. Example problem associated with the learning objective.	66
Figure 3.7. Provided solution to sample problem.	67
Figure 3.8. Sample A of simulated student work.	67
Figure 3.9. Quiz for Sample A.	68
Figure 3.10. Detailed study design diagram.	69
Figure 3.11. The structure of a FRAM function hexagon.	75
Figure 4.1. Idealized functional purpose level FRAM model.	92
Figure 4.2. Visual representation of an idealized instantiation of the grading system at the generalized function level.	99
Figure 4.3. Visual representation of the cognitive level functions	111
Figure 4.4. Visualization of content development functions.	115
Figure 4.5. Visualization of schedule setting functions.	119
Figure 4.6. Visualization of lesson planning functions.	121
Figure 4.7. Visualization of assessment task design functions.	125
Figure 4.8. Visualization of grading guideline design functions.	129
Figure 4.9. Visualization of training design functions.	134
Figure 4.10. Visualization of course content delivery functions.	138
Figure 4.11. Visualization of activity guidance functions.	138
Figure 4.12. Visualization of content learning and task performance functions.	144
Figure 4.13. Visualization of grader training functions.	151
Figure 4.14. Visualization of evaluation preparation functions. Note: this image has had empty space removed to conserve space.	157

Figure 4.15. Visualization of evaluation and scoring functions.	160
Figure 5.1. Work-as-imagined instantiation for “correct response” for LO 1.....	205
Figure 5.2. Work-as-imagined instantiation for “correct response” for LO 2.....	206
Figure 5.3. Work-as-imagined instantiation of grading LO 1 for sample 2.....	207
Figure 6.1. Recommendations for design of course learning objectives and ancillary content to reduce grading variability.....	244
Figure 6.2. Recommendations to support task development to reduce grading variability	245
Figure 6.3. Recommendations for the design of grading documents and guidelines to reduce grading variability	246
Figure 6.4. Recommendations for design of grader training to reduce grading variability	247
Figure 6.5. Recommendations for grading procedures to reduce grading variability	248
Figure 6.6. Recommendations for organizational policies to reduce grading variability	249
Figure H.1. Student sample 1 for LO 1 and LO 2.....	283
Figure H.2. Student sample 2 for LO 1 and LO 2.....	284
Figure H.3. Student sample 3 for LO 1 and LO 2.....	285
Figure H.4. Portion of student sample 1’s code for LO 3.	286
Figure H.5. Portion of student sample 3’s code for LO 3.	286
Figure H.6. Student sample 1 for LO 4 and LO 5.....	287
Figure H.7. Student sample 2 for LO 4 and LO 5.....	288
Figure H.8. Student sample 3 for LO 4 and LO 5.....	289
Figure H.9. Student sample 1 for LO 6 and LO 7.....	290
Figure H.10. Student sample 2 for LO 6 and LO 7.....	291
Figure H.11. Student sample 3 for LO 6 and LO 7.....	292
Figure H.12. Portion of student sample 1’s code for LO 8.	293
Figure H.13. Portion of student sample 2’s code for LO 8.	294
Figure H.14. Portion of student sample 3’s code for LO 8.	295
Figure H.15. Portion of student sample 1’s code for LO 9.	296
Figure H.16. Portion of student sample 1’s code for LO 10.	297
Figure H.17. Portion of student sample 2’s code for LO 9.	298
Figure H.18. Portion of student sample 2’s code for LO 10.	299

Figure H.19. Portion of student sample 3's code for LO 9.	300
Figure H.20. Portion of student sample 3's code for LO 10.	301

LIST OF ABBREVIATIONS

Abbreviation	Explanation
AERA	American Educational Research Association
APA	American Psychological Association
EI	Evidence Item
FRAM	Functional Resonance Analysis Method
GTA	Graduate Teaching Assistant
HRA	Human Reliability Analysis
IST	Instruction Support Team
LO	Learning Objective
MAbD	Mean Absolute Difference
MAcD	Mean Actual Difference
NCME	National Council on Measurement in Education
PS	Problem Set
RQ	Research Question
STEM	Science, Technology, Engineering, and Mathematics
UTA	Undergraduate Teaching Assistant
W2G	‘What to Grade’
WM	Working Memory

ABSTRACT

Grades serve as one of the primary indicators of student learning, directing subsequent actions for students, instructors, and administrators, alike. Therefore, grade validity—that is, the extent to which grades communicate a meaningful and credible representation of what they purport to measure—is of utmost importance. However, a grade cannot be valid if one cannot trust that it will consistently and reliably result in the same value, regardless of who makes a measure or when they make it. Unfortunately, such reliability becomes increasingly challenging to achieve with larger class sizes, especially when utilizing multiple evaluators, as is often the case with mandatory introductory courses at large universities. Reliability suffers further when evaluating open-ended tasks, as are prevalent in authentic, high-quality engineering coursework.

This study explores grading reliability in the context of a large, multi-section engineering course. Recognizing the number of people involved and the plethora of activities that affect grading outcomes, the study adopts a systems approach to conduct a human reliability analysis using the Functional Resonance Analysis Method. Through this method, a collection of data sources, including course materials and observational interviews with undergraduate teaching assistant graders, are synthesized to produce a general model for how actions vary and affect subsequent actions within the system under study. Using a course assignment and student responses, the model shows how differences in contextual variables affect expected actions within the system. Next, the model is applied to each of the observational interviews with undergraduate teaching assistants to demonstrate how these actions occur in practice and to compare graders to one another and with expected behaviors. These results are further related to the agreement in system outcomes, or grades, assigned by each grader to guide analysis of how actions within the system affect its outcome.

The results of this study connect and elaborate upon previous models of grader cognition by analyzing the phenomenon in engineering, a previously unexplored context. The model presented can be easily generalized and adapted to smaller systems with fewer actors to understand sources of variability and potential threats to outcome reliability. The analysis of observed outcome instantiations guides a set of recommendations for minimizing grading variability.

1. INTRODUCTION

Evaluation of student learning lies at the heart of the educational process. Through formal or informal assessment, evaluation communicates crucial information to multiple stakeholders (Knight, 2002; Marzano, 2000; Nitko, 2001). Students use evaluation to gain essential feedback regarding their achievement of learning goals and their need to make adjustments moving forward. Instructors use evaluation to learn where their students struggled, to identify which concepts to reiterate, and to make evidence-based pedagogical decisions. Parents use evaluation to monitor their child's progress and to identify the need for additional academic support. Guidance counselors, psychologists, and advisers use evaluation to flag students in need of special attention. Higher education institutions and employers use evaluation to estimate a student's mastery of content and skills. Evaluation even shapes educational policy at departmental, institutional, and national levels. Simply put, evaluation of student learning provides the necessary evidence to sustain a productive education system.

Evaluation of student learning often connects to the process of grading, which is the classification of student performance into discrete, ordinal categories (Johnson, 2008). Strong arguments exist against the use of grades. For instance, educational psychology research has suggested that attaching a grade level to feedback can diminish interest in content, hinder cognitive risk-taking, and reduce the quality of student thinking for all students, from elementary school through college, and regardless of discipline (Kohn, 2011). Marzano (2000) argues that the components that comprise grades and the standards for success in those components vary considerably across teachers that grades are effectively meaningless. Still, the process of grading is a necessary reality for most educators (Allen, 2005).

While the objections to grading might dictate that evaluation should focus on directed feedback to learners, the practice of grading has persisted partly because of its apparent ability to efficiently and succinctly communicate a broad overview of student performance. Many instructors consider the ability to fit students into categories of performance to be highly satisfying (Johnson, 2008). However, this efficiency of communication comes at a cost—summarizing student performance with a single letter grade or number results in the loss of crucial information (Allen, 2005). That loss of information makes it impossible to discern the specific components of a given grade, which can vary significantly from one instructor to another (Allen, 2005; Rust,

2011). As such, the entire argument for the use of grades in the evaluation of student learning hinges upon one vital component—the grade's validity.

1.1 Validity

Validity represents an evaluative judgment of the extent to which evidence and theory support the uses and interpretations of assessment scores (AERA/APA/NCME, 2014; Messick, 1995). Validity relates to the uses and interpretations of assessment scores; an assessment, inherently, can be neither valid nor invalid (Kane, 2013). It is also not appropriate to strictly classify a use or interpretation of an assessment score as either valid or invalid; rather, validity lies on a spectrum, and it can change by context and in response to new evidence, use cases, implications, or theoretical understandings (Kane, 2013). The sole purpose of grades is to communicate the levels of students' academic achievements, and validity relates to the accuracy and trustworthiness of those grades (Allen, 2005).

To determine if a use case or interpretation of an assessment score is valid, one must construct and evaluate arguments for and against that use or interpretation (AERA/APA/NCME, 2014; Kane, 2013). Use and interpretation arguments consist of claims based on assumptions that require evidentiary support, such that more ambitious claims require more evidence to support (Kane, 2013). These arguments are practical rather than logical, and one cannot evaluate them using traditional logic (Kane, 1992). Instead, an interpretation and use argument should be evaluated based on coherence, completeness, and the plausibility of its inferences and assumptions (Kane, 2013). One should use theory and empirical evidence to determine the plausibility of assumptions and inferences. Further, any whole argument should contain claims regarding the consequences of use, such that negative consequences are unacceptable (Kane, 2013).

In addition to considering consequences, use and interpretation arguments should include claims regarding the extent to which scores align with their intended meaning (what is traditionally known as construct validity), the fairness of the assessment and scoring practices, and the reliability with which scores are determined. No use or interpretation argument can exist without evidence to support the accurate and complete depiction of the underlying construct intended to be measured (AERA/APA/NCME, 2014). Additionally, many uses or interpretations cannot be considered ethical or accurate without evidence of fairness. However, even accurately represented constructs and fair assessment and scoring practices are irrelevant without evidence of consistency.

As such, reliability of scores is a necessary, though not individually sufficient, condition for validity.

1.2 Fairness and Reliability as Primary Contextual Challenges to Grading Validity

In the context of grading students in engineering coursework, there appear to be three primary challenges to validity, each related to fair and reliable grading: differences in faculty's understanding of validity and goals of assessment; trends to include open-ended items in assessment; and difficulties associated with increasing scale. The first two challenges are also exacerbated by increasing scale because differences in perspectives, knowledge, and interpretation become more relevant and pronounced as more people are involved.

1.2.1 Faculty understanding of assessment validity and goals

Modern experts consider validity to be the extent to which multiple sources of evidence support an argument for the use and interpretation of an assessment and its scores (AERA/APA/NCME, 2014; Kane, 2013). However, research suggests that many educators, including within the engineering education community, hold several misconceptions regarding validity (Douglas & Purzer, 2015). For instance, many retain the outdated idea of validity being one of three types (i.e., content, construct, and criterion). Further, engineering educators often overestimate the relationship between validity and psychometrics, overlooking the importance of the underlying theoretical basis of validity (Douglas & Purzer, 2015). These misconceptions likely result in sporadic and inconsistent attempts to administer valid assessment across engineering courses.

In addition to having limited knowledge about validity, many educators perpetuate invalid assessment and grading practices derived from the inconsistent and inadequate assessment practices that they observed during their educational experiences (Allen, 2005). For instance, educators frequently attempt to communicate more information through grades than can be reasonably represented by a single number or letter (Allen, 2005). Related to this issue is that educators often confound academic grading with non-academic measures of achievement, such as effort. This merging of academic and non-academic factors into academic grades is inherently subjective and susceptible to bias and unreliability due to the variance in educators' values (Allen,

2005; Thorndike, 1997). On the other hand, educators may also engage in sub-par assessment and grading practices as a matter of practicality—many educators may not have the time or resources necessary for greater validity. Regardless of the cause, grading validity relies on an unambiguous, consistently interpretable meaning, which is harder to obtain when educators have a weak understanding of assessment or evaluate based on professional judgments that may be nebulous and unarticulated.

1.2.2 Trends toward the use of open-ended assessment items

There is a general trend in education, and particularly in STEM disciplines, toward the use of open-ended, performance-related tasks in assessment, which may be advantageous for multiple reasons but are challenging to grade consistently. Darling-Hammond et al. (2013) and Hansen (2011) argue that high-quality assessment assesses higher-order cognitive skills in authentic, real-world contexts. The competencies desired in STEM fields as described by the Engineer of 2020 (National Academy of Engineering, 2004), the Next Generation Science Standards (National Research Council, 2014), and ABET (2016) support these trends. The push for more open-ended, performance-based assessment is occurring broadly across the nation (Gipps, 1999). Open-ended, performance-based assessment items are powerful as they can assess skills that are not easily accessible through closed-ended measures and can more effectively uncover student understanding (Arffman, 2015). For instance, multiple-choice questions are often unable to discriminate between students who obtain a correct answer for the right reason versus those who obtain the correct answer for a wrong reason or even those who obtained the "wrong" answer despite using an acceptable process (Berg & Smith, 1994). Of course, thoroughly developed and well-designed multiple-choice questions, like those included in concept inventories, can provide insights about misconceptions. However, the abstraction of a student communicating understanding through the selection of a single letter in a pure multiple-choice item obscures the perception of what motivates their selection.

Despite the benefits of a broader range of assessable competencies and the superior discriminating abilities of open-ended assessments, there are many challenges associated with reliability. Arffman (2015) identified three general threats to reliability with open-ended assessment items: unclear and complex tasks or questions; arbitrary and illogical coding rubrics (addressed in §1.4); and unclear and ambiguous student responses. Complex, ambiguous, or

unfamiliar wording can elicit unintended and unexpected student responses, which may be due to a reliance on the student possessing specific language and comprehension skills to achieve an appropriate interpretation of the task. Similarly, overly complicated questions sometimes require context-irrelevant concepts or skills to answer correctly. While any number of uncontrollable non-cognitive factors (e.g., fatigue, stress, wellness) might produce an inaccurate representation of a student's abilities, assessment questions are manipulable. As such, questions should not require knowledge or abilities that are unrelated to the construct when avoidable.

Open-ended assessments can also challenge reliability due to their ability to produce unexpected, unclear, and ambiguous student responses. On the one hand, the general openness of questions may allow for multiple, equally appropriate responses, some of which may not have been anticipated by the assessment designer and are not adequately handled by the chosen grading scheme (Johnson, 2008). These unanticipated responses demand more experienced graders, who may not always be available (National Research Council, 1993). It is possible to design questions that over-specify specific responses; however, this threatens the assessment's authenticity and can still fail to prevent misinterpretations nonetheless (Johnson, 2008). On the other hand, questions may require construct-irrelevant skills (e.g., strong English comprehension or writing skills), masking the students' true construct-related abilities (Arffman, 2015).

In addition to technical issues with the questions, rubrics, and responses, the scorers themselves are social beings whose experiences, values, knowledge, and perceptions shape their interpretations of both the students' work and the criteria for evaluation (Gipps, 1999). Differing backgrounds and problem contexts can cause scorers to develop different referential models for each level of competence (Johnson, 2008). Clearer models of competence run the risk of overly specific descriptions that reduce the generality and transferability of the assessed constructs (Johnson, 2008).

1.2.3 Difficulties with increasing scale

As class sizes increase, instruction can benefit from economies of scale that assessment cannot, both in terms of acceptable practices and time requirements (Gibbs, 2006). Institutions often do not account for class size in the assignment of class contact hours, but assessment loads are inherently proportional to class size (Gibbs, 2006). Administrators rarely provide instructors with reduced commitments to compensate for teaching larger classes. Similarly, universities often

use student fee income from large classes to subsidize lower enrollment courses, resulting in smaller resource allocations per student in large classes (Gibbs, 2006). Consequently, instructors regularly make concessions with assessments by reducing frequency or quality, reducing feedback volume or quality, or using alternative methods to score assessments, such as teaching assistants or self- and peer-assessment techniques.

In addition to individual class size, scale also relates to multiple sections of a single course. For instance, without regulation, there can be extensive variability of curriculum and assessment across sections and institutions (Gipps, 1999). For some institutions, up to 34% of the observed variability in final grades in introductory calculus, physics, and chemistry courses was accounted for by the students' class section (Ricco et al., 2012). The same study noted that lower variance existed for institutions that had policies to reduce differences in instruction and assessment practices across sections. Karimi (2015) found similar improvements in consistency by coordinating across sections of first-year engineering courses. Still, even measures to reduce differences might be insufficient. In a study on a large-scale assessment scoring program, Congdon and McQueen (2000) showed that the reliability across multiple graders varies throughout a given day or across days and that even the reliability of a grader with themselves can vary significantly throughout a grading cycle.

1.3 Consequences of Invalid Grading

The Standards for Educational and Psychological Testing (2014) identify several distinct groups who use assessment, including students, educators, administrators, researchers, psychologists, employers, and policymakers. Each group uses assessment scores to inform different types of decisions. Assessments can influence learning, future options, and personal wellbeing for students. The use of assessment plans or data often informs instructors' instructional decisions. Administrators use assessment results for accreditation and resource allocation. Thus, the previously discussed contextual factors lead to potential consequences that drive the need for extensive evidence to support validity, reliability, and fairness (Messick, 1995).

1.3.1 Consequences for students

Assessment can impact learning outcomes, personal wellbeing, and future opportunities for students, though these consequences are highly interrelated. When students perceive that they have performed poorly, their access to future opportunities is limited, and their attitudes, mental health, and willingness to engage in future learning opportunities are also harmed (Arnold, 2002). Receiving lower grades can result in reductions in ambition, confidence, and motivation (Arnold, 2002). Further, students who receive worse-than-expected grades experience significant reductions in self-esteem, affect, and identification with their chosen major (Crocker et al., 2003). These results appeared to be more pronounced for both engineering students (compared to psychology students) and women, although male students benefited the most from better-than-expected grades. In addition to personal factors such as reduced senses of self-efficacy and motivation, students' chances of acceptance to academic institutions, and chances of receiving scholarships, tuition assistance, or job offers are also affected by assessment scores (Allen, 2005).

1.3.2 Broader consequences

The consequences of assessment extend beyond individual students. In the context of engineering, students' successes with assessment and perceptions of fairness are essential factors for achieving two of the three goals set forth by the National Academies of Sciences, Engineering, and Medicine (2018): striving for equity, diversity, and inclusion; and ensuring adequate numbers of STEM professionals. The report states that unwelcoming disciplinary cultures and "chilly" departmental climates contribute to the continued underrepresentation of female, minority, disabled, and economically disadvantaged students. An earlier report (National Academies of Sciences, Engineering, and Medicine, 2016) suggests that the normative STEM culture, which views student abilities as genetically determined, causes many highly competitive introductory engineering courses to be barriers that discourage underrepresented students. Further, consistent discrimination faced by women due to harmful implicit biases, held by both men and women, cause women to be less likely to be hired, to receive less credit for identical achievements as men, and to be less likely to get the benefit of the doubt when information is scarce (National Academy of Sciences, National Academy of Engineering, and the Institute of Medicine, 2007). It is

reasonable for one to assume that other underrepresented student groups might face similar challenges.

The fact that disproportionately fewer students transfer into engineering from non-engineering majors than between other fields contributes to a dearth of engineering graduates (Main et al., 2015). These results are partly due to the grades students earn in introductory courses in the engineering curriculum. Students who perform well in introductory courses and earn higher GPAs have higher performance expectancies in engineering programs and are more likely to retain their intended major (Main et al., 2015). Therefore, unreliable or unfair assessment practices could lead to the attrition of capable students who might have had successful careers in engineering, potentially perpetuating underrepresented groups.

Assessments have implications that transcend individual disciplinary fields and move to broader society. For instance, policymakers use assessment data to make funding allocation decisions across different institutions (Johnson, 2008). Policymakers often hold mistaken assumptions about the meaning of grades and make invalid inferences (Johnson, 2008). Such misunderstandings, weak inferences, and the awareness of potential limitations to measurement accuracy raise public concerns about the credibility of assessment data (Newton, 2005), which may be particularly damaging in the current era of "fake news." Concern over the misuse of assessment is justifiable: invalid use of assessment potentially contributes to cultural reproduction and social stratification, as poor examination results can deny students access to higher education and social, political, and economic advancement (Gipps, 1999).

Given the collective consequences of assessment uses, assessment scores must be valid. Students' grades should reflect their knowledge and abilities and not be a consequence of unfair assessments or unreliable grading. Therefore, valid use of assessment is the fundamental contributor to the quality of data upon which students, instructors, and institutions make decisions that affect engineering and society at large. As such, it is beneficial to explore mechanisms to improve the validity of assessment use and grading.

1.4 Means to Improve the Validity of Grading

The Standards for Educational and Psychological Testing (2014) place the ultimate responsibility of valid use and interpretation of assessment on the user of the assessment and its scores. Related to the previously discussed issues of inconsistent components of grading, educators

should think carefully about what contributes to a grade and explicitly articulate those contributions. When multiple sections of a course exist, particularly at a single institution, administrators should remember that coordinated instruction and consistent assessment and grading standards across all sections improve reliability.

Standards-, competency-, or learning objective (LO)-based grading represents a set of methods that can strengthen the consistency of performance expectations across multi-section, multi-instructor courses in a fair and meaningful way (Betts & Costrell, 2001; Muñoz & Guskey, 2015). Holding students accountable to performance on explicitly stated outcomes informs instructors of the extent to which individual students or entire sections have mastered the content. The consistent application of grading criteria and standards allows the grades of courses using these approaches to convey more information and, therefore, be more meaningful (Muñoz & Guskey, 2015). From a fairness perspective, using these approaches also helps students at all incoming ability levels and backgrounds have more common expectations (Guskey, 2001). Further, standardized learning outcomes, assessments, and scoring procedures are vital preconditions to the technical reliability of assessment (Gipps, 1999).

In addition to establishing standard learning outcomes and assessments, Standards 6.8 and 6.9 of the Standards for Educational and Psychological Testing (2014) state that those responsible for scoring must establish scoring protocols. More specifically, responsible parties should establish rubrics, procedures, and criteria and provide adequate training and quality control whenever human judgment is involved. Further, the standards recommend documentation and correction of systematic sources of scoring error. While the Standards are rather strict and outline potentially unrealistic goals for everyday classroom assessments, the National Council on Measurement in Education (NCME) (2019) provides a more practical set of standards. The NCME standards state that classroom assessment should be unbiased and fair (i.e., unaffected by factors not associated with the skill or knowledge intended to be measured) and reliable and valid (i.e., consistent, dependable, and appropriate to support an interpretation of and decisions about student learning). Considering the push toward open-ended, performance-based assessment tasks that require human judgment to conduct evaluation, rubrics, intended to explicate criteria and levels of performance in evaluating student work, are a valuable tool for achieving reliable grading (Jonsson & Svingby, 2007). Price and Rust (1999) attempted to develop common assessment standards and rubrics across an entire academic department. They found the approach provided better guidance to

students about expectations (thereby improving the quality of their work), raised the quality and consistency of scoring for both individual scorers and scoring teams, and improved the quality of feedback given to learners.

Despite the explication of criteria and differentiation between performance levels provided by rubrics, studies have shown that inconsistencies persist. For instance, providing a bit of explanation for Congdon and McQueen's (2000) variable reliability of grading at large scales, Braun (1988) and Crisp (2010) found that inconsistencies in scoring can occur due to individual differences in individual leniency, the time of scoring within a day, the team scoring leader, or the scorer's experience, where novice scorers tend to be less consistent. Arffman (2015) also found that rubric use is less valid when scoring rubrics are vague, illogical, or arbitrary. Thus, criteria and distinctions between performance levels need to be clear, fair, reasonable, and meaningful. Arffman (2015) noted that too many performance levels or criteria specifying arbitrarily fine distinctions made it difficult for graders to discriminate between each level.

1.5 Grading Open-Ended Problems in Large University Courses

The challenges of creating, implementing, and interpreting valid assessment of performance-based tasks become more of a hurdle as the numbers of students being assessed increases. Open-ended assessment for large, multi-section courses involves many groups of people—assessment developers, rubric developers, graders, training developers, instructors, and course organizers. This set of roles, along with the artifacts created or applied by each role, clearly constitute a complex system. In some cases, few of the people in any one group are involved in any of the other groups, making effective communication an essential element of a successful system. In such a system, challenges to validity might stem from a variety of sources, including the assessments, the rubrics, or the training materials. On the other hand, a lack of alignment across any of the system's components or inconsistencies in their use may also threaten validity.

In addition to the many people involved in assessment in large classes, these contexts have other qualities that challenge assessment validity. These systems are more dynamic than smaller classes or large, standardized assessments like the SAT. Smaller courses are often taught by a single instructor who may repeatedly teach the course over multiple semesters, honing their grading expectations and procedures. With a single-instructor course, even if policies or procedures change from one semester to the next, the intra-rater reliability (i.e., reliability with

respect to oneself) is likely better than the inter-rater reliability needed for multiple graders. Not only are many graders needed in large courses, but there is often considerable turnover from one term to the next, losing the consistency and accumulation of knowledge and experience that occurs with a single instructor course.

Content and course materials also contribute to the challenge. When many students take a course, the assignments and associated rubrics often undergo significant changes each year not only for the sake of improvement but also to prevent issues of academic dishonesty, such as previous students sharing their work with current students. Having more students means there are more potential violators, making it harder to prevent dishonesty. In comparison, large-scale testing companies have the opportunity to screen and pilot future questions and exercise significant control over the leakage of content through restrictive testing procedures. Unlike large-scale testing contexts that often address relatively narrow ranges of competencies, the largest university courses are often introductory courses that possess a survey-like structure. As a result, these courses typically cover a broader spectrum of learning objectives across the semester than might be covered in a smaller, more focused course or assessed by standardized tests. As such, achieving consistency of grading interpretation in large, multi-section university courses, particularly when assessing open-ended performance tasks, can be extremely difficult. However, the potential consequences of invalid evaluation of student learning in these courses make the need to understand reliability and fairness in these contexts a critical problem to explore.

1.6 Research Questions

Grades, in all circumstances, should be fair and meaningful. This validation requires the development of a use and interpretation argument that is backed by substantial evidence. Kane (2013) suggests using many sources of evidence to build these arguments. However, the number of sources of evidence for valid use and interpretation is irrelevant if grading is unreliable. Unreliable grades render all other evidence meaningless.

This study focuses on the grading in a required engineering course spanning several sections of over 100 students at a large midwestern university. The course uses open-ended performance tasks and employs many graders. As such, the most significant obstacle to building a sound use and interpretation argument is the reliability of the grading. Other sources of evidence to support the use and interpretation argument, such as the extent to which the assessment tasks

represent the intended constructs, are well-handled by the faculty and staff who curate the course content. Given the known issues with assessing open-ended tasks already discussed and the observed inconsistencies across graders for this course (Hicks & Diefes-Dux, 2017), it is vital to determine the strength of the reliability evidence for the argument for using and interpreting course grades. Beyond this, because NCME (2019) argues the importance of unbiased, fair, reliable, and valid classroom assessments, it is necessary to collect evidence of achievement.

As evidence of reliability is necessary for grades to be valid and meaningful, there is a clear need to develop a deep understanding of how grading occurs in this specific context. This understanding can illuminate weaknesses (i.e., sources of unreliability or variability) and drive possible mechanisms for improvement. The need for understanding and mechanisms for improvement leads to the overarching research question: *What is the evidence for reliable grading of open-ended engineering tasks across many graders applying rubrics, and how can it be strengthened?*

As alluded to in the previous discussion, this study frames grading of open-ended tasks in large, multi-section courses as a complex system. A stable system, from this perspective, would produce a consistent output (i.e., grade) for any given set of inputs (i.e., problem characteristics, rubric characteristics, and student response), regardless of any inherent variability of internal functions in the system. Thus, the overarching research question seeks to identify possible ways to improve this type of grading system's stability.

Answering this overarching question requires a thorough understanding of the grading system and how it varies. In this context, the grading system effectively represents the grading process and the factors that contribute to that process. Unfortunately, the complexity of this system makes direct observation of the system difficult. Thus, aspects of the system were observed directly through a controlled environment to inform inferences about how the system functions and to explore the following research questions:

1. Based on experience-based perspectives of performing grading, teaching assistant, instructor, and content developer roles and observations of grading within a controlled environment (i.e., think-aloud interviews), what is a comprehensive process model of the grading system?
 - a. How do the cognitive grading components identified in the model extend previous research regarding the use of cognitive grading strategies?

- b. What processes in the grading system, beyond grading, might affect system variability?
 - c. What does the model say about possible variability within the system?
 - d. What does the model say about the propagation of variability in the system?
- 2. Based on observations of grading in a controlled environment, how do model instantiations vary?
 - a. How do variable outputs from the content developer, teaching team, and student aspects of the model affect ideal model instantiations of grading (i.e., work-as-imagined model instantiations)?
 - b. How do work-as-imagined model instantiations differ from actual instances of grading (i.e., work-as-completed model instantiations)?
 - c. Which variable outputs of the content developer, teaching team, and student aspects of the model contributed most to the actual variability of work-as-completed instantiations?
 - d. How resilient is the system? That is, how well does the system produce acceptable outputs despite variability observed across work-as-completed instantiations?
- 3. What are reasonable inferences about sources of variability within the system based on an analysis of the work-as-imagined and work-as-completed model instantiations?
 - a. What possible mechanisms might dampen the identified variability?

Each top-level question feeds into the next question to ultimately address the overarching question regarding the evidence for the reliable application of grading in this system and how it can be improved. The first set of questions relates to developing an understanding of how the system functions through the development of a general process model. The second set of questions uses direct observations to understand how the system may operate in practice. Collectively, the results of the first two sets of questions indicate the system's vulnerability to variability and how to strengthen it.

1.7 Summary

The evaluation of student learning is a vital part of the educational process and often portrays the learning of an entire semester of content into a single letter or numerical grade. Many groups of people use grades for various purposes that can have enormous consequences ranging

from the individual to the societal level. As such, it is imperative that grades be fair, that they can be relied upon to have the same value no matter who determines them or when, and that their actual meaning aligns with their intended meaning. In other words, to ensure that grading data accurately inform potentially weighty decisions, grades must be valid.

Unfortunately, some complications threaten the validity of grades and, despite strategies to improve reliability, existing evidence suggests that grades are not always adequately reliable. Even if some degree of subjectivity is unavoidable, it is necessary to hold the amount of variability that exists in the grading process to a minimum. The intention of this study, therefore, is to understand how a complex grading system can and does function, particularly in terms of how the system is susceptible to variability. Secondly, this understanding contributes to the validity argument for the use and interpretation of grades in the selected course by analyzing evidence of reliability and identifying the potential for improvement. Figure 1.1 summarizes the research questions used to achieve the goals of this research. While these questions will explore a specific grading system, the findings from this research may be generalizable to other assessment and evaluation contexts.

Overarching Question: What is the evidence for reliable grading of open-ended engineering tasks across many graders applying rubrics and how can it be strengthened?

RQ 1: What is a comprehensive process model of the grading system?

- a. What relevant processes occur outside of grading?
- b. How does model extend models of grader cognition?
- c. How might the system vary?
- d. How can variability propagate?

RQ 2: How do model instantiations vary?

- a. How does context affect the system (work-as-imagined)?
- b. How does completed work differ from imagined?
- c. Which contextual factors contribute most to observed variability?
- d. How resilient is the system outcome to internal variability?

RQ 3: What inferences can be made about variability in the system?

- a. What mechanisms might dampen variability?

Figure 1.1. Summary of research questions.

2. CONCEPTUAL FRAMEWORK AND LITERATURE REVIEW

Developing an understanding of the components of a grading system and considering how those components' performance might vary contribute to a logical theoretical foundation for a study hoping to reduce variable grading outcomes. To this end, this chapter starts with a set of literature related to socio-technical systems and how to study their variability as an overall template for the structure of this study. Following a presentation of the literature dictating the study's structure, this chapter includes two additional bodies of literature related to two essential components of the grading system: the grading documents, which relate to not only the grading schemes but also to the assignments and the corresponding student work; and the graders themselves, including the factors associated with their performance and theories of the cognitive processes involved in grading.

2.1 Assessment as a System

A system is an assembly of interacting components that may function dependently or independently and may be classified based on a few characteristics of the system (Ghaboussi & Insana, 2018). A system may be static or dynamic, depending on how elements on the system change over time and how those changes influence the outputs of parts of the system or the system as a whole (Hollnagel, 2012). Systems can be classified in multiple ways, one of which is based on the extent to which aspects of the system's performance can be known and predicted. This approach to classification highlights the complexity of the system, whereby the less knowledgeable we are about the system, the more complex (Hollnagel, 2012). Further, systems are considered either technical (also known as technological) or socio-technical, depending on whether humans are involved (Hollnagel, 2012).

Hollnagel (2012) explains that there are generally four assumptions associated with technological systems that aid in understanding them:

1. The system, or events that occur in the system, can be decomposed into simple parts or steps.
2. The parts or steps are either successful or failures.
3. The order of events within the system is predetermined and fixed.

4. Combinations of events are ordered and linear.

For purely technological systems, these assumptions may frequently be met, so knowledge of the system can be wholly, or at least nearly, complete. However, when a system's events include or are influenced by human behavior, which is relatively unpredictable, the validity of these assumptions are highly questionable (Hollnagel, 2012).

Socio-technical systems, as opposed to purely technological systems, are typically intractable (Hollnagel, 2012). This intractability stems from the number and complexity of system details, the rate of change, the comprehensibility, and the processes. In intractable systems, there are often many details that require elaborate descriptions. They can be highly dynamic, with component descriptions or system structures changing rapidly. Aspects of how components function may be at least partly unknown. Additionally, the processes may be heterogeneous and irregular.

Any grading that is not fully automated requires human activities and is, therefore, a socio-technical system (though one could argue that automated grading is still susceptible to human variability due to the programming process). When a single teacher is involved, the system is relatively simple. The assignment, which produces the student work, and any tools to assist grading, which are the technical components, are typically designed and employed, the social processes, by the same person who teaches the students how to complete the assignments. As such, the components can be well-aligned in purpose, and the processes can be consistent and predictable. The complexity of this system significantly increases at larger scales, as the people involved in each process may exhibit variable interpretation and application. Further, alignment may weaken, as the people responsible for designing components within the system may differ from those interacting with the components.

There are a few different approaches for designing grading systems that can mediate the challenges of variability when scales are large. One option is to design the assessments themselves to consist of closed-ended or multiple-choice items. Disregarding infrequent and easily correctable errors that might occur when graded by hand, multiple-choice assessments can be graded with complete objectivity and consistency. However, the multiple-choice questions themselves are susceptible to a wide range of threats to validity (Haladyna et al., 2002). Further, multiple-choice questions rely on a set of predefined answers and cannot possibly assess open-ended tasks. Thus, multiple-choice questions are at odds with Wyatt-Smith & Klenowski's (2013) noted push in

recent years toward more cognitively demanding and complex assessment tasks. This push is particularly relevant in engineering, where open-ended problem solving epitomizes engineering practice (Douglas et al., 2012).

Grading systems at large scales can also be implemented using self- and or peer-assessment. As the grading of open-ended tasks can be time-intensive and the time cost can become prohibitive when the student-to-grader ratio increases, self- and peer-grading can be an appealing option, despite their questionable reliability and validity (Jonsson & Svingby, 2007). Research suggests that self-grading helps students internalize learning criteria and strengthen learning; however, while self-grading produced grades can be accurate, they tend to be inflated (Jonsson & Svingby, 2007). Peer-grading can correlate well with expert or instructor scoring, but generally requires an aggregating of at least four peers for reliability and validity to near grading by an instructor and the method lacks some of the learning benefits that occur with self-grading (Jonsson & Svingby, 2007; Schunn et al., 2016).

As will be described more thoroughly in the next chapter, the grading system that is the focus of this study assesses open-ended tasks at such frequent rates for so many students that each section could not possibly be graded by just the instructors and graduate teaching assistants who must also organize and deliver content. As such, six near-peer undergraduate teaching assistants assume the bulk of the grading responsibility in each section, two of whom do not attend the class. Meanwhile, a team of instructional support staff and instructors design the assignments and rubrics. Thus, in the context of this research, assessment is a highly complex socio-technical system with many human components and most closely resembles, though not exactly, a peer-grading approach.

2.2 Aspects Associated with Grading Documents

Within the socio-technical system of grading, there are three primary “technical” components. Throughout this document, these artifacts will be referred to collectively as “grading documents,” and include the assignment itself, the student work, and the tools or schemes used to evaluate the student work. This chapter addresses the technical components of the system first because of their direct controllability; however, keep in mind that the social, human components of the system (i.e., having multiple graders who work under time constraints and experience external stressors) may contribute just as much as, if not more than, the technical components.

2.2.1 Assignments and student work

As the section about assessment as a system described, assessments may include a variety of question types, such as closed-ended multiple-choice items, or more open-ended performance tasks. The assignment's specific aspects affect how the assignment can be graded, which affects the variability of grading system output. One study performed generalizability theory analyses to explore sources of variability in the grading design projects (Menéndez-Varela & Gregori-Giralt, 2018). Their findings suggest that the type of assigned task contributed to up to 17.3% of the observed variability in grading.

Black et al. (2011) distinguish factors about questions that affect grading as being either directly manipulable, indirectly manipulable, or non-manipulable, where designers can control directly manipulable factors, partially control indirect factors, and only use non-manipulable to guide prediction. The two directly manipulable factors they identify are question features and mark scheme features (both of which could relate to the maximum allowable grades, intended difficulty for students, and the process for determining a “definitive” grade). All of these features affect the non-manipulable factor of student response features. Meanwhile, the grading task's organization, the grading technology's usability, and the physical work environment are indirectly manipulable. All of these features, including the grading strategy, affect the cognitive resources required and, ultimately, the grade's reliability.

Using empirical data, Black et al. (2011) identified several key features of questions. The questions may be written in any number of formats from objective, or constrained, items to short or extended subjective items, where the former are typically graded more reliably (Black et al., 2011). Questions may also require verbal, non-verbal, or mixed responses. Given the questions, the range or scope of acceptable answers is impactful, where a wider, more open set of possible answers may be less reliable. Questions that elicit long, open-ended answers demand that graders read lengthier responses with variable word choices or problem-solving approaches, increasing divergence of grading decisions. Similarly, the complexity of the acceptable answer, such as a simple recall of knowledge versus an intensive application, contributes to the difficulty of achieving consistency. Many of these factors increase the cognitive resources needed for graders to extract the intended meaning from the student's response and to evaluate the congruence between the student's work and the acceptable answers.

Suto and Nádas (2010) also found that more challenging questions were harder to grade consistently. Simple questions typically only require simple cognitive grading techniques, while more challenging questions demand deeper cognitive engagement and reflective judgment. Additionally, they found that when different sub-parts of a question are individually assigned marks, and those parts are dependent, graders' decisions are more likely to diverge. They also found that the need for the grader to have and apply content knowledge and the number of demands placed on the student by the question to have a notable, though smaller, effect.

Suto and Nádas (2010) identified the previously mentioned features separately for mathematics, physics, and biology assessments, and found that different domains were more or less likely to be affected by each of the question features. Suto and Nádas (2009) provided more detail of the technique they used to identify the features—the Kelly's Repertory Grid. With this approach, they found that the abstractness of the question content, the amount of algebra needed, the amount of mathematical phraseology, the allowance of alternative answers, the amount of description needed from the student, the use of diagrams or graphs, and the context of the question were all relevant in mathematics questions. On the other hand, the prompting of recall, the application of knowledge, the quantitative nature of the task, the amount of writing needed to answer, the reliance on external information from the student, the need for a diagram, and the amount of reading necessary for grading were relevant features in physics questions. These patterns suggest that different features of questions may be relevant in different disciplines or for different content; notably, all of these features relate to cognitive demand and extraneous cognitive load.

Black et al. (2011) also identified factors of the student work that affect grading variability. Most simply, the amount of physical space provided to the students for their answers can affect the quality and readability of their work. Less organized and harder to read answers are more difficult to grade and are graded less consistently. Responses that are more constrained are simpler to grade. However, less controllable are the spelling, clarity, legibility, and nature (i.e., expected versus unexpected) of the response. Low quality in any of these features of the student's work forces the grader to slow down to interpret the work, which generally results in decreased consistency.

In addition to the factors mentioned above, the quality of the student's work directly contributes to grading reliability. Russell et al. (2017) conducted a study comparing peer grading

with expert grading and found that, overall, peer grades agreed well with experts' grades. However, high- or very low-quality work produced the best agreement. More significant variation existed for peer ratings of mid- or low-quality work, even for peers who demonstrated higher grading competence throughout calibration training. Cooksey et al. (2007) reported similar findings concerning the consistency of mid-quality work.

2.2.2 Schemes and tools for grading

As suggested previously, open-ended or constructed-response assessment tasks require the judgment of at least one evaluator (and may be referred to as rater-mediated assessments) and are becoming increasingly popular in large-scale assessments (Wind & Peterson, 2018). Large scales necessitate the development of a marking scheme to facilitate evaluative judgments. The goal of any such marking scheme is to assign scores to student work that accurately represents how much and how well a student has learned (Ahmed & Pollitt, 2011). However, several obstacles make it difficult for scores to be properly representative.

When grading constructed-response student work, there are three primary threats to producing valid scores: wrong behaviors presented by students, inconsistent scoring by evaluators, and construct-irrelevant variance (i.e., variability in student performance due to a problem's requirement of knowledge or skills irrelevant to the construct under consideration) (Ahmed & Pollitt, 2011). When designing an assessment, the writer expects and anticipates students to produce a set of responses, ranging from low quality to high quality. Students do not always present all of these responses, but students regularly produce answers that were not anticipated by the assessment writers. Ideally, both of these quantities are minimal, but responses not anticipated by the grading scheme are most threatening to validity. The other two threats relate to the extent to which the graders understand the range of possible and observed answers. When graders do not understand student performance expectations and what represents high- or low-quality work, they are more likely to be inconsistent or assign scores based on factors not related to the relevant construct. However, as Wind and Peterson (2018) point out, construct-irrelevant variance may also occur when graders perceive unfairness in the grading scheme.

The threats to scoring validity can be mediated with well-developed grading schemes. Ahmed and Pollitt (2011) devised a scale ranging from level 0 to level 3 to indicate the extent to which a marking scheme assists with reliable scoring. Level 0 schemes provide no help, while

level 3 schemes offer guidance for scoring every possible response. They suggest that at level 1, there must be at least a description of what constitutes an acceptable performance from the students. Level 2 schemes should also provide descriptions of poor performances. Level 3 schemes must provide a means to discriminate between varying levels of performance and anticipate all possible answers. Ahmed and Pollitt (2011) note, however, that the ease of creating a top-level grading scheme may depend heavily on the extent to which the assessment task is constrained, as constrained tasks can be objectively right or wrong, but unconstrained tasks require guidance for judging quality.

There are, of course, a variety of grading schemes that have been proposed and used. Lench (2010) conducted a study comparing the consistency of four methods:

1. Assigning unconstrained points (i.e., a total number of possible points for the entire assignment with no further specifications).
2. Assigning restrained points (i.e., a specific number of points allotted for different aspects of the assignment).
3. Using generic rubrics (i.e., each aspect of the assignment has point allowances with some general specifications for scoring).
4. Using topic-specific rubrics (i.e., generic rubric, but with directions specific to the assignment).

Lench's study found that consistency was highest with topic-specific rubrics, followed by the point restrained method, followed by generic rubrics, and rounded out with the unconstrained assignment of points. Marzano (2002) performed a similar study and found the same order across the same methods regarding the amount of rater-by-person variability—topic-specific rubrics were the least susceptible to interactions between the evaluator and the student.

Various grading schemes are susceptible to systematic errors. Thompson et al. (2013) found that evaluators selecting a score from a specified maximum value (such as the unconstrained points method mentioned previously) systematically underestimates scores whereas evaluators simply selecting a letter grade leads to systematic overestimation of scores. Silvestri and Oescher (2006) supported these findings, suggesting that lacking a rubric leads to artificial grade inflation. Thompson et al. also considered providing criteria with either a simple four-point ordinal system with check-related symbols or with a full-scale range of integer scoring. Both of these approaches led to stronger discrimination between samples but were perceived to be difficult to use. Further,

while the four-point ordinal system led to fewer irrelevant or arbitrary reductions in scores, the full integer scale was considered to be most accurate.

Chapter 3 details the course associated with this research study more thoroughly. For now, note that the course uses topic-specific rubrics with additional text to guide graders for specific questions. With this in mind, and because rubrics occupy such a large portion of academic literature, it is helpful to begin by exploring rubric design and effectiveness more thoroughly. Note, also, that much of the extant literature does not address issues of scale. The synthesis at the end of this chapter ties these key issues together with the underlying notion of the challenges presented by increasing scale.

Rubrics

While it is commonly accepted that at least trace amounts of subjectivity are unavoidable when assessing complex, open-ended tasks, many believe rubrics can remediate inconsistencies (Andrade, 2000; Jonsson & Svingby, 2007; Stellmack et al., 2009). The argument is that, through a formalization of criteria and an explication of performance expectations, rubrics reduce variations due to variable subjective interpretations across graders (Moskal & Leydens, 2000). Some of the studies presented in the previous section support such a stance. It is helpful, therefore, to understand how rubrics work.

Rubrics, which may be either analytic or holistic, generally consist of a two-dimensional matrix with a list of criteria as rows (the standards or learning objectives being graded) and gradations of quality or performance as columns (Andrade, 2000). Depending on the perspective guiding their construction, rubrics may be one of two primary varieties. The reductionist perspective, sometimes referred to as “rational” or “criteria-driven” assessment, corresponds to analytic rubrics and asserts that evaluations can be rational judgments based on a common set of properties (Wyatt-Smith & Klenowski, 2013). Through training and calibration, evaluators can achieve inter-rater reliability when evaluating those properties. The second perspective, referred to as “global” assessment, claims that not all cases can be reduced to a set of pre-specified features, and should be evaluated using a holistic rather than an analytic approach (Wyatt-Smith & Klenowski, 2013).

Successful design and application of rubrics relies on a number of—potentially questionable—assumptions about how they are used. Rubrics rely on the assumption that criteria

can be developed without norming and can be written unambiguously to allow for consistent interpretation across all user groups (i.e., the graders and the students) (Bloxham et al., 2011). Further, rubrics should be able to be applied reliably, despite potential needs for graders to cognitively coordinate a complex set of criteria while analyzing student work (Bloxham et al., 2011). Wyatt-Smith and Klenowski (2013) argue, however, that the assumption that scoring consistency is an automatic consequence of criteria explication ignores the importance of judgment in the evaluation process. Wyatt-Smith and Klenowski assert that criteria are artificial and abstract constructs open to interpretation and that definitive, exhaustive checklists are rarely achievable and impractical. They also claim that there are two additional levels of criteria beyond those explicitly stated: latent criteria that are initially unspecified but become apparent during the grading process, and meta-criteria (i.e., unwritten criteria that dictate the use of explicit and latent criteria). The existence of latent and meta criteria threaten consistency when many graders are involved, particularly when there is a lack of metacognitive awareness of these criteria levels.

In theory, the amount of judgment necessary while applying a rubric should vary depending on the holistic or analytic character of the rubric, where analytic rubrics should theoretically require less judgment thanks to additional granularity. Sadler (2009b) argues, however, that the analytic approach is susceptible to indeterminacies associated with dual agendas of graders (i.e., looking at a student's work attempting to simultaneously develop an overall sense and identify key characteristics or deficiencies), discrepancies between perceived holistic judgment and analytically derived grades, the assumption that criteria are conceptually discrete, the uniqueness of specific situations defying pre-defined criteria, and individual graders' varying interpretations of the criteria. Still, despite his criticism, Sadler offers no alternative solution for achieving common grounds when many graders are grading across large numbers of students.

Research comparing reliability of holistic and analytic rubrics have also led to mixed results. For instance, Baird et al. (2017) found graders to produce significantly lower mean absolute score differences when using analytic rubrics than when using holistic rubrics. Barkaoui (2011), on the other hand, found holistic rubrics to have stronger interrater agreement. Still, Barkaoui noticed that graders were more self-consistent with analytic rubrics, despite leading to scores that are systematically more lenient than those with holistic rubrics. Despite the greater leniency Barkaoui witnessed with analytic rubrics, the severity divide between novice and expert

graders was less with these rubrics. Further, Barkaoui's work demonstrated that analytic rubrics may be better able to differentiate students into statistically distinct ability levels.

While a few studies exist that empirically probe rubrics, empirical support of rubric use is, unfortunately, rather limited. Reddy and Andrade (2010) and Rezaei and Lovorn (2010) both reviewed the literature and found that, despite widespread use of rubrics, very little empirical research has been conducted to investigate the validity of rubric use. Rezaei and Lovorn (2010) noted one study, for instance, where English faculty applied rubrics designed by several Education faculty to grade writing assignments and were still more swayed by the mechanical aspects of the writing than the content. As a result, they expressed concern for the design and use of rubrics by faculty outside of Education. Reddy and Andrade's (2010) also noted that the large majority of studies did not establish quality, as they failed to describe their process of rubric development. Based on their review, Reddy and Andrade found four areas they believe to be most in need of attention in rubric research: the use of more rigorous research methods and analyses, expanded geographic and cultural perspectives, more research on validity and reliability, and a closer focus on learning.

Rubric design

Despite the limited amount of empirical data, there have been a number of notable articles that provide recommendations for rubric design based on a combination of common-sense, personal experience, or reviews of rubric-related articles. Popham (1997), Moskal (2003), and Tierney and Simon (2004) all note the importance of focusing criteria on skill mastery rather than task mastery and the need for criteria to be specific enough to discriminate between performance levels. However, Popham (1997) and Tierney and Simon (2004) note the importance of concision of criteria to prevent details being overlooked. Further, the criteria should be free from bias, expressed in terms of observable behaviors, understandable to all users, and should be consistent and parallel across a rubric (Moskal, 2003; Tierney & Simon, 2004).

Some empirical research has been conducted that provides guidance for the design of criteria. Goldberg's (2014) study of engineering design rubrics led her to recommend that when a rubric is designed, one should ask if the rubric and corresponding assessment task adequately capture all aspects of a construct. However, Menéndez-Varela and Gregori-Giralt's (2018) generalizability study of rubric use led to conclusions that too many criteria can make student work

difficult to manage. Further, while not an empirical finding, Sadler (2010) warned that inclusion of too many elements might cause the problem opposite to inadequate construct coverage: construct-irrelevant variance. Indeed, Joe et al. (2011) concluded based on their study of rubric grading that over-complexity in a rubric may be the greatest threat to validity. They found that graders often abandoned portions of long rubrics, especially when different elements needed to be considered simultaneously.

Studies have also led to recommendations regarding performance levels and descriptors (i.e., the text that distinguishes each performance level for a given criterion). Goldberg's (2014) findings suggested the importance of evenly spaced performance levels that are defined with consistent, parallel, non-redundant descriptors. Though, once again, Menéndez-Varela and Gregori-Giralt's (2018) recognized that too many performance levels can make discrimination between levels difficult. This might align with Goldberg's (2014) recommendation to look for evidence that the number of performance levels should be expanded or reduced. In addition, Goldberg (2014) suggests investigating text across criteria or descriptors for notable ambiguities, redundancies, or deficiencies that might benefit from revisions that could help the grader. Joe et al. (2011) similarly note that rubrics should be simplified to focus on only the most critical features and constructs, and that those constructs should be clearly articulated.

A rubric's rating scale, both in terms of the number of achievement levels and the point values at each level, may also contribute to variability. Researching large-scale assessment of writing, Humphry and Heldsinger (2014) found what they called the "halo effect," which refers to the phenomenon of graders consistently selecting the same performance level for all criteria when all criteria have the same number of performance levels. They noted that graders would grade based on a general impression of the student or work rather than focusing on individual performance criteria. Thus, Humphry and Heldsinger (2014) recommended freedom to vary the number of performance levels as necessary across a rubric. In a similar study, Woodley et al. (2017) found that when a rubric is constructed such that the lowest possible score is a 1 rather than a 0, significantly more graders are willing to select the lowest performance level. They noted that graders' conflicted emotions about assigning scores of 0 to student work led to inconsistent grading decisions.

Fortunately, it seems that pairing a rubric with sample work can help reduce some of the inconsistencies that might develop for various reasons. Heldsinger and Humphry (2013) argue,

based on a study of teachers grading elementary school writing, that including calibrated exemplars to illustrate expectations at different performance levels for each criterion can improve reliability, particularly when extensive training is not feasible. The use of such exemplars might minimize the concerns Goldberg (2014) had regarding feelings of cognitive dissonance related to misfit between rubric scores and work quality. Still, Heldsinger and Humphry (2013) note that more research is necessary to generalize the benefit of calibrated exemplars to all classroom settings or to large-scale grading.

2.3 Aspects Associated with Graders

As noted previously, no matter how detailed a rubric, many argue that any grading of open-ended performance tasks inherently contains some degree of subjectivity and human judgment (Andrade, 2000; Cooksey et al., 2007; Jonsson & Svingby, 2007; Stellmack et al., 2009). While Menéndez-Varela and Gregori-Giralt's (2018) generalizability theory analysis observed relatively small main effects due to graders (explaining less than 5% of variability in most cases analyzed), they reported grader-by-item and grader-by-student interactions constituting as much as 18% of scoring variability.

There are many factors that might explain the relatively high variability graders can contribute to the process. Cooksey et al. (2007) claims, for instance, that grading relies on the integration of information, perceptions, memory, and training, all of which may vary from one grader to the next. Crisp (2010) argues that graders' judgments stem from their beliefs about the purpose of grading, perceptions of accepted practices and proper interpretations, tendency toward strict adherence to standards, and their mental models of varying work quality. Meanwhile, Griswold (2010) notes the influence of graders' values and beliefs, such as the importance of non-performance factors such as effort and the use of grades as punishment or rewards.

Evidence also suggests that graders tend to grade in a holistic manner, even when analytic rubrics are provided. Hay and Macdonald (2008) observed teachers conducting grading and noted that despite the presence of analytic rubrics, teachers often made judgments about student achievement at a holistic level, without referencing the rubric criteria. These teachers claimed to have sufficiently internalized the criteria and to have had an intuitive feel for achievement, but application of the rubric criteria to student performance suggested these teachers were overestimating their ability to judge accurately. Bloxham et al. (2011) also observed university

lecturers employing holistic judgments despite being given analytic criteria. A few lecturers did reference the criteria after making a holistic judgment in order to check or refine their judgment, but often performed norm referencing at the same time. Still fewer reviewed the criteria before marking, and those who did were the lecturers who had formal training in education. It is likely that these graders suffer from Meier et al. (2006) refer to as the “halo effect”—not to be confused with Humphry and Heldsinger’s (2014) halo effect, this version refers to the tendency to assign grades based on knowledge or perception of a student rather than actual performance.

Joe et al. (2011) corroborate the inattention to criteria, noting that graders tended to focus on less than half of explicated criteria. Further, while it is frequently assumed that graders will only focus on the stated criteria, Joe et al. found that graders regularly embedded their own criteria into the closest existing criteria. In other words, they deliberately use mental models of criteria to fit their personal expectations. Inexperienced graders tend to be more consistent in their scoring than experienced graders, in part because they tend to focus on a more consistent subset of criteria than experts, whose focuses relate to their specific areas of expertise (Joe et al., 2011).

Perhaps the most frequently documented tendency of grader error is the tendency for some graders to consistently grade either too leniently, too severely, or right down the middle (Cook et al., 2010; De Lima et al., 2013; Iramaneerat et al., 2008; Meier et al., 2006; Raymond et al., 2011). Meier et al. (2006) explain that some graders tend to consistently grade on the higher side of a rubric (leniency error), consistently on the lower side of the rubric (severity error), or consistently toward the middle of the rubric (central tendency error).

It should be noted that many of the previously mentioned factors, as well as the tendency to commit leniency or severity errors, tend to be at least partly a function of experience. In a study analyzing the work of graduate teaching assistants (GTAs), Doe, Gingerich, and Richards (2013) found that the more experienced GTAs better approximated scores assigned by expert graders. Further, less experienced graders have been observed to systematically grade more leniently, while having lower levels of agreement and self-consistency (Barkaoui, 2011; Sonner & Sharland, 1993). On the other hand, experienced graders are twice as likely to employ alternative grading strategies such as holistic and associative grading (i.e., grading through comparison with other work) than inexperienced graders, which leads to variable focus on specified criteria (Joe et al., 2011). Still, it is hard to argue that experience is detrimental to grading consistency, especially if the experience is developed through proper, process-oriented training.

2.3.1 Grader training

Following their review of the literature, Rezaei and Lovorn (2010) acknowledged that while rubrics help with grading reliability, the improvement is not inherent or guaranteed. Graders' evaluative decisions are inherently governed by cognitive frameworks and heuristics that will vary if not calibrated (Joe et al., 2011). This is, perhaps, why AERA/APA/NCME (2014) explicitly states that whenever complex responses are scored by humans, careful training is required (p. 112). The Standards recommend that this training consists of samples that exemplify varying levels of performance and also encourages regular monitoring to ensure continued performance.

Despite the requirement for training noted in the Standards (2014), the literature presents mixed findings related to the need of training. For instance, Brown et al.'s (2004) study of the reliability of scoring elementary writing and Bresciani et al.'s (2009) study of a rubric measuring research quality both argue that rubrics can attain high levels of reliability with little-to-no training. On the other hand, Alshuler's (2016) study of rubrics to evaluate students' reflective journals suggests that training benefits even faculty-level graders. Similarly, Baird et al.'s (2017) large-scale analysis of grader accuracy demonstrated a significant effect of training, especially in group training settings, where the group leader had a particularly significant impact. Their study extended the recommendations from the Standards (2014) to include presentations about interpreting questions and rubrics, followed by team discussions of exemplar work.

Given the context of this study, it is also important to acknowledge studies on the training of teaching assistants, which is mostly centered on graduate teaching assistants. Roehrig et al. (2003) found that at the time of their study, only 17% of universities had formal training of teaching assistants lasting more than one day and that training practices varies considerably across institutions. Some literature, such as Essick et al. (2016), does discuss alternative training approaches including weekly courses, training videos, and mentoring programs.

Surveys of graduate teaching assistants find that they have a fear of grading, and it is fair to assume that that fear would extend to undergraduate teaching assistants, as well (Melvin & Bullard, 2010). However, graduate and undergraduate teaching assistants are particularly valuable to the grading process, as they can typically provide timelier performance feedback than could professors and their perspectives more closely resemble the understanding of the students (Dickson et al., 2017). Still, despite the potentially outdated claims of Shannon et al. (1998), it is possible

that many teaching assistant training programs are not sufficiently structured or rigorous to provide the needed support.

To contrast the lacking teaching assistant training programs, a few institutions have developed relatively extensive training programs for UTAs in the first-year engineering programs. Verleger and Diefes-Dux (2013) and Marbouti et al. (2013) present multi-stage models for familiarizing UTAs with content, calibrating grading with experts, and providing feedback. Kecskemeti et al. (2015) followed a similar process at their university and used follow-up emails and meetings with major grading anomalies were identified.

2.3.2 Cognitive strategies in grading

A few researchers have delved into the cognitive processes involved in grading. At a very superficial level, Charney (1984) claimed that graders develop their own idiosyncratic interpretations of criteria and their accuracy is threatened by being thoughtful about a student's work. She argued, instead, that for graders to be reliable, they must perform the process quickly and superficially.

Lumley (2002) developed a more comprehensive view of the cognitive processes involved in grading, suggesting the process involved reconciliation between their overall impression of the student work, specific features of the work, and the wording in the rubric. Lumley noted, however, that the scale rarely accounts for all eventualities, and graders must develop coping strategies based on a tension between their complex intuitive impressions of the work and their understanding of the rules. He identified that managing, reading, and rating were the three general behavior types that occur during grading. Further, he broke the grading sequence into three stages: initial reading to gain an overall impression including global and local features by reading and commenting on salient features; rating each criterion by articulating and justifying score decisions with respect to scale descriptors; and considering the selected scores by confirming or revising score selections upon a final scan of the response. Lumley also acknowledged that graders tend to develop their own unique interpretations to rubric descriptors and must reconcile their perceptions of institutional expectations and any conflicts that may occur between the rubric and the work being graded.

The most extensive work regarding grader cognition has been conducted by Suto and colleagues out of Cambridge. Much of their work is framed by the dual-processing theory

popularized by Kahneman and Frederick (2002). This model breaks cognitive processing into two levels: system 1 processes are quick, relatively effortless, intuitive judgments (and likely correspond to low cognitive demand); and system 2 processes are slower, effortful, conscious, and reflective judgments (and likely correspond to high cognitive demand). Based on a series of think-aloud interviews, Greateorex and Suto (2006) identified five cognitive grading strategies: matching, scanning, evaluating, scrutinizing, and no response. Matching, scanning, no response and some evaluating use System 1 cognition, while scrutinizing and other instances of evaluating use System 2 cognition. Grading processes using System 2 cognition are more likely to be graded with lower consistency (Suto & Greateorex, 2008).

According to Greateorex and Suto (2006), a grader uses one or more of the aforementioned strategies based on a number of factors. At the personal level, the grader's experience teaching, general experience with grading, experience with grading the particular assignment, personal preferences, and directions from a more authoritative figure, such as a principal examiner, each affect the strategies that are chosen. Additionally, the assignment and student work also influence the chosen strategy based on what the problem asks the student to do, the grading scheme applied, what typical responses look like, and what the particular response looks like.

The simplest cognitive grading strategy is the "no response" strategy (Suto & Greateorex, 2006). While Suto and Greateorex (2006) do not say it, attempting to apply the "no response" strategy is certainly the first strategy used in every grading instance. The strategy only requires System 1 processing and is just the determination of whether or not the student has provided a response. If a response appears to be present, the grader selects a second strategy to determine a grade. If no response is present, the student is automatically given a zero and the grader moves on.

Matching is the simplest cognitive grading strategy when work is actually present to be graded (Suto & Greateorex, 2006). A System 1 process, matching consists primarily of comparing some portion or all of a student's answer, ideally in a pre-determined location in the assignment, to the stated "correct" answer. If the student's answer matches the correct answer or solution, points are immediately awarded, and the grader moves on. If they are not the same, the grader must either decide that the answer is definitively incorrect, look at other aspects of the student's work, or consider using a different strategy.

Scanning is a little more sophisticated than matching (Suto & Greateorex, 2006). Initially, it consists of looking through the whole space in which the student response is expected. This

many include multiple scans and the grader may look for one or more details at a time. At a System 1 level, this strategy may involve looking for recognizable visual patterns, such as one or more numbers, letters, or words. Points are then awarded based on whether or not the expected response is present. On the other hand, when a more complex statement, phrase, diagram, or calculation is expected, System 2 process may become involved. However, System 2 is only invoked if the more complex detail is identified, at which point the grader must move on to Evaluating.

During Evaluation, the grader has to figure out the meaning of what the student has written and apply knowledge and information from a combination of sources to determine if the response is accurate (Suto & Greateorex, 2006). When the grader has applied some level of System 2 judgment to determine accuracy, then points are allocated. This determination may require using an additional grading strategy to decide the level of response accuracy.

Scrutinizing is an exclusively System 2 process (Suto & Greateorex, 2006). This occurs when a response is unexpected, partially incorrect, or not aligned with what is given in the grading scheme. The goal is to reconstruct the student's line of reasoning to figure out what the student did correctly and incorrectly. Often, this includes looking through the response to identify specific points of error. Once the source of inaccuracy is identified, the grader must then determine the appropriate level of points to award the student.

Collectively, it is expected that the graders for this study will use cognitive strategies similar to those identified by Suto and Greateorex (2006). It bears noting that Suto and Nadas (2009) recognized that the distribution of the use of those strategies varies across graders grading in different disciplines. This variance in cognitive strategies most likely stems from the variability of the items being graded, where different disciplines (i.e., mathematics, physics, and business, in their study) tend to have different types of problems (e.g., more memorization versus analysis or evaluation). Still, what these cognitive strategies fail to capture are the steps that lead up to applying the strategies and are limited in the description of processes involved in switching between strategies.

2.4 Variability of Human Performance

While technical components in a system may cause failure due to variable lifespans or unexpectedly improper functioning, the primary source of variable performance in a socio-technical system is due to humans. Historically, this variable human performance has been referred

to as “human error” (Sharit, 2006). It should be noted, however, that referring to human actions as error, particularly regarding judgment and decision making, is contended by experts, who opt for terms like “erroneous actions” to refer to actions producing unexpected results or unwanted consequences (Sharit, 2006).

A single, precise definition of “error” is lacking across the literature (Sharit, 2006); still, there are some common themes amongst conceptualizations. Most importantly, erroneous actions are those which result in unwanted or adverse outcomes or consequences. This may also encompass near misses, which did not result in an adverse outcome but nearly could have. On the other hand, exploratory behavior or trial-and-error learning are not considered erroneous. Nor is an intentional violation of procedures; however, because this research regards the entire system, rather than the actions of one actor, intentional violations by one actor will be perceived here as resulting from erroneous actions of another actor.

Many factors may contribute to erroneous human actions, as Sharit (2006) demonstrates with his framework for understanding human error. In this framework, Sharit suggests that adverse outcomes are the result of errors that pass through systemic barriers and originate due to any number of contextual factors (e.g., administrative policies, organizational culture, time constraints, workload, knowledge demands, procedures, training, or communication) and human fallibility (e.g., sensory limitations, short- and long-term memory, biases, expertise, attention, fatigue, and affect). From a sociotechnical perspective, some number of errors may be unavoidable, and it is an organization’s responsibility to implement barriers in the form of policies, procedures, and culture that might hold errors in check (Sharit, 2006). Notably, however, the implementation of interventions as barriers can, themselves, create new opportunities for human fallibility. For example, lengthy procedural protocols could become overwhelming and encourage the development of time and effort saving, though error-prone, heuristics.

Identifying the root cause of an externally visible erroneous action can be difficult, as the same observation may stem from any number of legitimate explanations (Liu et al., 2017; Sharit, 2006). There are many different classification taxonomies for erroneous actions, such as the skill-, rule-, and knowledge-based model or the stage of information processing model (Sharit, 2006). Closely related to the information processing model is the macrocognitive function model (Liu et al., 2017). This model suggests five macrocognitive functions (detecting and noticing, understanding and sensemaking, decision making, action implementation, and team coordination),

each of which may result from failure mechanisms associated with performance influencing factors. While tracing the particular performance influencing factor may be difficult (Liu et al., 2017), identification of the macrocognitive function and likely error mechanism does help narrow down possible causes. Regardless of the taxonomy used, it seems variability in human performance on a given task will relate to their ability to take in, handle, and make decisions based on information. Thus, factors such as cognitive demand, cognitive load, and issues related to decision making can help illuminate sources of variability.

2.4.1 Cognitive demand

The idea of cognitive demand generally suggests that some tasks are inherently more demanding of cognitive resources than other tasks. Tasks have been categorized using a number of different taxonomies, perhaps most famously by Bloom (1956). In an extension to Bloom's work, Anderson and Krathwohl (2001) created a two-dimensional system consisting of a cognitive domain and a knowledge domain. The cognitive domain suggested cognitive demand increased across tasks that demand remembering, understanding, applying, analyzing, evaluating, and creating. Meanwhile, any of these tasks can occur across the increasingly demanding knowledge domain of factual, conceptual, procedural, or metacognitive tasks.

Smith and Stein (1998) presented a classification system in mathematics contexts that consisted of four levels of cognitive demand: (1) memorization, (2) procedures without connections to concepts or meaning, (3) procedures with connection to concepts and meaning, and (4) doing mathematics. Increasing levels of demand corresponded to less directional guidance and greater procedural ambiguity. The most cognitive demanding task required complex, nonalgorithmic thinking, an understanding of relevant concepts, and some degree of self-regulation of cognition. However, it is important to note that the level of demand is not purely a function of the task, but also the prior knowledge and experience of the person performing the tasks. In their study, Smith and Stein found that groups of teachers could achieve strong agreement of sorting relative levels of cognitive demand across several different tasks.

Extending upon Smith and Stein (1998), Tekkumru-Kisa et al. (2015) created the Task Analysis Guide in Science. In this system, the lowest level of cognitive demand remained memorization tasks. However, the next two levels were altered to tasks involving scripts and tasks involving guidance for understanding, the latter being split into two possible levels of cognitive

demand. The highest level was transformed to “doing science” rather than mathematics. Additionally, the new taxonomy incorporated a second dimension of integration, such that tasks could be scientific practices, science content, or integration of both content and practice. This allowed for the differentiation of memorized practices, memorized content, scripted practices, scripted content, scripted integration, guided practice, guided content, guided integration, and doing science. From this perspective, guided integration requires a higher level of cognitive demand than does guided practice or content. Recently, Douglas et al. (2017) applied this framework to an engineering context by replacing “science” with “science/engineering” and defining “doing engineering” as “developing a solution combining content and practice.”

2.4.2 Cognitive load

Cognitive load is discussed differently in the literature than cognitive demand, though one could reasonably infer some conceptual overlap. Analysis of cognitive demand often attempts to characterize the inherent difficulty of a task, ignoring the effect of additional components that may impose cognitive effort. Cognitive load, on the other hand, looks at the total sum of cognitive resources imposed on someone at a given time. Sweller (1994) describes a task's intrinsic cognitive load as the number of individual “elements” that must be handled by someone simultaneously in order to perform the task. That is, tasks with higher intrinsic cognitive load have greater informational complexity due to the interconnectedness and interactivity of ideas they depend upon. Sweller (1994) also notes that intrinsic cognitive load is significantly dependent upon the individual, as increased knowledge and experience lead to development and automation of cognitive schema that require fewer cognitive resources to handle sets of concepts. In this sense, the idea of intrinsic cognitive load relates strongly to cognitive demand.

In addition to intrinsic cognitive load, cognitive load theory also includes extraneous cognitive load. Extraneous cognitive load is purely a function of how the task is communicated rather than with the task itself (Sweller, 2010). Extraneous load imposes demands on someone's cognition that is not germane to the task at hand. For example, jargon can constitute extraneous load if the use of jargon is not specifically necessary for the task, as the person performing the task must translate the meaning of the jargon in addition to performing the task itself. From an instructional perspective, any cognitive load imposed by a problem that is extraneous to the task at hand decreases the “germane cognitive load,” which is the portion of cognitive resources

devoted to acquiring knowledge. Regardless of whether the load is intrinsic or extraneous, someone to whom the content is less familiar is “not in a position to distinguish” between the type of load (Sweller, 2010). Everyone’s working memory is limited to handling a relatively small number of elements, so if the total cognitive load (intrinsic + extrinsic) is too high, the person will not be able to effectively perform the task.

While the design of the task can impose extraneous cognitive load, it is also important to recognize that the physical environment can impose additional load. Characteristics such as visual or auditory noise, smells, thermal conditions, and lighting conditions can all distract attention and limit the resources in working memory available to perform a task (Choi et al., 2014). The effect of the environment on a person can be physiological or affective in nature. Further, given that tools and technology are part of the environment and may constitute a state of distributed cognition (i.e., external holders of information that allow one to handle more elements at once), the design of course materials can produce negative emotional responses that also limit available cognitive resources. As a result, the aesthetic design and orientation or presentation of information may affect performance of a task.

2.4.3 Decision making

While attentional cognitive limitations affect the accuracy of human performance, factors associated processing information and making decisions also play a prominent role, particularly when the task being performed can insight an emotional response. Decision making is a complex process heavily influenced by preferences, values, past experiences, personal dispositions, and mood (Forsythe et al., 2015; Lerner et al., 2015). A considerable number of studies in the field of neuroscience have found numerous strong connections between decisions made and the brain’s reward circuit (Forsythe et al., 2015). Unfortunately, it is difficult to predict a person’s behavior because stimulation of the reward circuit is strongly related to variable qualities such as differences in altruism, risk aversion, testosterone levels, positive or negative associations with related past experiences, or feelings about the subject of the decision.

Despite the difficulties of predicting decision making behavior, there are general trends about the decisions people make related to social connections and perceptions of fairness (Forsythe et al., 2015). For instance, when people perceive unfairness toward themselves, they often exhibit the same neural responses as when angry or disgusted. Individuals with more testosterone

experienced unfairness akin to reactions of potential confrontation. However, these reactions can be mediated and suppressed when unfairness is coupled with potential reward. On the other hand, situations perceived as fair showed stimulation of the reward circuit. When the unfairness is directed toward others, people experience empathy, exhibiting similar neural responses, if the subject of unfairness is considered likeable. However, when the subject of unfairness is unlikeable, no empathetic response occurs and may even lead to activation of the reward circuit (i.e., feelings of satisfaction or pleasure) in some males. In a related context, when given the opportunity to be charitable (e.g., with money, time, energy, or kindness), individuals experienced stimulation of the reward circuit for giving, but were twice as likely to engage in the behavior if they were altruistic rather than egoistic.

Taking a step back from the neuroscientific perspective to the psychological, emotions can have a potent beneficial or harmful influence over decision making behavior, often acting as the dominant driver even in high stakes decisions (Lerner et al., 2015). Effectively, decisions are made primarily to avoid negative feelings or increase positive feelings and to trigger time-tested responses to similar situations that can save cognitive effort. Emotions that affect decisions may be related to the anticipated outcome of the decision at hand (i.e., integral emotions) or may be entirely unrelated and carried over from a previous experience (i.e., incidental emotions). Unfortunately, decision making is not simply a function of emotional valence (i.e., whether an emotion is positive or negative), but also depends on the specific type of emotion and the overall level of arousal. More specifically, emotions shape goals that direct decisions. For example, a feeling of anger may intensify focus and incite a desire to change or overcome the situation while a feeling of safety may lower inhibitions and promote the use of heuristic decision making.

While the emotion-based framework, as with neuroscience, cannot produce exact predictions of decision making due to idiosyncratic differences, the general trends can be used to generally reduce the unwanted effects of emotions. Ultimately, people make decisions based on conscious or subconscious evaluations of characteristics of alternative options, influenced by their personality and preferences as well as integral or incidental emotions (Lerner et al., 2015). Still, there are strategies to reduce emotional influences. Primarily, this may be done by decreasing the magnitude of the emotional response through time, reappraisal, or counteractive emotional states, or by insulating the decision from the emotion by increasing awareness of misattribution or modifying choice architecture. Reappraisal, or reframing the meaning of stimuli that produced an

emotional response, has been shown to be an effective approach. Increasing awareness of misattribution, through reminders to self-monitor emotions and to focus attention on relevant information and filter out irrelevant emotional influences, may be effective, but requires the decision maker to have strong self-awareness and sufficient motivation without significant additional cognitive burden.

2.5 Methods to Analyze Variability in Socio-Technical Systems

Broadly speaking, socio-technical systems can be quantitatively and qualitatively evaluated for the impact of human erroneous actions on system performance using what are known as human reliability analyses (HRAs) (Baziuk et al., 2018). Many HRA methods have been developed over the years and these methods are generally divided into first, second, and third generation methods (Di Pasquale et al., 2015). The first generation, consisting of techniques such as the Technique for Human Error Rate Prediction (THERP) and the Accident Sequence Evaluation Program (ASEP), was strongly rooted in quantification of success and failure probabilities, with little focus on the underlying causes or reason for behavior. The second generation, including methods such as A Technique for Human Error ANalysis (ATHEANA) and the Cognitive Reliability and Error Analysis Method (CREAM), shifted to conceptual cognitive models focused on causes of erroneous actions rather than strict calculations of probability. These methods are more elaborate and sophisticated but are lacking in empirical validation. The newest generation of methods, such as the Information-Decision-Action Crew (IDAC) model and the Functional Resonance Analysis Method (FRAM), attempt to address the limitations of the second generation and are designed to handle more dynamic systems.

The FRAM represents a shift from a focus on system architecture and components to how the system functions (Hollnagel, 2012). Rather than considering a system to be in either a “normal” or “failed” state as in most other methods, FRAM considers the variability of the system’s functioning, partly due to the recognition that human judgment is not appropriately viewed as “failure.” The FRAM also acknowledges that all complex socio-technical systems consist of some inherent variability but concerns itself primarily with whether or not that variability will resonate and/or propagate to produce an unacceptable outcome. By recognizing the dynamic interrelationship between functions within a system, the FRAM provides flexibility for analyzing

systems that can be highly variable in implementation and identifying how variable outputs of some functions contribute to variable outcomes of the whole system.

The Functional Resonance Analysis Method is built on four underlying principles (Hollnagel, 2012). First, failures and successes are equivalent in that they are always, at least one hopes, the product of someone intending to do the right thing. A failure is, therefore, a result of an unexpected input to the function or the control to the function being insufficiently robust. Further, an error can only be identified through hindsight, when an unexpected outcome occurs. Second, individual and collective human performance is adjusted to match conditions. That is, humans often use variability of performance as an asset in response to changing internal conditions (including physiological and psychological conditions) and external conditions (organizational, social, contextual, and environmental factors). Third, system outcomes are emergent, not resultant. In other words, when an unexpected outcome occurs, it is likely not explainable using decomposition and causality, but instead emerges from a non-linear and partly intractable system. Finally, complex socio-technical systems may not occur through a predetermined set of cause-effect links but instead consist of coupled and interdependent functions that may develop differently from one specific situation to another. Taken together, the FRAM suggests that socio-technical systems vary due to purposeful human behavior intended to handle varying conditions, which may propagate or resonate throughout the system. Thus, the FRAM is ideal for analyzing systems fitting these circumstances.

As will be elaborated upon in a subsequent section, in the context of a grading system, graders select and employ some number of cognitive processes depending on a number of variables related to the nature of the content being graded and the quality of the students' responses (Black et al., 2011). As such, a grading system presents variable conditions that result in variable use of different cognitive functions which will vary across from one situation to the next. The earlier generation HRA techniques provide less flexibility to analyze a system as dynamic as grading seems to be and limit the extent to which sources of variability can be identified and understood. The FRAM, which will be detailed more thoroughly in the Methodology chapter, therefore seems like the most appropriate HRA technique for this particular system.

It should be noted, however, that later generation HRA techniques generally dismiss quantitative analyses partly due to the nature of the tasks to which they are typically applied. These techniques come from fields like industrial and nuclear engineering where unexpected outcomes

due to error are relatively rare and can be extremely costly (Rasmussen, 1985). It might be a reasonable question to ask why techniques developed for industrial applications such as Risk Assessment would be relevant or appropriate in the context of educational research. Grading, especially in the context of this study, results in a far greater frequency of relatively low-cost erroneous actions. This means that the data available to analyze are far richer than in traditional applications. Further, just as in industrial systems, grading systems “fail” due to issues with either the equipment (i.e., rubrics and assignments) or the users of the equipment (i.e., graders). This means that this analysis can be done even more effectively in this setting than in more traditional settings. Additionally, it will be meaningful to compute at least rudimentary quantitative measures of the consistency of outcomes for the system. This is particularly possible when the situational factors are nearly identical, as can be done in experimental settings.

2.5.1 Measuring variability in grading

One common approach to measuring reliability of grading is inter-rater reliability. Inter-rater reliability can be classified within three general categories: consensus estimates, consistency estimates, and measurement estimates (Oakleaf, 2009). Consensus estimates assume that reasonable graders can agree exactly on how to use a rubric (Oakleaf, 2009). Consistency estimates allow for variation of interpretation, as long as each grader’s scales are consistent with one another (Oakleaf, 2009). Measurement estimates, on the other hand, develop a summary score of multiple ratings to incorporate all discrepant interpretations (Oakleaf, 2009).

The type of estimate used dictates the calculations that must be performed to determine reliability. Acceptable calculations for consensus estimates include the percent agreement, Kendall’s coefficient of concordance, and Cohen’s kappa (Oakleaf, 2009). Acceptable consistency estimates include Pearson’s r , Spearman’s ρ , and Cronbach’s α (Pantzare, 2015; Stemler, 2004). Measurement estimates are more complex, including approaches such as principle components analysis, generalizability theory, and facet rater severity indices and fit statistics (Stemler, 2004). Individual estimates, however, can potentially be misleading, so it is often good practice to conduct multiple estimates (Stemler, 2004).

Two other consistency estimates that have been used when analyzing grading are Mean Actual Difference (MAcD) and Mean Absolute Difference (MAbD) (Suto & Nádas, 2007). The MAcD finds the difference for each score between the “definitive score”—that is, the score that is

determined by a principal examiner—and the grader’s score and averages across all grading instances. This indicates whether the grader is, on average, more stringent or more lenient than would be expected or desired. The MAbD takes the absolute value of each of those differences and averages across all grading instances. This measure indicates the average magnitude of difference between scores and could produce a notable value even if the MAcD is zero.

These different estimates highlight that inter-rater reliability does not necessarily imply inter-rater agreement, given that the latter requires consensus where the former only demands consistency (Pantzare, 2015). Consistency may be sufficient for achieving fairness from the students’ perspectives, but consensus of grader interpretations and conclusions is necessary for the goal of confidently evaluating performance of specific learning objectives (Pantzare, 2015). As such, it may be useful to consider both consistency and consensus measures.

These measures of inter-rater reliability adopt the observed ratings tradition, but alternative approaches to measuring grading quality adopt what is known as the scaled ratings tradition (Wind & Peterson, 2018). The observed ratings tradition espoused by typical inter-rater reliability metrics assumes that ratings can easily be decomposed to identify specific sources of measurement error. The scaled ratings tradition, on the other hand, suggests that grading is a nonlinear process that is influenced by facets of the graders, the students, and the items. The aggregate level analysis of inter-rater reliability measures alone is insufficient to understand individual graders or to improve grading quality (Wind & Peterson, 2018). It is possible, however, that a combination of the two traditions can lead to richer findings. As such, an initial investigation of grading behaviors with the FRAM, followed by quantitative measures summarizing the findings may lead to stronger understanding of what contributes most to grading quality.

2.6 Synthesis and Summary

The grading process as a complex system consisting of multiple human and non-human elements, each of which may contribute to the variable performance of the system. Thus, it is reasonably classified as a socio-technical system. As such, I have selected a method, the Functional Resonance Analysis Method (FRAM), that was created for the purpose of understanding how variability occurs within a socio-technical system. The FRAM is a highly adaptable and dynamic approach for exploring and modeling complex systems.

The system contains multiple agents (i.e., types of people). The individuals who design the course materials (assignments and rubrics), the team who teaches the content, the students who complete the assignments and produce the work that is graded, and the graders. The assignments, rubrics, and student work can all vary as a result of variable actions taken by the corresponding agents. For instance, the assignments can vary in terms of length, open-endedness, and difficulty. These factors affect the intrinsic cognitive load—which, as a measure of inherent complexity, approximates the measure of cognitive demand—and the extraneous cognitive load imposed on the students to complete the tasks. According to the research presented in this chapter, these factors should affect the breadth of the quality of work produced by the students, where greater cognitive demand and extraneous cognitive load are likely to lead to a wider spectrum of student work quality.

Similar to the assignments, rubric variability may lead to subsequently variable actions. The rubrics vary with respect to their length, clarity, complexity, and robustness, all of which also affect the cognitive load and demand imposed upon the graders while they grade, in part by designating the types of cognitive strategies they must employ. Additionally, as the graders have to make evaluative judgments and decisions, the environment (which can impose extraneous cognitive load), their personal traits and dispositions (e.g., testosterone levels, altruism, and how they feel about those whom they grade) and their emotional states (either pre-existing or anticipated as a result of their impending decisions) can affect their grading decisions. Ultimately, as humans making decisions, they can be expected to make the decisions that produce positive feelings and minimize negative feelings.

Of these factors, some are directly manipulable, some are indirectly manipulable, and some are entirely non-manipulable. A model of all the functions involved in the grading process and exploring where and how it varies facilitates identification of the potential causes of variability that are directly or indirectly manipulable. Applying that model to direct observation, incorporating multiple measures of function and system outcome reliability, will highlight which of these sources of variability are most relevant.

Most of the considerations addressed throughout this chapter present an idealistic perspective that likely disappears in the context of grading with many graders at a large scale. Many of the ideas concerning assignment and rubric design become increasingly difficult when more instructors, students, and graders interact with the materials. Having more agents,

particularly in a heterogenous population, increases the likelihood that content will be misunderstood or misinterpreted. Further, increasing the likelihood across a larger number of instances translates to a growth in the overall number of unfavorable outcomes. This rings true both with respect to interactions with grading documents and natural variability of human behavior, particularly due to randomly occurring erroneous actions and those due to external influences. Suffice it to say, scale complicates and accentuates all of the challenges to reliability within the system. Some complications will always be beyond control, but an understanding of the gamut of challenges to reliability highlights what can be manipulated to improve system consistency, both ideally and practically.

3. METHODS

Given the adopted conceptual framework, the development of an understanding of the grading system requires an appropriate application of a socio-technical systems analysis approach. As such, this study employs the Functional Resonance Analysis Method (FRAM), intended to fully describe the system and its potential sources of variability (Hollnagel, 2012). This chapter details how the data were collected and analyzed to develop the overall mode, work-as-imagined instantiations, and work-as-completed instantiations based on direct observation.

3.1 Research Design

As illustrated in Figure 3.1, this study consists of two main stages of analysis: the development of a general cognitive process model of the grading process and work-as-imagined instantiations of the model using the FRAM and the identification and analysis of work-as-completed instantiations of the model based on observable grading events (collected through think-aloud interviews). That is, this study first focuses on exploring rich qualitative data coupled with personal experiences and knowledge of the system to identify functions within the grading process, allowing for general descriptions of potential variability in the system. It then further explores the think-aloud interview data to apply the model to separate observed instances to determine how the system actually varies in practice. A final third stage synthesizes the findings of stages 1 and 2 to identify the greatest sources of variability and generate possible control mechanisms.

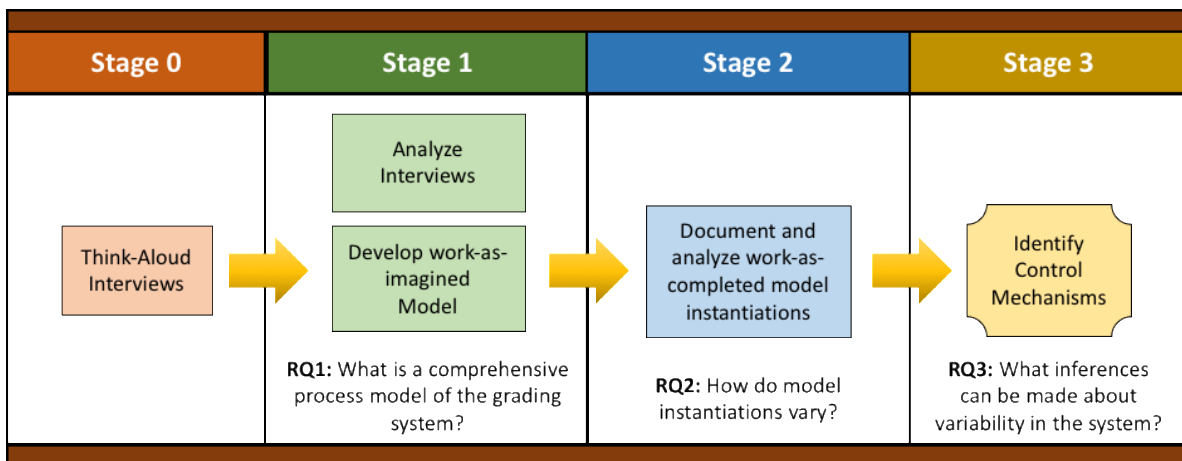


Figure 3.1. High-level overview of research design aligned with research questions.

The first stage of this research is primarily concerned with understanding the general processes in the grading system. As such, specific details about the frequency of variable outcomes of functions within the system are not important; rather, it is helpful to understand how the system operates under ideal conditions given the outcome of some background functions. However, because the number of grading instances and probability of variable judgment is high in this context, some amount of quantitative analysis is reasonable. Thus, the second stage of research analyzes actual instantiations of the model (i.e., work-as-completed instantiations) to understand how the system actually varies in practice and how that variability affects the variability of the overall system output. Finally, once the work-as-completed instantiations from the second stage of the project have been completed to highlight the most relevant factors from the model, the final step of the FRAM can be completed. This step takes knowledge of model instantiations to identify possible explanations for observed variability and produce recommendations for control mechanisms.

3.2 Context

Many large engineering programs have required first-year engineering courses. At the university where data was collected a small number of students enroll in honors or service-based alternatives; however, the majority of students take a two-semester course sequence. The first course covers topics including descriptive statistics, modeling, and design while the second course covers logic, programming tools, and applications of descriptive statistics, modeling, and analysis. This study focused on the grading performed in the second course.

3.2.1 Course details

In the spring semesters, when most students take the second course in the sequence, the course has over a dozen sections. Each section full section of 120 students utilizes an instructional team consisting of an instructor, a graduate teaching assistant (GTA), four undergraduate peer teachers (PTs), and two undergraduate graders. While the instructor and GTA will occasionally deliver up to two sections, the undergraduate PTs and graders work in only one section.

In an effort to make assessments meaningful, the course coordinators have attempted to design the assessments to align strongly with the course's learning objectives. Throughout the

course, students work on problem sets, project milestones, and exams, all centered around the course learning objectives. Collectively, the course covers approximately 20 major learning objectives, each with a set of sub-learning objectives, totaling close to 90 distinct learning objectives (the exact number changes each term as learning objectives are revised). Some of these sub-learning objectives are graded multiple times throughout the semester, while a few just provide guidance to students and are never graded directly.

3.2.2 Grading

Grading, particularly on the problem sets, is performed primarily by the undergraduate PTs and graders. While it varies by section and problem set, the graders typically will grade the majority of the assignments and the remaining assignments are distributed amongst the PTs for grading. Each graded assignment consists of a number of possible points (typically 10 for problem sets), spread across some number of relevant learning objectives.

Each learning objective is graded individually based on the level of achievement of that learning objective. A rating of “Proficient” represents full achievement of the learning objective and earns full points. A rating of “Developing” is awarded to work that is close to full achievement and receives 80% of the total possible points for that learning objective. A rating of “Emerging” shows at least partial achievement of the learning objective but corresponds to demonstrating about 50% of the requirements. “Insufficient Evidence” is given to students who attempt to answer the problem but show little to no evidence that they have achieved the learning objective. Finally, a student response is classified as “No Attempt” if the student did not provide a remotely relevant response. Both “Insufficient Evidence” and “No Attempt” receive 0 points but are separated to allow finer granularity of evaluation data.

3.2.3 Rubrics

The rubrics, an example of which is shown in Figure 3.2, are designed to indicate the relevant learning objective, the specific portion of student work that is to be evaluated with the rubric (e.g., the linearization of the power function in problem 2, step 5), and a list of “evidence items” (i.e., specific pieces of evidence associated with achievement of the overall learning objective) that must be demonstrated to achieve a “Proficient” rating, as well as the number of

evidence items necessary to achieve each of the lower ratings. The “What to Grade” portion also includes a suggested solution in red text and additional instructions or guidance in blue text. Further, within the list of evidence items, black text is used for evidence items that are generally associated with the specific learning objective (and would remain the same if applied to a different problem) while blue text is used to give specific instructions relevant to the problem at hand.

As the example in Figure 3.2 shows, this course’s rubrics all establish a comprehensive set of observable pieces of evidence (i.e., evidence items) that constitute proficient performance of the learning objective. The rubrics then differentiate performance levels based on the number of evidence items not observed within the student’s work, in the area specified by the “What to Grade” portion. This is not necessarily a typical format for a rubric—it is more common for each performance level to have its own unique description. In fact, this particular style of rubric was not encountered in the review of the literature.

Learning Objective	13.07 Linearize and plot data appropriately					
What to Grade:	<div>PS07_beach_logins.pdf > LINEARIZED DATA</div> <div>Grade the linearized data on the linearized data plot. NOTE: do not grade the regression line or any formatting other than what is below.</div> <div>% linearize the data for use in power function log_offshore = log10(offshore); % log of offshore distance log_depth = log10(depth); % log of water depth</div> <div>% Plot linearized data figure(2) plot(log_offshore,log_depth,'g*') xlabel('Log (Offshore Distance in Meters)') ylabel('Log (Depth in Meters)')</div>				Prob 2, Step 5	
Proficient	Developing	Emerging	Insufficient Evidence	No Attempt		
1 pt	0.8 pt	0.5 pt	0 pt	0 pt		
Evidence items for proficiency: 1. Linearize the independent variable data correctly based on the diagnosed function type <ul style="list-style-type: none">Power: log of independent data 2. Linearize the dependent variable data correctly based on the diagnosed function type <ul style="list-style-type: none">Power: log of dependent data 3. Axes labels (description and units) are correct based on the plotted data You will need to see what they plotted to see if their units match what's in their plot command. You'll grade whether or not they plotted the correct information in the next LO	1 (of 3) missing or incorrect item from the proficient list	2 (of 3) missing or incorrect items from the proficient list	3 (of 3) missing or incorrect items from the proficient list	Did not attempt the graded item		

Figure 3.2. Example of rubric used for grading in the course.

3.2.4 Training

The week before students submitted each problem set, the undergraduate PTs and graders were expected to engage in online training modules for each new learning objective (Figure 3.3) and corresponding rubric item (Figure 3.4), as well as lists of common mistakes and related topics that were not covered by the learning objective (Figure 3.5). Each module included an example

problem (Figure 3.6) and a correct solution for that problem (Figure 3.7). Additionally, for each module, the training gave two samples of simulated student work (Figure 3.8). For each sample of student work, the PTs and graders were asked to complete a quiz in which they selected the achievement level and the evidence items they perceived as not adequately demonstrated (Figure 3.9). They were also expected to write what written feedback they would give the student.

Learning Objective (LO): 13.05 Create plots with linear and/or log axis scales (Excel)
Evidence of Proficiency Requires:
Plots of data using different axis scales to show relationships useful for function discovery
Linear scale: linear scale on x-axis, linear scale on y-axis
Log-linear scale: log scale on x-axis, linear scale on y-axis
Linear-log scale: linear scale on x-axis, log scale on y-axis
Log-log scale: log scale on x-axis, log scale on y-axis
Function discovery plots display original independent and dependent data (i.e., non-linearized data) whose relationship is being examined
Each plot has x- and y-axis labels that reference the data in the plot and do not reference the type of scale used
Show the minor gridlines on log scaled axes
Manage the horizontal axis crosses option so that the x-axis tick labels are at the bottom of the plot
Manage the decimal places shown on the x and y axis tick marks

Figure 3.3. Learning objective as shown in training.

It is important to note that due to the development and timing needs of assignments and rubrics, it was not always possible for the sample problem to be identical to the problem that the PTs and graders actually graded in the problem sets. In some cases, the training modules were designed before the assignment were finalized. Further, when problems covered multiple learning objectives, only the learning objectives that had not been previously assessed were included in the training. For these new problems, there were no previous actual student responses, so it was necessary for the instructional team to simulate artificial student work based on expected responses.

LO Grading Instructions and Rubric

The student solution must be evaluated for each of the 9 items of proficiency evidence.

Note: When grading the some pieces of evidence for this learning objective, you might be directed to just look at one of the plots (rather than all four). For this training assume that all four plots must demonstrate each evidence item. If any of the fours plots does not demonstrate a particular piece of required evidence, the evidence item is not met.

Proficient	Developing	Emerging	Insufficient Evidence
<ul style="list-style-type: none"> Plots of data using different axis scales to show relationships useful for function discovery <ul style="list-style-type: none"> Linear scale: linear scale on x-axis, linear scale on y-axis Log-linear scale: log scale on x-axis, linear scale on y-axis ... 	1-2 (of 9) missing or incorrect items from the proficient list	3-4 (of 9) missing or incorrect items from the proficient list	5 or more (of 9) missing or incorrect items from the proficient list

Figure 3.4. The learning objective description and rubric item from training.

Not Assessed by this LO:
Format of the plot for technical presentation

Common Student Mistakes:
Students will label the x and y axis as $\log(x)$ or $\log(y)$ if log scaling is used, where x and y would be specific to the context of the problem. It must be just x and y as there has been no transformation of the data.

Figure 3.5. What is not assessed by the learning objective and common student mistakes.

Problem

The student is asked to use function discovery and data transformation to model the relationship between earthquake intensity and magnitude. The student is provided with a dataset containing moment (in giganewton-meters or GN-m) and magnitude.

The student must plot the data on all combinations of linear and log scale axes.

Figure 3.6. Example problem associated with the learning objective.

Solution

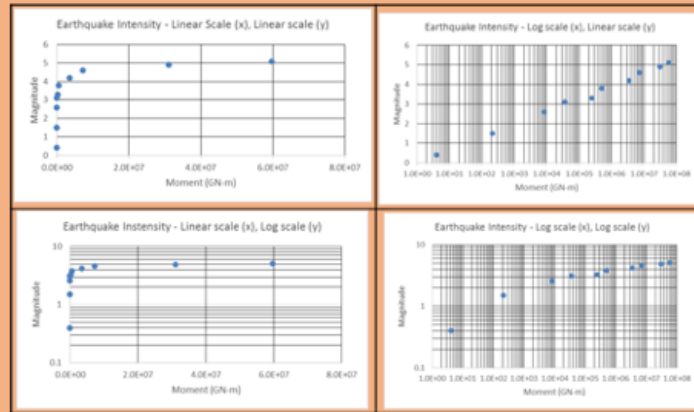


Figure 3.7. Provided solution to sample problem.

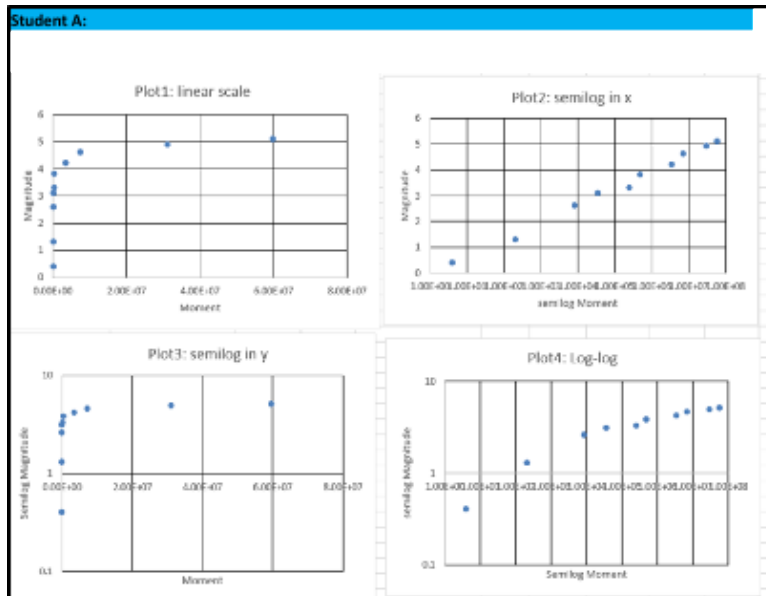


Figure 3.8. Sample A of simulated student work.

Preview Test: LO 13.05 - Quiz 1

QUESTION 1

Mark the level of achievement demonstrated by Student A's work:

- ☐ Proficient
- ☐ Developing
- ☐ Emerging
- ☐ Insufficient evidence

QUESTION 2

For Student A's work, which of the following pieces of evidence will you need to provide a comment on? Select all that apply.

- ☐ Log-log scale plot: log scale on x-axis, linear scale on y-axis.
- ☐ Manage the decimal places shown on the x and y axis tick marks.
- ☐ Show the minor gridlines on log scaled axes.
- ☐ Linear scale plot: linear scale on x-axis, linear scale on y-axis.
- ☐ Manage the horizontal axis crosses option so that the x-axis tick labels are at the bottom of the plot.
- ☐ Function discovery plots display original independent and dependent data (i.e., non-linearized data) whose relationship is being examined.
- ☐ Log-linear scale plot: log scale on x-axis, linear scale on y-axis.
- ☐ Each plot has x- and y-axis labels that reference the data in the plot and do not reference the type of scale used.
- ☐ Linear-log scale plot: linear scale on x-axis, log scale on y-axis.

Figure 3.9. Quiz for Sample A.

3.3 Study Participants and Data Sources

The think-aloud interviews in the first stage were conducted in the spring semester of 2017. First, an initial protocol was developed for the think-aloud interviews. Two pilot interviews were conducted with two graduate teaching assistants who were contacted individually and willing to participate. These pilot interviews informed revisions to the layout and design of the think-aloud documents that would allow participants to more efficiently grade a larger number of student samples. The interview documents will be discussed more thoroughly in a subsequent section.

Following the pilot interviews, all undergraduate teaching assistants were emailed near the end of the semester to ask for participants in the study (the original email is included in appendix A). Participants were ensured that their participation would be anonymous and not reported to any members of the instructional team to encourage authentic participation and honest perspectives. They were also offered \$20 for their participation. In total, 76 undergraduate TAs were contacted, 21 responded, and 17 ultimately participated in the study. In addition, interviews were conducted with three faculty instructors and an instructional support team member who was involved in the development of the assignments and rubrics.

3.4 Detailed Design

Figure 3.10 provides a more detailed diagram of the study design, including additional steps within each stage. Each of these steps will be explained in detail in the subsequent subsections. The alignment between each stage of the research, the research questions, the general purposes, and sources of data used or generated are summarized in Table 3.1.

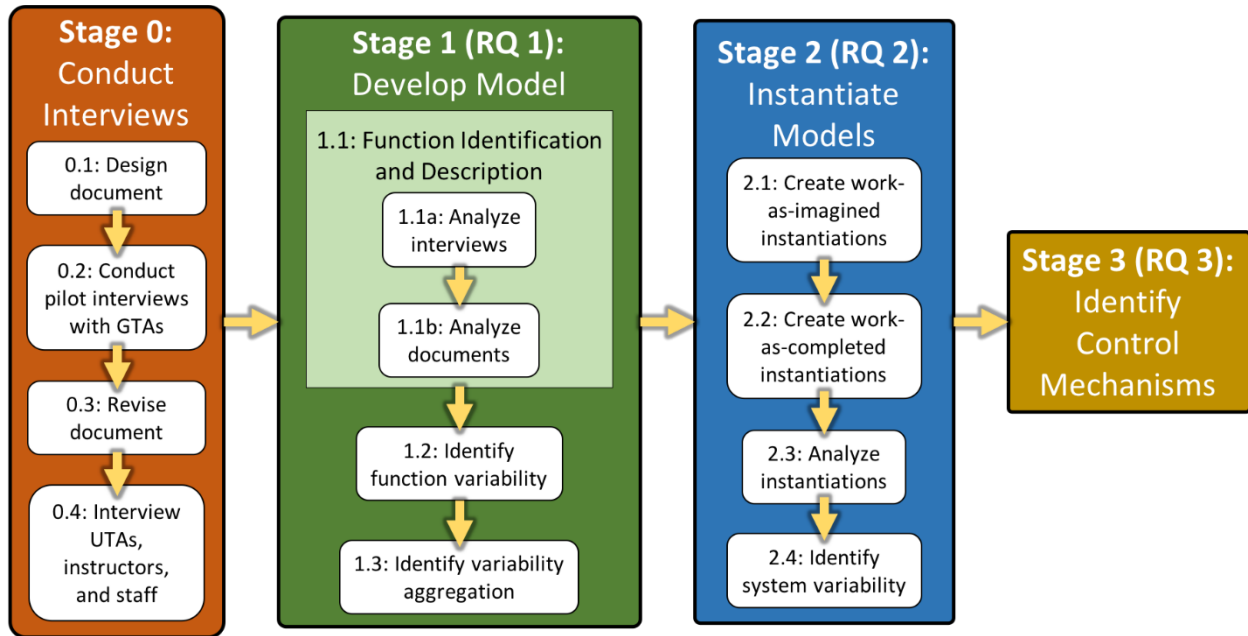


Figure 3.10. Detailed study design diagram.

3.4.1 Stage 0: Think-aloud interviews

In order to conduct the FRAM, it is beneficial to directly observe specific instances of the grading process. The purpose of the FRAM analysis is to understand the overall system and the way individual implementations of the functions within the system vary in a controlled environment with several known variables (i.e., the assignment, rubric, and student work). By controlling features that could each contribute to system variability, it is possible to observe the variability introduced by the graders, themselves. Thus, controlled, observational data effectively provides insight for the FRAM model. To identify the functions the graders are using, a structured set of think-aloud interviews were conducted.

Table 3.1. Summary of processes and purposes aligned with research questions and data sources.

Stage	Research Question(s)	Purpose	Data Source
Stage 0: Conduct think-aloud interviews	RQ 1: What is a comprehensive process model of the grading system?	Interview graders to provide direct observation of the grading process in a controlled environment, reducing factors that cause variability.	Created from Spring 2017 course materials
Stage 1.1a: Analyze interviews	RQ 1a: How does the model extend existing models of grader cognition?	Analyze the interviews qualitatively to identify the different cognitive functions being utilized by the graders while grading.	Think-aloud interview audio and text (<i>stage 0 output</i>); literature
Stage 1.1b: Analyze documents	RQ 1b: What relevant processes occur outside of grading?	Analyze the documents qualitatively to identify the different cognitive processes (i.e., functions) involved in their design and development.	Think-aloud document (<i>stage 0</i>)
Stage 1.2: Identify potential variability	RQ 1c: How can the system vary?	Examine the outputs of each function to identify all the possible ways each of the functions in the grading system could theoretically vary.	Model functions (<i>stage 1.1 output</i>); literature
Stage 1.3: Identify variability aggregation	RQ 1d: How can variability propagate?	Examine possible connections between functions to identify the potential resonant effects of function variability on overall system performance.	Model functions (<i>stage 1.1 and 1.2 outputs</i>)
Stage 2.1: Create work-as-imagined instantiations	RQ 2a: How does context affect the system? RQ 2b: How does work-as-imagined differ from work-as-completed?	Make ideal model instantiations for each rubric-sample pair to represent grading process for maximally accurate outcome.	Think-aloud document (<i>stage 0</i>)
Stage 2.2: Create work-as-completed instantiations		Apply the FRAM model developed to each grading instance observed in the interviews to develop all work-as-completed instantiations and allow for comparison with work-as-imagined instantiations.	Model (<i>stage 1 output</i>) applied to think-aloud interviews (<i>stage 0 output</i>); work-as-imagined models (<i>stage 2.1</i>)
Stage 2.3: Analyze instantiations	RQ 2c: Which contextual factors contribute most to variability?	Compare work-as-completed instantiations to work-as-imagined instantiations in groups based on characteristics of the contextual factors to identify factors with the greatest impact on variability.	Model instantiations (<i>stage 2.2 output</i>)
Stage 2.4: Identify system variability	RQ 2d: How resilient is the system to internal variability?	Examine the relationship between process variability and the accuracy of the overall output of the system to identify how robust the system is to internal variability.	
Stage 3: Identify control mechanisms	RQ 3a: What mechanisms might dampen variability?	Using the collective analyses, determine which factors contribute most to unreliable grading outcomes to identify possible mechanisms to reduce variability that are grounded in the academic literature.	Observed variability (<i>stage 2.3 and 2.4 outputs</i>); literature

Think-aloud interviews are an excellent way to gain insights into the implicit thought processes people utilize as they work through a cognitively oriented activity. For this project, the think-aloud interviews were conducted following the recommendations of Boren and Ramey (2000). As Krahmer and Ummelen (2004) note, the approach of Boren and Ramey (2000) allows the interviewer more room to interject to ask for clarifications in the event that a participant is being unclear or insufficiently communicative than is allowed by the more traditional think-aloud method of Ericsson and Simon (1993). In either approach, however, it is important for the interviewer to play a minimal role beyond encouraging verbalization from the participant so as to reduce their impact on the participants' thinking. As such, during these interviews, the primary interjections were to encourage the participants to continue to vocalize their thoughts, with the occasional need to ask the participants to clarify or elaborate on their thinking when the reasons behind their decisions were, in the moment, seemingly unclear.

Beyond obtaining just the thoughts of the participants, it was also important to understand what features of the rubrics and student samples the graders were paying attention to or found confusing. To capture this information during the interviews, the participants were given an iPad with the think-aloud document loaded into the Notability app. The Notability app allowed the participants' voices to be recorded as they verbalized their thinking (audio was also recorded on the interviewer's iPhone for redundancy). Meanwhile, the Notability app allowed the participants to use a stylus to make annotations on the document, which were synchronized with the audio recording upon playback. As the participants were also instructed to highlight aspects of the student work that contributed to their grading decisions and to indicate their final grading decisions on the document, the participants' annotations and audio recordings serve as a rich source of data to identify the participants' thoughts and decisions.

Stage 0.1: Document design

The first document to be developed for the think-aloud interviews was the problem set document. When the problem set was selected, only the first 10 problem sets had been completed. TA performance on training was used as a proxy for likelihood of grader error across a problem set's learning objectives in order to select the problem set that appeared most challenging for the graders to grade accurately.

The ranks of four different measures of performance were averaged: the percent agreement (i.e., the percentage of TA assigned scores that matched the definitive grade), the Mean Absolute Difference (i.e., the average of the absolute values of the differences between assigned scores and definitive scores), the average number of evidence items incorrectly identified (i.e., the total number of evidence items marked as achieved when the definitive mark considered that item not achieved, and vice versa), and the standard deviation of TA's selected scores (such that a small standard deviation would suggest stronger agreement across graders and a large standard deviation would indicate variable scoring decisions). Problem Set 7 (PS07) and Problem Set 9 (PS09) both had scores of 7.25 using this system. The advisers for this study recommended using PS09, as its focus on user-defined functions was more open-ended than PS07, which focused on plotting with linear and non-linear regressions.

The next aspect of the document that needed to be designed was the selection of the student samples. With access to the submitted work from one section of the course, three samples were identified for each learning objective that demonstrated different student responses of varying quality. The intention of this process was to meaningfully represent a variety of solutions presented by students that could potentially be interpreted variably by the graders.

Finally, the protocol for the interviews needed to be designed. The previously identified problems and samples were initially organized into a single PDF document that was uploaded into the Notability app. The order of the documents was originally a full problem set rubric, followed by a set of student responses to each problem, repeated three times. The problem set itself and the suggested solutions to the problem set were printed and given on the side so that the participant would have an easier ability to juggle through the documents.

Stage 0.2: Pilot interviews

Two pilot interviews were conducted with graduate teaching assistants who volunteered to help. Upon these interviews, it was determined that the arrangement of the documents in the PDF required the participant to shift back and forth through many pages constantly. This was extremely inefficient and severely limited the number of samples the GTAs were able to grade in an hour. The GTAs recommended rearranging the document such that less time would be needed to navigate back and forth through the document.

Stage 0.3: Revised document design

Following the pilot interviews, the interview document was revised such that a single rubric item (or pair of rubric items, when both corresponded to the same portion of work) were shown at once, followed by a single sample of student work. This was then repeated three times for each rubric item or pair of rubric items. As a result, participants could move back and forth between a page or two when looking at a rubric and looking at the sample work rather than having to move through many pages with the previous document structure.

Stage 0.4: Data collection

Once the interview document was finalized, the actual interviews were conducted over the course of two weeks. In total, 17 undergraduate TAs participated, as well as three faculty instructors, and one instructional support team member who had been involved in the development of the assignments. The TA interviews ranged from 30 minutes to one hour, with most participants requiring the full hour. A few participants were unable to complete the grading in one hour and the support team member and each instructor required more than an hour to complete the grading. Participants' notes and utterances were recorded using an iPad and the Notability app, as noted previously.

3.4.2 Stage 1: Model Development

The data collected through the think-aloud interviews, along with knowledge of and experience performing roles of other agents in the system (i.e., instructor, content developer, and teaching assistant), collectively contributed to the development of a FRAM model of the grading system. The first three steps of the FRAM include identifying and describing the functions in the process, identifying sources of variability, and considering how variability aggregates throughout the system. These steps occur in a concurrent fashion based on experiential knowledge of the system, which can be informed by observational evidence (Hollnagel, 2012). Through initial model development, a “work-as-imagined” general model was created. Then, specific observations of instances of the system occurring allowed for “work-as-completed” instantiations of the model to be created to represent those instances (Hollnagel, 2012). Thus, while the first three steps of the FRAM (stages 1.1, 1.2, and 1.3, respectively) are presented linearly, they occurred concurrently

rather than sequentially. Finally, it should be noted that Chapter 4 will be devoted to the results of Stage 1, as it thoroughly describes all of the functions within the model, how they can theoretically vary, and how they can relate to one another.

Stage 1.1: Function identification and description (FRAM step 1)

The first step of the FRAM is to identify and describe the functions involved in the process being modeled (Hollnagel, 2012). The term “function” refers to the activities conducted in order to achieve a specific outcome—which, in the context of this project, is to evaluate a student’s performance for achievement of learning outcomes. The term can also refer to procedural activities established by an organization or a process performed by a technological system, either independently or in collaboration with one or more humans. Functions can either occur in the foreground or in the background. Foreground functions are the primary functions in the process being investigated. Background functions, on the other hand, may affect the foreground functions by altering context, but are not the major aspects being explored.

The FRAM approach represents functions using hexagons, with each corner of the hexagon representing a different aspect of the function (see Figure 3.11). These aspects are described by Hollnagel (2012) as follows:

- Input (I): an entity that is processed or transformed by the function, or a state that initiates the function.
- Output (O): the result of the function, which may be an entity or a state change.
- Preconditions (P): required conditions for a function to proceed.
- Resources or Executive Conditions (R): what is consumed or used (e.g., operational procedures) as the function proceeds.
- Time (T): time-based constraints on the function, in terms of start time, finish time, or duration.
- Control (C): monitors or controls for the function.

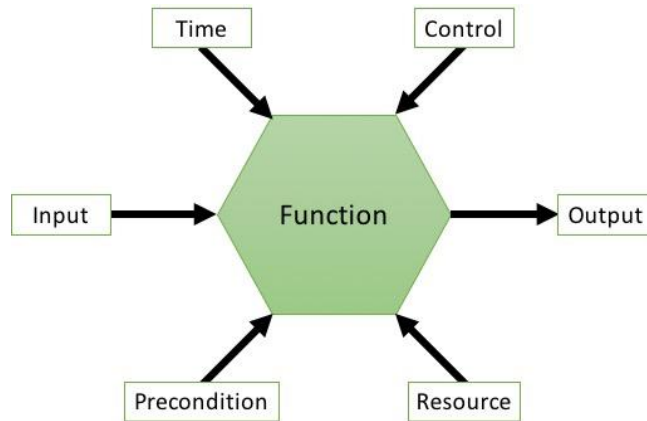


Figure 3.11. The structure of a FRAM function hexagon.

The aspects for each function are identified and organized using FRAM frames (see Table 3.2, below) (Hollnagel, 2012). The frame is a table with a row for each aspect for a particular function. Not every aspect is necessary or relevant for each function; for instance, some functions may not occur under a time constraint or may not require a precondition. Further, the specific designation of some aspects may be debatable, such as whether a particular constraint represents an executive condition or a control. In these cases, the decision to designate a particular aspect as one versus another has minimal impact on the model, as a whole—that is, functionally, the different aspects do not have unique impacts on the function. In terms of understanding the system, it does not matter whether or not it is designated a control or an executive condition. What is important is recognizing that variability of that aspect can affect the performance of the function in question, and either designation captures that effect.

Table 3.2. An empty FRAM frame

Name of function	
Description	
Input	
Output	
Precondition	
Executive Condition	
Control	
Time	

In order to identify and describe the functions that comprise the grading system, it was first necessary to qualitatively analyze each aspect of the system. While the literature presented in Chapter 2 provides a set of possible expected grading behaviors and aspects of assignments and student responses that affect grading outcomes, Hollnagel (2012) specifies that function identification should not be an *a priori* process. Thus, it was necessary to explore the phenomena of grading with an open mind to observe how all the elements of the system interact and to identify the purpose of each. To achieve the open qualitative analysis of the grading process, think-aloud interviews and protocol documents were coded using an iterative, multi-step procedure.

Stage 1.1a: Interview analysis

Four interviews were randomly selected to be coded using an initial process coding technique (Saldaña, 2016, pp. 110–119). That is, specific, descriptive gerund-based codes were assigned for each behavior or task observed in the interviews that related to grading, keeping an open mind for behaviors that may not have been initially anticipated from personal experience. The initial open-coding was followed by a focused-coding process to categorize the initial codes based on similar general behaviors or tasks that occurred during the grading process. Some nuance was retained with sub-codes that represented minor differences within a focused code when it seemed relevant or meaningful to do so (Saldaña, 2016, pp. 239–244). These focused codes were applied to four new randomly selected interviews to ensure that the set of codes were sufficiently representative of the phenomena present.

To verify that the actions represented by the codes were understandable and representative of real-life grading, were present in the interviews, and were not failing to capture important tasks or behaviors, an undergraduate research assistant with extensive context-specific grading experience reviewed the set of codes. This included reading the code descriptions and listening to five randomly selected interviews, separate from the initial four, to verify that the codes and their descriptions were understandable. This process was intended to contribute to the trustworthiness of the identified codes and to improve the quality of the qualitative analysis. Following this process, codes were extensively discussed to achieve consensus with the undergraduate assistant to develop a comprehensive and understandable set of actions utilized in the grading process.

The initial open and focused coding stages were conducted independently and then checked for validity by the undergraduate research assistant, as described previously. As Hollnagel (2012)

emphasizes that the development of a FRAM model is most appropriately performed by a team of knowledgeable members, the codes developed were then translated to FRAM functions through direct collaboration with the same undergraduate research assistant who was highly knowledgeable and experienced as a grader within the course. To translate the codes into functions, it was necessary to consider all of the inputs, outputs, preconditions, controls, executive conditions, and time aspects that were relevant for each function. Further, to make a complete model, any non-output aspect had to be the generated output of another (either foreground or background) function. This required extensive collaboration and discussion about the cognitive processes observed in the interviews in conjunction with analysis of the grading documents and informed by experiential knowledge of the system. Ultimately, no functions were defined without a consensus of interpretation.

Stage 1.1b: Document analysis

Translating the observed behaviors into functions led to a need for other functions to produce the necessary aspects to describe each of those observed functions. For example, if the graders must engage in a function of interpreting the rubric, someone must first write the rubric that will be interpreted by the graders. Thus, as the behaviors of the grading process were being articulated and written as functions, it was necessary to go back and investigate each of the documents with which the graders interact and consider the actions that must have been taken to generate each of those items and who conducted those actions. These actions were not directly observed, but were inferred based on necessity—for instance, the grader would not have a piece of student work to evaluate if a student had not performed the task, first—and based on personal experiences acting in the various roles. As with the function identification for the actual grading process, the background grading system functions were articulated through discussion with the undergraduate research assistant to ensure clarity.

Function definitions and abstraction hierarchies

The initial focused codes were transformed from simple qualitative descriptions into the format of FRAM functions (similar to an axial coding process (Saldaña, 2016, pp. 244–250) but fit within the FRAM framework). This meant that the functions needed to be revised to ensure that

each function represented an action taken by a member within the system and needed relationships to the six aspects of FRAM functions (i.e., inputs, outputs, resources/executive conditions, controls, preconditions, and time). Some of the initial focused codes did not constitute specific actions taken by a member of the system and needed to be revised; for example, one of the initial focused codes was “error spotting,” to represent instances when a grader unexpectedly encounters an error in the student’s response. It was recognized that this was not a unique action compared to other actions, such as “scanning,” but rather a variable output of the scanning function—that is, the output of the scanning function is a determination of whether or not there is an error, which may be performed accurately or inaccurately depending on the alignment of their determination and the student’s work.

In other cases, the initial focused codes went to a level of nuance that did not need to be retained. For example, there were initially five different styles of “matching” identified (block matching, exact matching, individual word matching, memory matching, and number matching). It was ultimately decided that this would be best condensed into only two functions: “determining if a response matches the solution model exactly,” and “determining if a response effectively matches the solution model.” Thus, the translation of focused codes to FRAM functions generally consisted of more thoroughly investigating the identified codes using logic and observations from the interviews to identify which actions constituted essential differences in task versus variable output, followed by careful consideration of what, in an ideal implementation, is required input in order to perform the function, what controls, executive conditions, preconditions, or time constraints affect performance of that function, and what the output of the function should be. This process effectively condensed and reorganized the actions represented in the focused codes.

When considering the foreground and background functions identified, the system and number of people involved is quite large. To help organize the model, FRAM was hybridized with hierarchical analysis to clearly situate who performs each task within the system and identify their underlying goals (see Patriarca et al., 2017). In this approach, functions are defined at different levels of abstraction and with respect to different agents or actors. At the highest level of abstraction, the intention was to describe the functional purpose of each agent involved in the grading process—that is, any person or group involved with the production of artifacts or the delivery of required information, including the assignments, the responses to the assignments, understanding of content, training, and the rubrics. The second layer of abstraction of the system

is the set of high-level actions that must be performed to evaluate the general purpose of each agent. For instance, the functional purpose level task of “designing the course materials” must be broken down into a separate task associated with the design of each artifact used in the grading process, as well as the development of any other items that affect functions performed by other agents in the system. The final, deepest layer of abstraction is the set of cognitive functions. These functions represented the specific tasks required to achieve the generalized functions, primarily in terms of the cognitive processing or decisions made by the members of the system.

While the think-aloud interviews only provided direct evidence of actions taken by graders, it was necessary to identify cognitive functions performed by other members of the system to create a complete model at deepest level of abstraction with sufficient inclusion of necessary inputs. Thus, processes performed by other agents were inferred through personal experiences acting in each role and based on logical necessity of the necessary inputs for grading functions. For example, though there was no direct observation of a student performing an assigned task, it clearly had to occur for the student work to exist.

Stage 1.2: Variability identification (FRAM step 2)

Variability identification is the second step in the FRAM (Hollnagel, 2012). The goal is to consider both potential, expected, and actual variability. The key aspect of concern is the variability of the output of each function. Hollnagel (2012) notes that output variability can result from either inherent variability of the function itself, variability of the environment in which the function is performed, or variability of upstream functions (i.e., variability in inputs, preconditions, resources, controls, or time). These sources of variability can occur in isolation or simultaneously. At this point in the process, only inherent function variability and environmental conditions are generally considered.

Hollnagel (2012) explains that functions may manifest output variability in a number of different ways. The dimensions he identifies are speed (too fast/slow), distance (too far/short), sequence (reversal/repetition/commission/intrusion), object (wrong actor/object), force (too much/little), duration (too long/short), direction (wrong direction), and timing (too early/late, omission). In the context of a grading system, the typical interpretations of these factors may not make perfect sense. Dimensions such as timing (for instance, omitting or neglecting to evaluate a portion of the sample) or object (for instance, grading the wrong part of an answer) can be

traditionally interpreted. However, the development of this model forced consideration of other meaningful ways that the output could vary. For instance, a student's presented performance could vary in terms of predictability (how closely it matches the provided solution), effectiveness (how effectively it achieves the goals of the specified task), or clarity (how clearly the response is presented).

To complete this stage, it was necessary to consider all the ways each outcome of each function could theoretically vary. The think-aloud interview data, as well as extensive discussions with the undergraduate research assistant, were used to be as comprehensive in the identification of possible variation. Additionally, the literature (i.e., Black et al., 2011; Greateorex & Suto, 2006; Suto & Greateorex, 2008; Suto & Nádas, 2010) regarding features that affect grading accuracy, such as the difficulty of the evaluative task, the open-endedness, length, or complexity of expected and acceptable answers, or the length, clarity, or typicality of the student's response were incorporated into potential variability for the basic model. As with Hollnagel's (2012) examples of output variability, these potential variabilities were classified in ways that could be simply operationalized and delineated. This process occurred in an iterative fashion as the functions were developed and articulated. Despite how much the identification of variability was informed by literature, experience, and observation, it is important to note that the possible variabilities identified for each function were not entirely exhaustive.

Stage 1.3: Variability aggregation (FRAM step 3)

The third step of the FRAM is to examine how variability aggregates through the system as a function of upstream-downstream coupling (Hollnagel, 2012). In other words, this process considers how a variable output of one function might affect variability of subsequent functions. The variability introduced by one function may dampen, amplify, or have no bearing on future functions. Thus, this stage looked at the possible variabilities identified in the functions during the previous stage and considered how that variability could possibly affect subsequent functions, in part based on the estimated likelihood of outcomes provided by Hollnagel (2012). Again, this was developed concurrently with the previous stages in an iterative nature based on logic, experience, and observations from the think-aloud interviews. Further, similar to the previous stage, it is likely that the potential variability aggregations identified were not exhaustive, especially given the dynamic and highly variable nature of each model instantiation being difficult to fully anticipate.

3.4.3 Stage 2: Model Instantiations

The second stage of this study is intended to identify how the function may vary in practice, by focusing on a specific assignment, its rubrics, selected student samples, and the way graders grade them. This first consists of looking at the assignments, rubrics, and student samples to imagine an ideal instantiation of grading. Then coding the interview observations with respect to the work-as-imagined instantiations to create work-as-completed instantiations. Finally, it involves an analysis of the work-as-completed instantiations to understand how the system varied in practice.

Stage 2.1: Work-as-imagined instantiations

The development of the general model in the previous stage, while informed by observation and literature, is intended to encapsulate all of the possible ways that each function can vary. What it does not do is show what actual variability occurs within the system or directly how the variability of one function affects subsequent functions, in practice. The sequence of actions taken by the graders should depend on the situational features established by the background functions in the system. As such, it is necessary to consider how the grading should occur for each situation (i.e., each LO-sample pairing) to create work-as-imagined instantiations of the model that demonstrate how the system should change as a result of more directly manipulable variability.

To develop the work-as-imagined instantiations, the functions developed in stage 1 were applied to the materials used within the think-aloud interviews from stage 0 (i.e., the problems, the rubrics, and the sample responses). As each LO specified different evaluative tasks and each sample response achieved the task to a different extent, all 30 LO-sample pairings require unique work-as-imagined instantiations. These instantiations were developed independently through analysis of the documents based on two important principles: (1) the rubric must be interpreted strictly and literally, as any interpretive- or value-based judgments would be expected to vary across graders, and (2) no shortcuts can be taken (i.e., no evaluations skipped) for efficiency's sake, as the primary goal of grading is assumed to be providing completed and thorough feedback to the student, which would require an exhaustive evaluation. After applying these principles to each pair, discussions were performed with an undergraduate research assistant who had four years of

experience with the system to achieve consensus. The “definitive mark” for each sample was also determined and verified through the same process.

To illustrate the development of the work-as-imagined instantiations, consider the following example. In one of the problems used in the think-aloud interviews, the students were asked to develop an exhaustive set of test cases for a programming problem; that is, they need to identify possible values that could be input into the program that should, if the program performs properly, result in each possible output of the program. In this problem, a sufficiently exhaustive list required seven unique test cases. For an ideal grading process of a correct student sample, the grader would need to analyze all seven test cases to ensure that each path through the program would be executed if the seven tests were run. Alternatively, if a student sample only contained four test cases, the grader could theoretically know the objective was not achieved by identifying that fewer than seven cases were present. However, the assumed perspective for an ideal grading process would include giving complete feedback. As such, a grader would still need to look for each required test case to identify which of the test cases was missing. In this way, the model work-as-imagined instantiation for the complete, correct sample would differ from the work-as-imagined instantiation for the incomplete response because the result of the searching for each test case would differ when the test case being sought is absent.

In addition to identifying which functions are used in the work-as-imagined instantiations, it is also necessary to indicate how the output of the background functions vary along the dimensions identified in stage 1. Generally, all variability of outputs from the observed cases were classified relative to one another. That is, for a given variable dimension, unless all the observed cases are indistinguishable from one another, they will be placed onto the spectrum out variable outcomes comparatively. That is important to keep in mind, as what is considered “low” on a given spectrum for this specific context, may not be “low” if the context were expanded to other assignments, rubrics, or student responses within the same course, or expanded to similar items in other courses. Objective measures were used whenever possible to compare the observed cases, but some outputs required more subjective judgment. As it would not be meaningful to describe the process for each dimension of variability across all functions before the functions are defined, these processes are described more thoroughly as variability is presented in chapter 5.

Stage 2.2: Work-as-completed instantiations

Once the work-as-imagined instantiations were created for each LO-sample pairing, each was coded within Excel, as shown in Appendix E. These served as the baseline for coding each of the observed grading instances in the 17 think-aloud interviews. Functions used by the graders as expected in the work-as-imagined instantiations were left unaltered in the Excel spreadsheet. If the expected functions were observed and resulted in an unexpected output, they were highlighted yellow and the cell was added to include how the output varied unexpectedly. If the expected function was not used, the cell was highlighted red. If the functions were used that were not expected based on the work-as-imagined instantiations, they were added into the line for the corresponding evidence item and highlighted in green. As is an inherent limitation of think-aloud interviews, even the most self-aware and diligent participants occasionally fail to vocalize all of their thoughts. As such, some functions' usages were not certain but had to be inferred. In these cases, the cells were highlighted blue. Appendix F shows an example of this coding.

These interviews were coded independently and separately from the undergraduate research assistant. After coding was complete for each LO, the codes were compared and discussed until consensus was achieved across all of the instances. All final codes for each think-aloud grader were compiled into a single spreadsheet for each LO. This allowed for comparisons in the next two stages.

Stages 2.3 and 2.4: Instantiation analysis and system variability analysis

Once the model was applied to code all of the interviews and the work-as-completed instantiations were coded into Excel, some higher-level analysis was possible. Each instantiation was quantified in terms of the number of expected functions not used, the number of unexpected functions used, the number of functions that led to unexpected outcomes, the number of functions demonstrating the need to review the documents, the number of functions demonstrating confusion, and whether or not a holistic approach was used. The number of expected functions not used, unexpected functions used, and number of functions with unexpected outputs were added to determine the total number of deviations from the work-as-imagined instantiations.

The variability of each background function determined during stage 2.1 was also added to the Excel document for each LO and/or EI. By enumerating and compiling all of this data into

a single spreadsheet, sorting and filtering could be used to get a better qualitative sense of how variability in one function contributed to use and variability of subsequent functions. This analysis, along with direct observations of actions taken by the graders during interviews, was considered the most meaningful data for interpreting how the system operates. However, because the number of functions and extensive amount of data available for analysis, three backward multiple regressions were performed to help guide interpretation and presentation of variability in the system: background function variability on deviation from the work-as-imagined instantiations; background function variability on the use of holistic grading; and background functions and deviation from work-as-imagined on the “actual difference” between the definitive mark and the mark assigned by the graders.

While the results of these regressions helped to provide direction for discussion, specific values were not reported—the limited number of observed cases for several of the background variables and the large number of variables causing potentially confounded interpretations prevent the results from being useful beyond guiding discussion of the specific observations. Further, it is important to recognize that the observed variability is specific to the course under study and even further to the assignment and samples selected. As such, the qualitative observations of grader behavior, which can more easily be connected to variable outputs of background functions, provide more generalizable and meaningful insight.

3.4.4 Stage 3: Control mechanism identification (FRAM step 4)

The final stage of the FRAM is to identify the factors or functions that contribute most to unfavorable variability (Hollnagel, 2012). By exploring a rich understanding of how the functions contribute to the variability (as established through the interview analyses), the points of greatest need can receive the most attention. Hollnagel (2012) provides a number of possible explanations for observed variability and mechanisms to eliminate or dampen variability. By incorporating abstraction hierarchies, identifying which functions contribute significantly to variability can help to apply Hollnagel’s suggested mechanisms to specific actions related to rubric design, training procedures, or other broader organizational structures or procedures. As such, after all of the analyses are complete, a prioritized set of recommendations for improving grading reliability will be developed and presented in the discussion. As the results of this final portion of the FRAM will tie together the developed model (chapter 4), the observed instantiations of the model (chapter 5),

and the literature (chapter 2), the possible control mechanisms are presented along with the general discussion (chapter 6).

3.5 Ecological Validity

Research in cognitive engineering or cognitive psychology is typically concerned with the ecological validity of the study (Hoc, 2001). Ecological validity refers to the extent to which one can transfer the findings from a study from the “artificial” context of the study to a more “natural” real-life situation. This form of validity is dependent upon the fidelity of the research context—that is, how closely the context simulates real work situations. The greatest fidelity occurs with field studies, but for simulation studies it is important to have a theoretically grounded objective of generalization. Hoc (2001) argues against conducting overly specific and contextualized research with no generalizability.

Research can generally be broken into two types: basic research and applied research (Hoc, 2001). In cognitive engineering, basic research aims to discover elementary or “microscopic” cognitive mechanisms that are applicable to large classes of situations. These studies often allow for more direct conclusions of causality. Applied research, on the other hand, relates to more specific, more complex contexts and cognitive mechanisms at a macrocognitive level, such as human error. Application deals with specific instantiations of a broader class of situations, which limits the generalizability. However, applied research can frame the specific context as a natural situation of a more general problem, and can reasonably inform basic research, as a result. Still, there are limitations of predictability in applied research due to challenges with making sufficiently accurate models or having sufficient information to specify all variables when models are sufficiently accurate and detailed. Given the differences between basic and applied research, this project clearly constitutes applied research.

Noting that this is applied research, it is necessary to analyze the work domain to understand how closely the artificial situation of the study matches a natural situation. This will ground the generalizability claims of the findings. The analysis of the natural work context and determination differences with the artificial context requires identification of the primary characteristics that were and were not reproduced with a logical or theoretical link to how the differences may influence the observations (Hoc, 2001). It is important to note that, arguably, there can never be a fully natural context for any study, as any act of observation inherently introduces artificiality. To

understand cognition, it is necessary to externalize internal cognitive mechanisms through methods like verbal reports, which unavoidably transform any natural situation.

This study collected observational data through verbal reports in think-aloud interviews. Think-aloud data is susceptible to bias due to the effects of being observed in a laboratory setting. That is, the knowledge of being observed is enough to potentially alter behavior or cognition. Further, the process of verbalizing cognition arguably transforms one's thought processes in automatic, subconscious ways—Ericsson and Simon (1998) illustrate notable differences in conceptual cognitive models of silent thinking versus thinking aloud or describing and explaining one's thinking. Additionally, the setting of the interviews and format of the grading document did not align with the way the participants would naturally grade, though the grading document alterations were necessary to collect a wider data set.

Despite concerns about altered cognition and setting, the procedures in this study are justifiable. While Ericsson and Simon (1998) acknowledge that think-aloud studies do not reflect natural and spontaneous thinking with complete accuracy, their review of several studies suggests that a participant who is asked to verbalize does not experience a systematic alteration in their sequence of thoughts compared to a silently thinking counterpart. Though they acknowledged participants asked to describe and explain their thinking often do change and improve their performance, participants in the think-aloud interviews conducted for this project were simply asked to verbalize their thinking, with minimal interventions beyond reminders to verbalize when necessary. Noting Vygotsky's (1962) argument that verbalizations are disconnected and incomplete representations of inner thought, demanding a task analysis to adequately explicate the complete cognitive process, Ericsson and Simon (1998) claim that observed verbalized thought sequences are often consistent with task analysis models. Still, the generation of functions based on expectations and observations outlined in the FRAM process develops a task analysis model to more holistically capture what might be difficult to observe through the interviews. Lastly, Ericsson and Simon (1998) suggest that natural performance can be, at least adequately, reproduced in laboratory settings.

To further justify, Joe et al. (2011) specifically studied the effects of different approaches of verbal reports on raters evaluating performance assessments. They noted through their literature review that verbal reports were the most widely used method for studies into rater cognition in the contexts of test development, validity studies, and technology usability studies. They argued that

using verbal reports is one of the strongest forms of validity evidence available for assessment studies. In their study, they compared the use of verbal reporting of raters using either the concurrent condition (i.e., verbalizing while simultaneously performing the evaluation), as was conducted in the present study, or the retrospective condition (i.e., verbalizing the cognitive process after performing the evaluation). They found that raters put forth more effort in their rating decisions during the concurrent condition, but often struggled to remember the rationale for their rating decisions in the retrospective condition. Thus, they concluded that the concurrent condition, provides a richer source of information to understand cognition, despite influencing effort.

Beyond the effects of the observation protocol, there are many uncontrollable and difficult-to-measure factors within the grading system that can affect variability in a natural context that are not present in the artificial setting of this study. For instance, the literature suggests that personal factors, such as knowledge of the subject, grading experience, personality traits, mood, or fatigue, can influence grading accuracy (Suto et al., 2011), but many of these factors will vary widely across graders and grading instances in actual grading contexts. Thus, these factors will be accounted for in the potential variability identified in articulating the system but will not contribute to observed variability in the work-as-completed instantiations.

Given the inherent limitations of the data collection technique, the necessary artificiality of the situational context of the grading for the data collection, and the lack of ability to control or observe many everyday variables that can affect grading from the collected data, one could reasonably be concerned about the ecological validity of the study. However, because the primary output of this research is a model, the most severe limitation to ecological validity is within the work-as-completed instantiations and the resulting inferences. The overall model and the work-as-imagined instantiations, informed through experience and theory in addition to the collected observations, will still provide an understanding of how the system can vary in a general sense. The streamlined format of the documents and additional cognitive efforts imposed by the verbal reporting protocol would likely reduce the variability in the system that could be observed, which means that the variability detected within this study are likely a subset of the variabilities that exist in natural contexts. Further, as the focus of the study is on understanding the effects of variability on the system that are controllable through design of documents and training, the effects on variability due to uncontrollable factors will likely always be present; however, articulating their potential impacts in the overall model can still inform resulting recommendations for grading.

The last remaining question is the extent to which the findings of this particular grading system and context can be generalized to other contexts. Certainly, there are specific aspects of the design of the rubrics and the assignments that could limit the generalizability to other grading contexts. However, the intention in articulating the functions within the system was to make them abstract enough to represent design, organizational, and grading decisions that would occur for any system. As such, the work-as-completed instantiation analysis will be most informative for the specific context under study but understanding of potential and actual variability observed can still be indicative of other grading systems given the abstract nature of the functions, themselves.

4. THE FRAM MODEL

The results of this study are split into two chapters: stage 1, the developed model, is presented in the present chapter and stage 2, instantiations of the model, is presented in the following chapter. In other words, this chapter is devoted to answering the research questions repeated in abbreviated form below:

RQ 1: What is a comprehensive process model of the grading system?

RQ 1a: What relevant processes occur outside of grading?

RQ 1b: How does the model extend existing models of grader cognition?

RQ 1c: How might the system vary?

RQ 1d: How can variability propagate?

As described previously, model development began through an initial open coding process (see Appendix B). The initial codes were then used to develop a set of focused codes (see Appendix C), which were revised to a set of high-level focused codes and corresponding nuanced sub-codes (see Appendix D). These focused codes were analyzed through discussion and consensus-seeking with the undergraduate research assistant and with the assistance of the think-aloud interviews and personal experiences in different roles to identify and define the functions that comprised the FRAM model. These functions will be described in the following sub-sections, from the highest level of abstraction down to the lowest and sorted by the agents performing each function.

Each function has an action-based name and description for what the function achieves (Hollnagel, 2012). Each function is characterized with between one and six aspects: input(s), output(s), precondition(s), resource(s)/execution condition(s), time, and control(s). Not every aspect is relevant for every function; for instance, background functions do not have inputs, but generally produce outputs that are used by the foreground functions. It is not always clear cut how any given contribution to a function should be categorized within the six aspects. However, it functionally does not matter—they simply serve as an organizational schema to identify all of the important variables that in some way affect the performance of a function. Ultimately, it does not change the overall understanding of the model whether a contribution to a function is classified as a precondition, execution condition, or control, what matters is a recognition that the variability of that contribution can affect the performance and output of the function.

While there are occasional ambiguities when classifying aspects, Hollnagel (2012) makes considerable effort to differentiate each. Inputs are what the functions use or transform. They are what start or activate the function and may be the data that is used or a signal for the function to begin. Outputs are the result of what a function does and are expected to have a range of potential variability. Preconditions are conditions that should be true or ought to be verified before a function occurs. To differentiate preconditions from inputs, Hollnagel (2012) gives the example of an airplane requiring the input of clearance for departure to initiate take-off, while performing the take-off checklist is a precondition to departure. That is, completing a take-off checklist should be performed before departure, but departure will not occur without clearance. Resources and execution conditions are either consumed or used *while* the function is carried out (as opposed to something that should occur *before* a function, as with preconditions). Hollnagel (2012) suggests execution conditions are typically tools or technology, such as a hammer or computer, but could be something less tangible like a skill or competence (such as an ability to code in MATLAB). Controls are often the result of functions conducted by organizational agents and are intended to regulate or supervise other functions to help achieve the desired outputs (e.g., plans, schedules, procedures, guidelines/instructions). Time could be viewed as a resource or control but is given its own category because of the unique ways the sequencing or deadlines imposed on functions could affect outputs.

4.1 Functional Purpose Level

The highest level of abstraction is the functional purpose level. This level defines the overarching functional purpose of each agent group in the system (Patriarca et al., 2017). In the context of this project, an agent is any person or group involved with the production of artifacts or the delivery of required information, including the assignments, the responses to the assignments, understanding of content, training, and the rubrics. Four unique agent groups were identified within the system: the Instructional Support Team (IST) and course curators, the teaching teams in each section, the students, and the graders. While some individuals may fall into more than one of these categories (e.g., instructors could act as both a member of the teaching team and a course curator), they fall into whichever agent group controls a given function as they perform that task. Although the graders were observed directly through interviews, functions for all agents at this

level of abstraction were developed through an understanding of the system based on personal experiences acting in each role rather than based on observed behavior.

A full model at each level of abstraction is primarily the list of functions included at that level. However, at the highest levels of abstraction, it is easy to provide a visual instantiation of how the functions should ideally interact. This helps to show the aspects of each function, how outputs may vary, and how the variability of those outputs may affect subsequent functions. Figure 4.1 shows the visual FRAM model at the functional purpose level. The instantiation shows each of the four agent groups and their corresponding functions (the IST in yellow, the teaching team in blue, the students in orange, and the graders in green). Each function is a hexagon with each corner corresponding to an aspect type, designated by the first letter of that aspect (except Execution Conditions are represented by the letter of its sister aspect, Resources). Note that not every function has every aspect and that some have many components of a single aspect. For instance, the IST function of “Design course materials” has no inputs or controlling conditions but has several unique outputs, some of which (i.e., the class schedule) affect multiple functions. The red circle around the grader’s output is a consequence of the fact that the function has an output (i.e., an evaluation of the student’s performance), but there are no subsequent functions that utilize that output. Similarly, the IST function has a grey tint to indicate that it is a background function, with only outputs.

While Figure 4.1 shows the idealized system operations at the highest level of abstraction, the actual occurrence of the event represented by the model may not always look this way. For instance, some of the output may be missing or some of the functions may not occur or may not occur properly. Thus, the actual presentation of the model is through a description of each of the functions, including how those functions may vary. Table 4.1 shows the functions that are included at the functional purpose level along with the agents responsible and descriptions. Subsequent subsections go into more detail about each agent group’s functions.

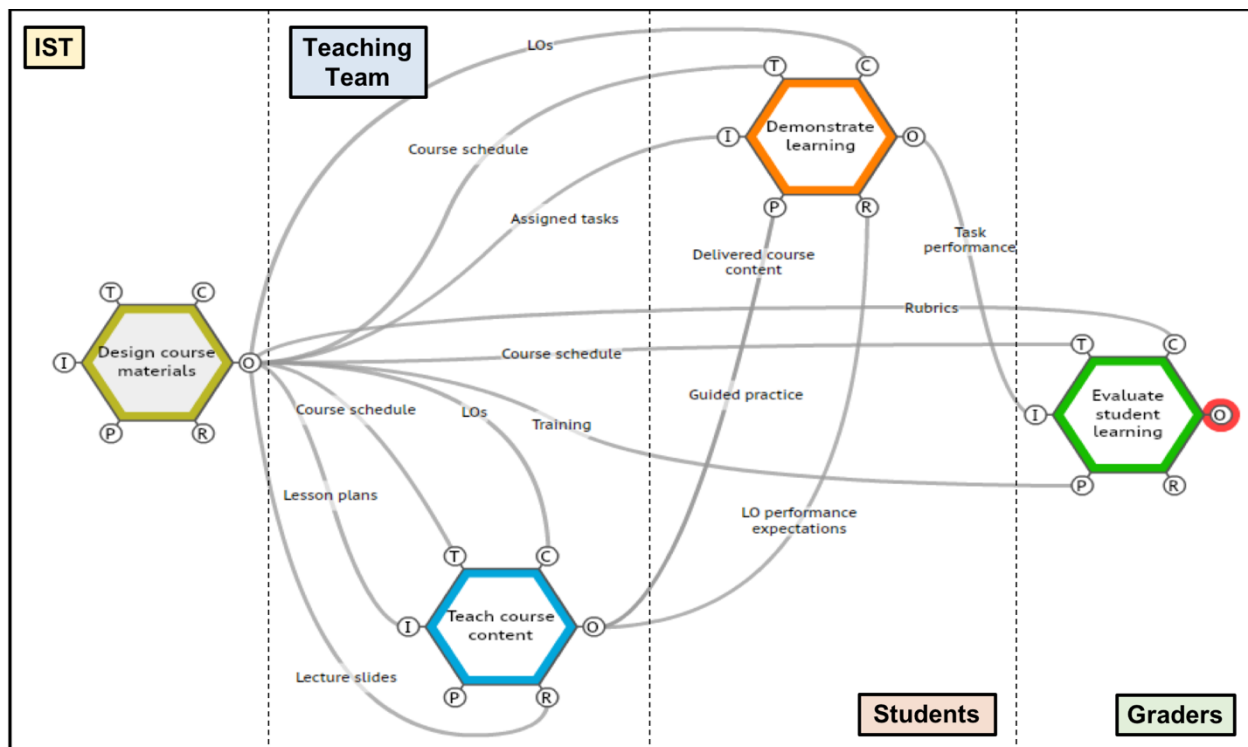


Figure 4.1. Idealized functional purpose level FRAM model.

Table 4.1. Overview of functional purposes

Agent	Function Name	Description
IST	Design course materials	Identify and articulate course LOs and LO EIs, design assignments, rubrics, and training, course schedule
Teaching team	Teach course content	Teach course content and develop LO proficiency through instruction and guided practice
Students	Demonstrate learning	Perform tasks dictated by assessments at the achieved level of LO proficiency
Graders	Evaluate student learning	Assign a quantitative score to student based on evaluation of task performance on each LO

4.1.1 IST's functional purpose

Table 4.2 shows the functional purpose of the IST. At the highest level, this group is responsible for designing all of the course materials. As such, the IST agent group encompasses any individuals who contribute to the development of course materials, even if they also exist in other agent groups (e.g., some instructors or GTAs who are also in the teaching team). Designing course materials includes identifying and articulating the learning objectives for the course and the observable actions that demonstrate achievement of those learning objectives. It also includes

designing a schedule of the course content, the assignments that provide students practice with the course content, the rubrics used to evaluate student performance on the assignments, and the training used to prepare the graders to apply the rubrics. While there are actions that may occur at a higher level that dictate IST's actions (e.g., setting the university's calendar), IST's actions will be considered the initiator for all subsequent functions for this model. Thus, there are no input aspects and only outputs: the envisioned learning objectives, the class schedule, the lesson plans, the assigned tasks, the rubric, and the training materials.

As the IST provides the core background function in this model, the potential variability stems primarily from internal variability of the function itself. That is, there are no upstream functions that could affect variability. Still, the outputs of the IST's function interact with and could affect all other functions (see Figure 4.1 for function interactions). The possible variabilities of each output will be elaborated upon in deeper abstraction levels, but it is important to understand at this level that these outputs can affect downstream functions. For example, if an LO varies in clarity, it can lead to greater variability in the LO performance expectations communicated by the teaching team and the interpretability for the students when completing their assigned tasks. Similarly, if the course schedule varies with respect to the time it provides, too little time could lead to variable content delivery by the teaching team, rushed assignment completion by the students, or rushed grading by the graders.

Table 4.2. Functional purpose of the IST and course curators

Agent:	IST
Function	Design course materials
Description	Identify and articulate course LOs and LO EIs, design assignments, rubrics, training, and course schedule
Input	---
Output	<ul style="list-style-type: none"> • LOs • Lesson plans • Assigned tasks • Rubrics • Training • Course schedule • Lecture slides
Precondition	---
Resource/E.C.	---
Control	---
Time	---

4.1.2 Teaching team's functional purpose

Table 4.3 shows the functional purpose of the teaching team to be the delivery of course content through instruction and guided practice to help students develop the capacity to proficiently demonstrate achievement of the learning objectives (LOs). The teaching team consists of the instructor, the GTA, and the PTs. Note that while the GTA and PTs can also be graders, they serve a different function in the instructional role than in the grading role. Thus, they are seen as functionally separate agents.

In order to accomplish the instructional task, the teaching team utilizes several outputs of the IST function. The teaching team transforms the lesson plan provided by the IST into a set of experiences represented by the delivered course content, the LO performance expectations, and the guided practice. At their disposal, the teaching team has provided lecture slides and is guided by the articulated LOs the students are expected to achieve, within the time dictated by the course schedule.

The teaching team functional purpose is susceptible to multiple forms of variability. External variability occurs through variability of the input, execution condition, control, and time aspects. The function can also vary internally through what experiences and concepts the teaching team delivers and how. For instance, while the lecture slides are provided to the instructor, they have the freedom to modify the slides or activities in class to meet the lesson plan how they believe is best. The GTAs and PTs can provide variable levels of support to the students while they engage in the guided in-class practice activities. Further, there is a possibility that personal interpretation of the LOs and course content on the part of each member of the teaching team can vary and lead to differences in the course content and performance expectations that are communicated to the students. These outputs can have a direct impact on student performance, in particular.

Table 4.3. Functional purpose of each teaching team.

Agent:	Teaching Team
Function	Teach course content
Description	Teach course content and develop LO proficiency through instruction and guided practice
Input	<ul style="list-style-type: none"> • Lesson plans • Delivered course content
Output	<ul style="list-style-type: none"> • LO performance expectations • Guided practice
Precondition	---
Resource/E.C.	<ul style="list-style-type: none"> • Lecture slides
Control	<ul style="list-style-type: none"> • LOs
Time	<ul style="list-style-type: none"> • Course schedule

4.1.3 Students' functional purpose

The overall functional purpose of students within the scope of this investigation is to demonstrate their learning of the course content through the performance of assigned tasks (see Table 4.4). Ideally, the student's work should be representative of their actual level of proficiency. In order to perform this function, students take the assigned task produced by the IST as input to produce their performance of that task as output. It is expected that the teaching team has delivered course content and guided the students through some form of practice during class time before the students begin performing the assigned task. The LOs are provided to the students on each assignment so they are aware of how their performance will be evaluated. The students also should perform the task following their understanding of the performance expectations for each LO that was communicated to them during class. The performance must be completed by the deadline within the course schedule created by the IST.

Each student's demonstration of learning is subject to considerable internal and external variability. Externally, the student's performance can be affected by upstream functions, such as variable delivery of course content or performance expectations, the quality of guidance during in-class practice, or a lack of clarity in the written LOs. Further, performance might be hindered if the deadline in the course schedule is unrealistically early. The function can also vary internally, as whether or not the students receive the delivered course content, guided practice, and performance expectations or the quality to which they do will depend, partly, on their literal and figurative presence in the classroom. There is variability in how well the students understand the

course content and variability in the effort the students will put forth into performing the assigned tasks. Of course, some of the variability in student performance may also be the result of external factors, including, but not limited to, personal fatigue, well-being, and other personal obligations. Ultimately, the students' performances will vary in terms of clarity, overall quality, and how well their performance aligns with expected or anticipated performances.

Table 4.4. Functional purpose of the students

Agent:	Students
Function	Demonstrate learning
Description	Perform tasks dictated by assessments at the achieved level of LO proficiency
Input	<ul style="list-style-type: none"> • Assigned tasks
Output	<ul style="list-style-type: none"> • Task performance
Precondition	<ul style="list-style-type: none"> • Delivered course content • Guided practice
Resource/E.C.	<ul style="list-style-type: none"> • LO performance expectations
Control	<ul style="list-style-type: none"> • LOs
Time	<ul style="list-style-type: none"> • Course schedule

4.1.4 Graders' functional purpose

Table 4.5 shows the graders' functional purpose: to provide a quantitative proficiency score to represent how well a student's task performance indicates achievement of the learning objective. Part of this process includes providing feedback based on the students' performances, which is itself an important consideration, but outside the scope of this research and is not being modeled, here. In order to evaluate student performance, the graders take the student's performance on the task as input and use the rubrics provided by IST as a control to guide the grading decisions. The graders are expected to complete training provided by the IST for each learning objective as a precondition for performing the evaluation and they are limited in time to grade by the deadlines imposed by IST's course schedule.

This function is subject to significant variability as has already been demonstrated through previous analysis of grading results (Hicks & Diefes-Dux, 2017). From upstream functions, the quality of the rubric and training could affect how well the grader understands performance expectations. Interacting with the rubric and the training, the extent to which the task performance aligns with the anticipated responses and the robustness of the rubric to handle unexpected

responses may affect how easily the grader can perform the evaluation function. Additionally, mid-level work has been shown consistently to be more difficult to grade accurately (Cooksey et al., 2007). Internally, this function may vary based on how well the grader understands the expectations for student performance, which could be related to how well the grader embraced or learned from the training, how well they, themselves, understand the content being evaluated, or the performance expectations communicated by their instructor and GTA (Charney, 1984; Cooksey et al., 2007). Further, there can be a considerable impact related to graders' values, beliefs, perspectives, personal tendencies toward leniency, or from external variables that affect the grader's current physiological state (Crisp, 2010; Griswold, 2010; Meier et al., 2006).

4.2 Generalized Function Level

The second level of abstraction is the generalized function level, which consists of the set of functions that comprise set of structural behaviors required to achieve the functional purposes identified in the previous layer of abstraction (Patriarca et al., 2017). For instance, IST's functional purpose level task of "designing the course materials" consists of separate tasks associated with the design of each artifact used in the grading process, as well as the development of any other items that affect functions performed by other agents in the system. Examples of items developed by the IST that affect grading functions include the schedules that impose deadlines on assignment submission and grading completion. The output of the function of developing a grading schedule is the grading schedule itself, which serves as a time aspect for several other functions in the system.

Table 4.5. Functional purpose of the graders

Agent:	Graders
Function	Evaluate student learning
Description	Assign a quantitative score to student based on evaluation of task performance on each LO
Input	• Task performance
Output	• Performance evaluation
Precondition	• Training
Resource/E.C.	---
Control	• Rubrics
Time	• Course schedule

As noted previously, the data for this phase only provided direct evidence to construct the grader functions. As such, the construction of the functions at the generalized function level helped to organize the functions identified through the analysis of the interviews and establish the necessary background functions to define the grading functions that were more directly observed through the interviews. That is, articulation of functions at the higher level helped to track the aspects of the functions and to ensure that any inputs, preconditions, controls, time, or resources needed for a function were provided elsewhere by another function. In these cases, direct observation of the action did not occur, but could be inferred through experience with and knowledge of the system (e.g., for a course schedule to exist, someone on the IST must have performed the function of creating the schedule).

The visual representation of the collection of functions at the generalized function level of abstraction are shown in an idealized system in Figure 4.2 and the set of functions are summarized in Table 4.6 (elaborated upon in subsequent subsections). The larger number of functions and interactions at the generalized function level make it significantly more complex than the functional purpose level. The visual model illustrates how functions interact; however, number of interactions does not necessarily indicate how strongly a function affects the system's outcome.

Table 4.6. Overview of generalized functions

Agent	Function Name	Description
IST	Create content	Develop learning objectives taught in the course
	Set course schedule	Set schedule of course topics and deadlines for students and graders
	Design lesson plan	Create lesson plans for lecture, including slides and activities
	Design assessment tasks	Develop assignments to elicit performance of LO achievement
	Design grading guidelines	Develop rubrics to guide the grading of LO performances
	Design grader training	Develop materials to train graders to grade each LO
Teaching Team	Deliver course content	Communicate content to the students through lecture
	Guide student activity and practice	Guide student learning through discussion/practice
Students	Learn content	Learn the knowledge or skills associated with course LOs
	Perform assigned tasks to demonstrate learning	Demonstrate LOs through the tasks assigned in assignments
Graders	Train to calibrate grading	Participate in training to calibrate decisions made using rubrics
	Prepare to evaluate task performance	Prepare to evaluate student work by bringing necessary information to working memory (WM)
	Evaluate task performance	Apply rubrics to evaluate student performance of LOs on assigned tasks
	Record grade	Document overall grade

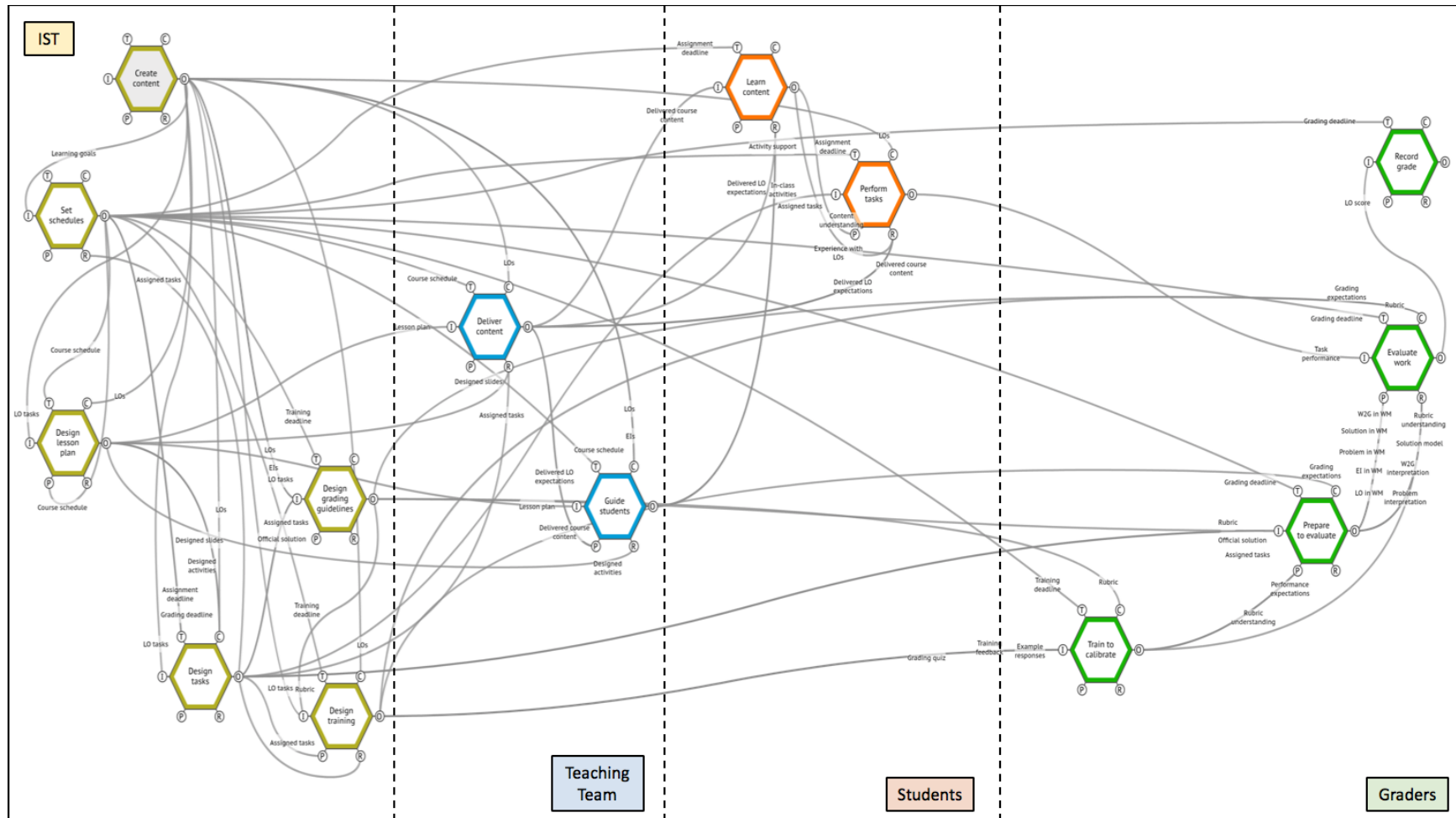


Figure 4.2. Visual representation of an idealized instantiation of the grading system at the generalized function level.

4.2.1 IST's generalized functions

The IST's functional purpose of designing course materials includes separate generalized functions associated with the design of each different component. Table 4.7 shows the FRAM frame for the IST generalized functions associated with developing the course content. Some of these functions occur more frequently than others or at different times throughout the semester, whereas the functions performed by the other agents all occur following a more regular cycle.

The second column in Table 4.7 shows one of the primary functions of the IST: creating the course's content and learning objectives. Clearly articulating what the students are expected to learn as a result of taking a course is the foundational activity in the development of a course, following Wiggins and McTighe's (2005) framework of backward design. Given the magnitude of the course and its requirement as a foundational course in the engineering curriculum, many stakeholders may contribute to the needed learning objectives; however, for the scope of this model, it is represented as receiving no inputs and serves as one of the model's most important background functions, despite the fact that in most semesters, the previous term's learning objectives may be only slightly modified or not changed at all. In addition to articulating the LOs themselves, this process should also include operationalizing the LOs into a set of evidence items (EIs) that can be observed as representing achievement of the LO. Further, the ideal practice, as articulated by the psychological and educational measurement and evaluation community (Thorndike & Thorndike-Christ, 2010, pg. 69), suggests that a set of possible tasks that would elicit each LO should be generated alongside the LO's development. Note that the outputs of this function directly affect eight other functions in the system. Potential variabilities will be elaborated upon in the cognitive function level of the model.

Table 4.7. The IST generalized functions associated with content development

Function	Create content	Set course schedule	Design lesson plan
Description	Develop learning objectives taught in the course	Set schedule of course topics and deadlines for students and graders	Create lesson plans for lecture, including slides and activities
Input	---	<ul style="list-style-type: none"> • Learning goals 	<ul style="list-style-type: none"> • LO-based tasks
Output	<ul style="list-style-type: none"> • Learning goals • Learning objectives • Evidence items • LO-based tasks 	<ul style="list-style-type: none"> • Assignment deadlines • Training deadlines • Grading deadlines • Course schedule 	<ul style="list-style-type: none"> • Lesson plan • Designed slides • Designed activities
Precondition	---	---	<ul style="list-style-type: none"> • Course schedule
Resource/E.C.	---	<ul style="list-style-type: none"> • Assigned task 	---
Control	---	---	<ul style="list-style-type: none"> • Learning objectives
Time	---	---	<ul style="list-style-type: none"> • Course schedule

The third column in Table 4.7 shows the setting course schedule function. Through this function, the IST distributes the timing of the course's content across the semester to create the course schedule. Like the creation of LOs, the course schedule will be influenced by a number of external factors (e.g., the university's calendar); however, for the scope of this model, those factors are not included making the course schedule function a background function in the model. The course schedule helps to dictate the timing and deadlines for the students to take or complete assessments throughout the course. With the assessment submission dates set, the IST can develop a training and grading schedule that will be used by the graders. This function also directly affects eight other functions and could have a considerable effect on the system, if the schedules set are not realistic or are altered by major external events. Fortunately, while the schedule is set once, prior to the start of the semester, there is enough flexibility to revise each of these outputs, if necessary.

Using the course schedule and the identified LOs, the IST produces general lesson plans for each class session (see the last column in Table 4.7). Following Wiggins and McTighe's (2005) framework, the IST should use the possible assessment tasks to create a set of lecture slides and in-class activities, such as small-group discussions or problem solving, to teach the content. Collectively, the lesson plans, slides, and in-class activities are available to each teaching team. The instructors in each section have the freedom to alter the slides and activities, so the degree to which these lesson plans are adhered to is relatively variable, but it is an expectation that the any content covered in the slides provided by the IST should be covered, in some form, by the

instructors, even if additional information is included or the content is presented in an alternative format.

Beyond creating the content and designing its delivery, the IST is also responsible for creating the assessments and establishing grading expectations, the functions for which are summarized in Table 4.8. The IST designs assessments including weekly assignments and exams (see the second column in Table 4.8). Ideally, the IST should have developed a set of multiple general tasks that should elicit performance of the course's LOs during the "create LOs" function which can then be transformed into a specific set of assigned tasks. While backward design (Wiggins & McTighe, 2005) suggests assessments should be designed before instructional activities, there may be an iterative element, such that no assigned tasks exceed the expectations communicated through instruction. In other words, there should be alignment between what is taught in class and what tasks the students are asked to perform, so the LOs, slides, and activities may serve as controls in the design of the assigned tasks. Additionally, the IST should also generate a solution or model response to the assigned task as the task is developed.

Table 4.8. The IST generalized functions associated with assignments and grading

Function	Design assessment tasks	Design grading guidelines	Design grader training
Description	Develop assignments to elicit performance of LO achievement	Develop rubrics to guide the grading of LO performances	Develop materials to train graders to grade each LO
Input	<ul style="list-style-type: none"> • LO-based tasks 	<ul style="list-style-type: none"> • Learning objectives • Evidence items • LO-based tasks • Assigned task • Official task solution 	<ul style="list-style-type: none"> • Rubric • LO-based tasks
Output	<ul style="list-style-type: none"> • Assigned tasks • Official task solution 	<ul style="list-style-type: none"> • Rubric 	<ul style="list-style-type: none"> • Grading quiz • Grading expectations
Precondition	---	---	<ul style="list-style-type: none"> • Assigned tasks
Resource/E.C.	---	---	<ul style="list-style-type: none"> • Official task solution
Control	<ul style="list-style-type: none"> • Learning objectives • Designed slides • Designed activities 	---	<ul style="list-style-type: none"> • Learning objectives
Time	<ul style="list-style-type: none"> • Assignment deadlines • Grading deadlines 	<ul style="list-style-type: none"> • Training deadlines 	<ul style="list-style-type: none"> • Training deadlines

In addition to creating the assignments themselves, the IST creates guidelines for evaluating performance of learning objectives based on students' responses to the assigned tasks (see the third column in Table 4.8). The grading guidelines ultimately take the form of a rubric and

require LO that is being evaluated and the operational components of that LO (i.e., the EIs). Generally, learning objective-based rubrics should be de-contextualized from the specific assessment to which they are being applied (Popham, 1997); however, to facilitate grading, the IST also includes additional, context-specific text within the rubrics for each unique problem for which the LO is used, so the specific assigned task and model responses are used in the creation of the rubric (see the example rubric provided in §3.2.3). This function has important potential variabilities that will be discussed in detail with the relevant sub-functions in the cognitive level model.

Finally, the IST also designs the training that will calibrate the graders' decisions while applying the rubrics to evaluate students' achievements of learning objectives (see the final column in Table 4.8). To achieve this function, the IST uses the rubric and the model response to the assigned tasks to find or generate exemplar cases to use in the training. With the exemplar cases, the IST produces an opportunity for the graders to practice applying the rubric and to get feedback on the accuracy of their grading decisions with respect to the exemplar cases.

4.2.2 Teaching team's generalized functions

At the generalized function level, the teaching team is responsible for communicating the course content to the students and fostering the achievement of the course's learning objectives through guided activities and practice. Table 4.9 shows the two generalized functions necessary to achieve the overall functional purpose: delivering the course content and guiding student activities. Both of these functions are highly dependent on the functions performed by IST.

Table 4.9. The teaching team's generalized functions

Function	Deliver course content to students	Guide student activities and practice
Description	Communicate content to the students through lecture	Guide student learning through discussion/practice
Input	<ul style="list-style-type: none"> • Lesson plan 	<ul style="list-style-type: none"> • Lesson plan
Output	<ul style="list-style-type: none"> • Delivered course content • Delivered LO expectations 	<ul style="list-style-type: none"> • In-class activities • Activity support • Delivered course content • Delivered LO expectations
Precondition	---	
Resource	<ul style="list-style-type: none"> • Designed slides • Assigned tasks 	<ul style="list-style-type: none"> • Designed activities
Control	<ul style="list-style-type: none"> • Learning objectives 	<ul style="list-style-type: none"> • Learning objectives • Evidence items
Time	<ul style="list-style-type: none"> • Course schedule 	<ul style="list-style-type: none"> • Course schedule

The first function, delivering course content to the students, converts the lesson plan provided by the IST into the course content and performance expectations that are delivered to the students (see the second column in Table 4.9). The teaching team is given a set of lecture slides, which the instructor has the freedom to modify to deliver the content, which should be aimed at helping the students to perform the tasks they are assigned in their problem sets. Ultimately, however, the teaching team is expected to facilitate the students in achieving the course's specific learning objectives within the schedule provided by IST. The extent to which the teaching team deviates from the intended lesson plan is the major source of internal variation for this function that could result in variability of the course content and performance expectations that are delivered and communicated to the students. This could lead to direct downstream effects on the students' functions.

The second function, guiding student activities and practice, uses the lesson plan provided by the IST to provide activities that help students practice the course content (see the final column in Table 4.9). Like the delivery of course content, this function possesses internal variability in that the instructor has the freedom to modify various parameters about the activity (e.g., the amount of time allowed, how the activities are presented or supported, how the students are debriefed afterward), which can affect the quality of the guided practice the students receive. The students are expected to have been presented whatever information is necessary to be able to engage with the activities prior to the activities, themselves, within the timeframe allotted by the course schedule. This function has the additional internal variability in how the multiple members of the

teaching team (including GTA and PTs) support and communicate LO performance expectations through feedback while interacting with students during these activities. As the same rubric will be used to evaluate all students across all sections, the evidence items that operationalize the learning objectives should guide the feedback given to the students throughout the activities. Further, not only may the teaching team’s provision of support differ from one section of the course to another, but there is potential variability of interpretations of the evidence items or abilities to support students within members of the teaching team; as a result, students could receive variable or contradictory feedback during the guided activities.

4.2.3 Students’ generalized functions

At the generalized function level, the students are responsible for two functions in order to achieve their general purpose: learning the content and performing the assigned tasks (summarized in Table 4.10). It is assumed, however appropriately, that students first learn the content before attempting to perform the tasks assigned in the assignments. Overall, these functions are most influenced by the teaching team’s functions and produce the most important component within the entire system—the performance that is evaluated by the grader.

Table 4.10. Generalized student functions

Function	Learn content	Perform assigned tasks to demonstrate learning
Description	Learn the knowledge or skills associated with course LOs	Demonstrate LOs through the tasks assigned in assignments
Input	<ul style="list-style-type: none"> • Delivered course content 	<ul style="list-style-type: none"> • Assigned tasks
Output	<ul style="list-style-type: none"> • Content understanding • Experience with LOs 	<ul style="list-style-type: none"> • Task performance
Precondition	---	<ul style="list-style-type: none"> • Content understanding
Resource	<ul style="list-style-type: none"> • In-class activities • Activity support • Delivered LO expectations 	<ul style="list-style-type: none"> • Delivered course content • In-class activities • Activity support • Delivered LO expectations
Control	---	<ul style="list-style-type: none"> • Learning objectives
Time	<ul style="list-style-type: none"> • Assignment deadlines 	<ul style="list-style-type: none"> • Assignment deadlines

The second column in Table 4.10 shows the first generalized student function: learning the content. In this function, the students take the content that is delivered by the teaching team and convert it to their own personal understanding of the content. Throughout this conversion of

content delivered to understanding of content, the guided practice and feedback they receive in class, along with the expectations for performance communicated to them, is to learn the knowledge or skills associated with the course learning objectives. This function involves taking as input the LOs that are communicated by the teaching team, using the resources of the lecture slides and in-class activities, as presented by the teaching team, to develop the output of the student's understanding of the content. Just as the individual members of the teaching team may vary in their interpretations of the LOs, which can vary the communicated LO input, individual students are subject to their own interpretations of the content based on any number of internal and external variabilities (e.g., intrinsic motivation to learn, past experiences, time available, competing obligations or emotions, external distractions). This means that the output of student understanding is subject to considerable variation both based on the learning function itself, but also due to upstream-downstream coupling.

Once the students have learned the content, they are expected to demonstrate their achievement of the learning objectives through performance on the assigned tasks (see Table 4.10). The student's performance on the task is a product of their understanding of the content. As with learning the content, the student's performance is time limited by the course schedule. The constraint imposed by the assignment deadline, in conjunction with internal variables, such as motivation, and external variables, such as competing demands on the student's time, means that the student's performance on the task does not necessarily perfectly represent their actual learning of the content. Similarly, students could misunderstand either verbal or written directions or expectations of the assigned task, resulting in additional variability.

4.2.4 Graders' generalized functions

The graders have three tasks at the generalized function level: training to calibrate grading, evaluating task performance, and recording grades (summarized, collectively, in Table 4.11). At the time the data was collected for this study, the training was designed to occur for each individual learning objective the week before the first assignment to use that learning objective was presented in class. The evaluation function assumes the completion of training and is the heart of this study, as all of the interview observations and the greatest amount of variability occur within this one generalized level function. Recording grades, at this level, serves as an endpoint to the system;

however, in practice, this is done for every learning objective for each assignment each student submits.

Table 4.11. Grader generalized functions

Function	Train to calibrate grading	Prepare to evaluate task performance	Evaluate task performance	Record grade
Description	Participate in training to calibrate decisions made using rubrics	Prepare to evaluate student work by bringing necessary information to working memory (WM)	Apply rubrics to evaluate student performance of LOs on assigned tasks	Document overall grade
Input	<ul style="list-style-type: none"> • Grading quiz 	<ul style="list-style-type: none"> • Assigned task • Official task solution • Rubric • ‘What to grade’ interpretation • Solution model 	<ul style="list-style-type: none"> • Task performance 	<ul style="list-style-type: none"> • LO score
Output	<ul style="list-style-type: none"> • Rubric understanding • Performance expectations 	<ul style="list-style-type: none"> • Problem interpretation • Problem in WM • Solution in WM • ‘What to grade’ in WM • LO in WM • EI in WM 	<ul style="list-style-type: none"> • LO score 	---
Precondition	---	<ul style="list-style-type: none"> • Rubric understanding • Performance expectations 	<ul style="list-style-type: none"> • Problem in WM • What to grade’ in WM • Solution in WM • LO in WM • EI in WM • Rubric understanding • Solution model 	---
Resource/ E.C.	---	---	<ul style="list-style-type: none"> • Problem interpretation • ‘What to grade’ interpretation • Rubric 	---
Control	<ul style="list-style-type: none"> • Rubric 	<ul style="list-style-type: none"> • Grading expectations 	<ul style="list-style-type: none"> • Grading expectations 	---
Time	<ul style="list-style-type: none"> • Training deadlines 	<ul style="list-style-type: none"> • Grading deadlines 	<ul style="list-style-type: none"> • Grading deadlines 	<ul style="list-style-type: none"> • Grading deadlines

The first function the graders are responsible for performing, as shown in Table 4.11, is using the training materials provided by the IST to calibrate their understanding of how to apply the rubrics. Using the training materials, the graders obtain an exposure to examples and an understanding of the rubric. The rubric operates as a resource to guide the grader’s grading

decisions. The training should also be completed by training deadlines, which are set before the students begin working on the assigned tasks. The extent to which the graders engage seriously with the training materials likely varies, even within a single grader from one training to the next, due to differences in personality, emotional state, or competing obligations and time constraints. Further, the design of the training in the context of this study was that the students would take two “quizzes” and would be prepared to grade afterward, regardless of performance on the training. As such, there is no guarantee of consistent interpretation or learning outcomes from the training.

The third column of Table 4.11 shows the core overarching function of the grading system—evaluating student achievement of learning objectives based on performance on assigned tasks. This function takes in the student’s performance on the task and produces a score for each learning objective. As training is required before grading, exposure to examples is an expected precondition to the evaluation function and an understanding of the rubric developed through training should be used as a resource in making grading decisions; however, as noted previously, variability in understanding is possible. Ultimately, the graders are expected to follow the rubric to the best of their ability to make their decisions and complete their grading prior to the grading deadline set by the IST. Thus, there are many internal and external sources of variability for this function, which will be elaborated upon in subsequent sections.

The final function within this entire model of the grading process, disregarding the regrade requests from students and scoring adjustments made by the graduate teaching assistant or instructor, is recording the grade (the final column of Table 4.11). This function simply documents the LO score that was decided through the evaluation of the performance. Any other action is considered external to the process, as the score is the final output of the system. While it would seem that this function should have no variability, occasionally students do not receive grades for random learning objectives they demonstrated. This may be a fault of the learning management system in which the grades are recorded, an intended click on the part of the grader but a failure to actually do so, or the entire LO being overlooked. It is also theoretically possible for a grader to accidentally click the wrong score.

4.3 Cognitive Function Level

In Patriarca et al.’s (2017) work, the deepest layer of abstraction is referred to as the “Physical Function” layer. They define these as the “specific processes related to sets of interacting

components and their properties.” In the context of railways, they suggest that one physical function of the rail driver is to “shut off power.” While this is a physical action, other functions at the same level are the signaler’s functions to “identify late train,” and “detect a track circuit block,” which are clearly cognitive tasks. As the context of this project is almost entirely consisting of cognitive tasks, the final layer of abstraction for this model will be referred to as cognitive functions rather than physical functions.

Functions at the cognitive layer include the specific tasks required to achieve the generalized functions, primarily in terms of the cognitive processing or decisions made by the members of the system. While the only agents directly observed through the think-aloud interviews were the graders, identification of cognitive functions performed by other members of the system was necessary to create a full and complete model at that level of abstraction. The inputs needed for the grader functions in the model required cognitive level functions of each agent in the system. Thus, all other processes had to be inferred through logical necessity and experience with the system. For example, though there was no direct observation of a student performing the assigned task, it clearly had to occur for the student work to exist.

While Hollnagel (2012) argues that FRAM models and functions should be developed openly, without pre-defined functions, the lack of direct observational evidence of the processes conducted by agents beyond the graders required the functions to be grounded, at least partially, in theory. Fortunately, the works of Black et al. (2011) and Suto and Nadas (2009) identified a number of features of questions and responses that contribute to grading difficulty. These factors provided a base set of possible codes; however, features of the rubrics, as well as additional features of the responses and assignment questions were revealed by the way the graders interacted with the documents during the think-aloud interviews. This back-and-forth process of considering documents and interviews demonstrated the need to create and interpret variability of functions theoretically grounded in the effects of cognitive load, cognitive demand, and emotion on cognitive ability and decision making.

Figure 4.3 shows a visual representation of the consistently used functions at the cognitive level of abstraction. At this level of abstraction, the number of functions makes it very difficult to read anything when the entire visual representation is presented. Thus, the following sub-sections will show portions of the overall visual representation of the system to provide a clearer image of the functions that are being discussed in that sub-section. This will also help to illustrate how the

potential variability of the different functions can affect the system, particularly in-terms of direct upstream-downstream interactions. It is also important to note that one of the grading functions—evaluate student performance—has extremely large internal variability dependent upon the outputs of the functions that precede it in the system. The evaluation function consists of many sub-functions that will be discussed in this section but will not be visualized until the next chapter when specific contexts can help illustrate the variability across work-as-imagined and work-as-completed instantiations.

Also note that throughout this section, there will be discussions and tables presented about the potential variability of the outputs of each function. In the tables, these are presented as either dichotomous or trichotomous variables (e.g., a task being either closed-ended or open-ended or directions being clear or unclear). It is acknowledged that in most cases, these variables actually lie on a more continuous spectrum. The polar nature expressed in these tables is intended to simplify the presentation of the model and demonstrate that as the variable moves in one direction or the other along the spectrum, it may amplify or dampen variation in downstream functions. It is not intended to suggest that the variables are truly dichotomous.

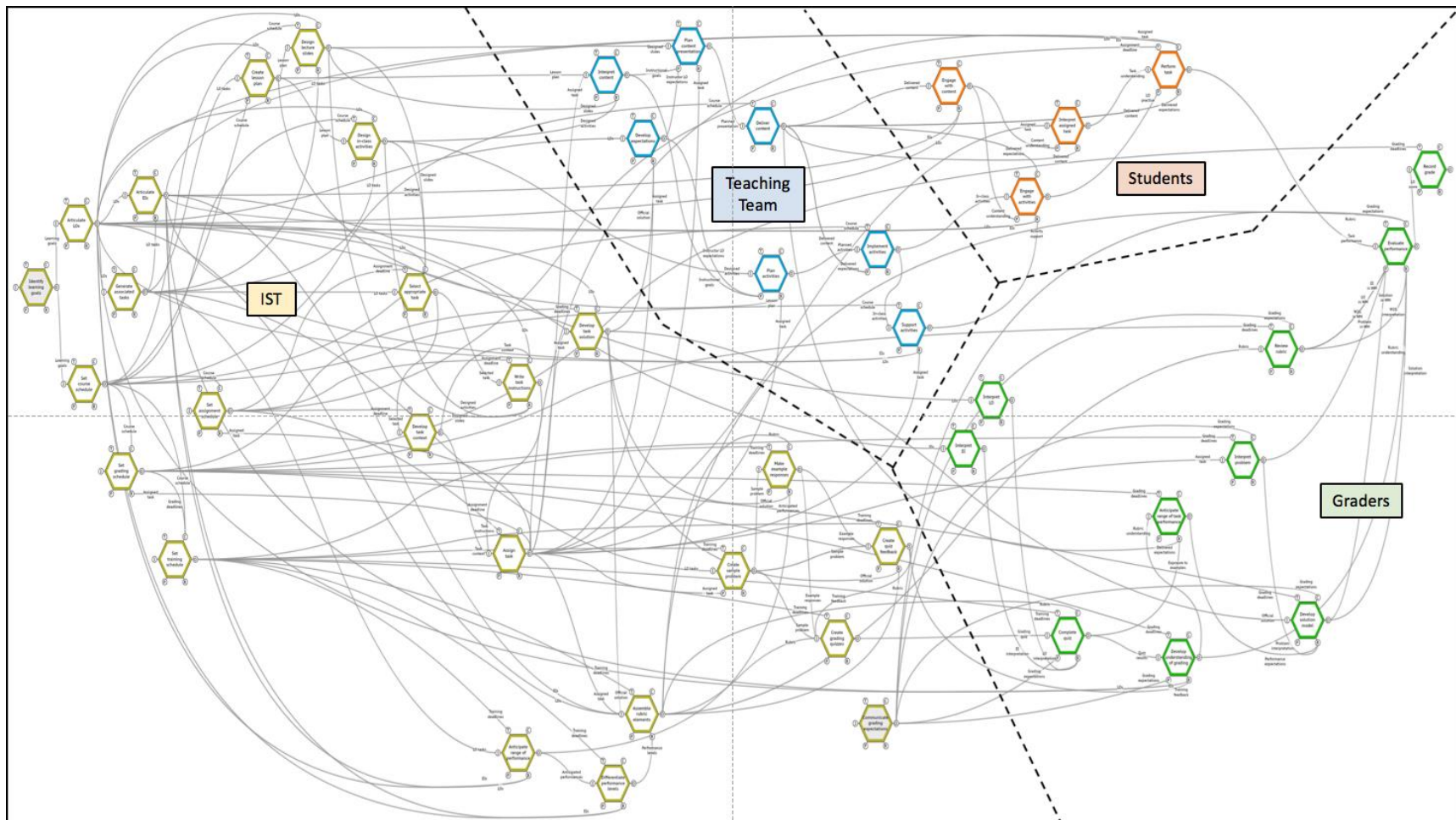


Figure 4.3. Visual representation of the cognitive level functions

4.3.1 IST's cognitive functions

There are 24 IST functions at the cognitive level of abstraction, many of which can have significant impacts on the system. Across the IST functions that will be discussed, some occur several times throughout a semester while others may only occur every few semesters. Still, the effects of these functions are large, and an undesirable output can propagate throughout the entire system across all sections of the course. This section will explore each of these functions and how they might affect the overall system. Also note, these functions, and the ways the outputs can vary were created with the intention of minimizing overlap. For example, if it is stated that a functions output can vary along a few different dimensions, the intention is that these are distinct, unidimensional constructs that do not correlate with one another.

Cognitive functions associated with developing course content

The “Create content” generalized function comprises the four cognitive level functions summarized in Table 4.12. The most fundamental step, following the Wiggins and McTighe (2005) backward design model, is the first function: identifying the general course learning goals. In the broader context beyond the course itself, these learning goals can vary by being more or less appropriate for the students who will take the course (see Table 4.13). An example of an inappropriate learning goal would be one that assumes students have prerequisite skills or knowledge that at least some subpopulations of the students do not possess. At a large university like the one where this study was conducted, this function likely coincides with course creation and requires a change in course number to perform any significant changes to the course learning goals.

Table 4.12. Cognitive level IST functions involved in developing course content

Function	Identify course learning goals	Articulate LOs	Articulate EIs	Generate associated tasks
Description	Identify the desired learning goals for the students after the course is completed	Articulate a set of learning objectives based on the learning goals of the course	Operationalize the achievement of the learning objective into a set of observable features of a performance	Generate a range of possible performance tasks that would elicit proficiency of LO
Input	---	• Learning goals	• Learning objectives	• Learning objectives
Output	• Learning goals	• Learning objectives	• Evidence items	• LO-based tasks
Precondition	---	---	---	---
Resource/ E.C.	---	---	• LO-based tasks	---
Control	---	---	---	---
Time	---	---	---	---

Table 4.13. Potential variability of course content development functions

IST	Generalized function: Create content		
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Identify course learning goals	Appropriateness	<u>Unclear</u> : LO does not clearly communicate intended construct, making interpretation more difficult [V↑]	Possible, likely
		<u>Adequate</u> : LO adequately communicates intended construct [V↔]	Typical
		<u>Clear</u> : LO is very clearly communicated, reducing likelihood of misinterpretation [V↓]	Possible, unlikely
Articulate LOs	Clarity	<u>Broad</u> : LO spans a large set of behaviors and requires many evidence items; graders will be less likely to pay sufficient attention to all evidence items during grading [V↑]	Possible, unlikely
		<u>Narrow</u> : LO spans a small set of behaviors and requires few evidence items; graders are more likely to properly evaluate the small number of evidence items [V↓]	Typical
		<u>Misaligned</u> : EIs are not aligned well with the LO or task, making student performance of evidence item and grading behavior more variable [V↑]	Possible
	Breadth	<u>Aligned</u> : EIs align with the LO and task [V↔] <u>Unclear</u> : EIs are unclear or difficult to understand, leading to variable interpretation for all agents [V↑]	Possible Possible
Articulate EIs	Alignment	<u>Adequate</u> : EIs are generally able to be understood and interpretations are consistent [V↔]	Possible, unlikely
		<u>Clear</u> : EIs are very clear, unlikely to be misunderstood [V↓]	Typical
	Clarity	<u>Imprecise</u> : EIs are not adequately specific or overlap with one another, leading to variable interpretations and grading behavior [V↑]	Possible, unlikely
		<u>Adequate</u> : EIs adequately specify construct and do not overlap with one another [V↔]	Typical
		<u>Insufficient coverage</u> : The LO's construct is not fully represented by EIs, leading to potential emotional reactions from graders due to misrepresentation of score [V↑]	Possible
	Precision	<u>Sufficient coverage</u> : EIs fully represent LO construct [V↔]	Possible, likely
		<u>System 1</u> : Requires simple cognitive processing to evaluate [V↓]	Typical
	Coverage	<u>System 2</u> : Requires complex cognitive processing to evaluate [V↑] <u>Unaligned</u> : Tasks do not sufficiently sample or align with learning objective, making grading decisions difficult [V↑]	Possible, unlikely Possible, likely
Generate associated tasks	Alignment	<u>Aligned</u> : Tasks align with learning objectives [V↔]	Likely
		<u>Unclear</u> : LO does not clearly communicate intended construct, making interpretation more difficult [V↑]	Likely
		<u>Adequate</u> : LO adequately communicates intended construct [V↔] <u>Clear</u> : LO is very clearly communicated, reducing likelihood of misinterpretation [V↓]	Possible, unlikely Typical

The second and third functions, shown as the third and fourth columns in Table 4.12, are used to translate the general idea of the course's learning goals into intentionally articulated, measurable learning outcomes. The 'Articulate LOs' function translates the learning goal into a more specific learning objective or outcome that the students will be expected to achieve after completing the course. For example, a general learning goal might be for students to learn non-sequential programming algorithms, which could translate into a few more specific learning objectives. One such learning objective might be that students will be able to write the code for a selection structure in MATLAB. The two major types of variability of the LO articulation function is how clearly the learning objective is articulated and the breadth of observable behaviors are spanned by the learning objective (see Table 4.13 for summary). The learning objective is used directly by 16 other functions (see Figure 4.4), so an unclear learning objective can lead to large variability throughout the system. Further, an overly broad learning objective may impose excessive cognitive load on graders and lead to greater variability in grading decisions.



Figure 4.4. Visualization of content development functions.

The ‘Articulate EIs’ function breaks the learning objective into a set of properties that can be directly observed in a student’s work and provide evidence that the student has achieved some degree of the learning objective. From this perspective, if every evidence item is observed in the student’s work, it can be assumed that the student has fully achieved the desired learning outcome. To continue the previous example, one piece of evidence that a student has learned how to code a selection structure in MATLAB is if they start a selection structure with the “if” command and use “elseif” or “else” for alternative paths.

There are several ways that the evidence items produced can vary, which given the number of direct and indirect interactions with other functions, can have a large effect on the system (see Table 4.13 and Figure 4.4). First, the way the learning objective is operationalized into evidence items may vary in their alignment with the tasks that are expected to elicit the behaviors, such that poor alignment may make it more difficult to elicit or identify the evidence items. Second, like the learning objectives themselves, the evidence items may vary in terms of their clarity, whereby lower clarity can lead to difficulty achieving consistent interpretation, which must be done by the teaching team, the students, and the graders. Variation of interpretation across and within these groups can lead to significant aggregation of variability of outputs throughout the system. Further, the evidence items can vary in precision (in this context, how specifically they address a construct without overlapping one another) and coverage (in this context, how sufficiently they cover and represent the learning objective they represent). Variation across these dimensions can conflict with student performances and lead to variable interpretations and grading decisions. Finally, the evidence items can vary in the inherent complexity of the evaluative task needed to identify them (i.e., do they require system 1 or system 2 cognitive processing to evaluate?). As the literature shows, more complex evaluative tasks are less consistent in outcome (Suto & Greatorex, 2008).

Along with articulating the learning objectives and evidence items, one should also ‘Generate associated tasks.’ Wiggins and McTighe’s (2005) framework for course design recommends identifying how students will demonstrate achievement of learning objectives immediately after identifying the learning objectives in the first place. Thorndike and Thorndike-Christ (2010) also advocate for exploring the entire breadth of possible tasks for each learning objective, which could help to illuminate the level of performance of the learning objective that will be appropriate for the students in the course, to the extent that it can realistically be done in practice. For example, recognizing that students will only need to create a selection structure

within the language of MATLAB helps to establish boundaries for the course. Similarly, one might consider a task with a selection structure that is far more complex than would be necessary for beginning programmers to accomplish, providing boundaries to the articulation of the learning objective and evidence items. The major opportunity for variability of this function is the alignment between the tasks generated and the learning objectives, which could create conflicts when the teaching team, students, or graders deal with learning objectives and the assigned tasks that come from the potential tasks.

Cognitive functions involved in setting schedules and deadlines

Table 4.14 shows the four different functions that are involved in setting schedules to manage course logistics. There is a degree of appropriate sequencing to these functions, given that some deadlines need to be set based on others and are generally completed prior to the beginning of each semester (see Figure 4.5 for interactions). The ‘Set course schedule’ function should be completed first, as the sequencing and time devoted to each topic and learning goal within the course should dictate when each assignment will be given. The ‘Set assignment schedule’ should then be used to ensure that students have been exposed to the relevant content and have sufficient time to properly complete the assigned tasks. Next, the ‘Set grading schedule’ function should be used to create a schedule based on when students will submit the assignments. Like the assignment schedule, this should be completed with knowledge of the assignments to properly estimate the time needed to produce high quality grades. Finally, the ‘Set training schedule’ function should be performed such that training will be completed prior to the start of grading for any assignment.

Table 4.14. Cognitive level IST functions involved in setting schedules

Function	Set course schedule	Set assignment schedule	Set grading schedule	Set training schedule
Description	Develop a schedule for the dissemination of the course content, considering the time needed for adequate coverage	Develop a schedule and submission deadlines for course assignments, considering tasks	Develop a schedule of deadlines for assignment grading, considering tasks	Develop a schedule for completing training, based on grading and course schedule
Input	• Learning goals	---	---	---
Output	• Course schedule	• Assignment deadlines	• Grading deadlines	• Training deadlines
Precondition	---	---	---	---
Resource/ E.C.	---	• Assigned tasks	• Assigned task	---
Control	---	• Course schedule	• Assignment deadlines • Course schedule	• Grading deadlines • Course schedule
Time	---	---	---	---



Figure 4.5. Visualization of schedule setting functions.

With each of these functions, the major source of variability is the precision of the schedule—that is, whether or not it is created to allow sufficient time for the relevant functions to be completed properly (see Table 4.15 for summaries). As the schedules and calendars function as time inputs for subsequent functions, schedules or deadlines being too tight can force tasks to be performed at lower levels of quality (i.e., having too little time to each a topic, complete an assignment, grade, or train). These can result be the result of completing the tasks out of sequence or without helpful information. For instance, if the grading schedule is set before the assignment schedule is completed, there may not be enough time allocated for grading an assignment after it is due for the students. Similarly, if the assignment or grading schedules are made without knowledge of the assigned tasks, too little time might be available for the task if it ended up being more challenging to complete than anticipated. Fortunately, there is a good degree of flexibility

with scheduling such that schedules could, in theory, be adjusted throughout a semester; although, adjustments may cause rippling effects for subsequent deadlines.

Table 4.15. Potential variability of schedule setting functions

Generalized function: Set schedules			
IST			
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood of variability
Set course schedule	Precision	<u>Imprecise</u> : Not enough time provided by schedule to properly perform subsequent functions [V↑]	Possible, unlikely
		<u>Adequate</u> : Time provided by schedule is sufficient to properly perform subsequent functions [V↔]	Typical
Set assignment schedule	Precision	<u>Imprecise</u> : Not enough time provided for students to adequately complete assignments [V↑]	Possible, unlikely
		<u>Adequate</u> : Time provided is sufficient for students to complete assignments up to ability level [V↔]	Typical
Set grading schedule	Precision	<u>Imprecise</u> : Not enough time provided for graders to adequately complete grading [V↑]	Possible, unlikely
		<u>Adequate</u> : Time provided is sufficient for graders to grade up to their ability level [V↔]	Typical
Set training schedule	Precision	<u>Imprecise</u> : Not enough time provided for graders to adequately complete training [V↑]	Possible, unlikely
		<u>Adequate</u> : Time provided is sufficient for graders to train up to their ability level [V↔]	Typical

Cognitive functions involved in developing class sessions

The IST engages in three primary tasks related to developing class sessions, as shown in Table 4.16. First, the course schedule and an understanding of tasks associated with the expected learning outcomes can be used to generate a basic plan for each class session. That plan can then be used, with the learning objectives and corresponding tasks in mind, to develop a set of general lecture slides for the class session. Using the same inputs, the IST also develops a set of recommended activities that are embedded within the lecture slides. Figure 4.6 shows a visual representation of how these functions interact with one another. Note that there are not an enormous number of output paths, as these functions most directly interact with the teaching teams, who then have the opportunity to adjust, to some extent, the lecture slides and designed activities as they see fit for their class.

Table 4.16. Cognitive level IST functions involved in developing class sessions

Function	Create lesson plan	Design lecture slides	Design in-class activities
Description	Create the overall plan for what content is covered in each class session	Design the basic lecture slides to communicate the course content	Design discussions or practice activities to guide student learning during class sessions
Input	• LO-based tasks	• Lesson plan	• Lesson plan
Output	• Lesson plan	• Designed slides	• Designed activities
Precondition	• Course schedule	---	---
Resource/ E.C.	---	• LO-based tasks	• LO-based tasks
Control	• Learning objectives • Evidence items	• Learning objectives	• Learning objectives
Time	---	• Course schedule	• Course schedule

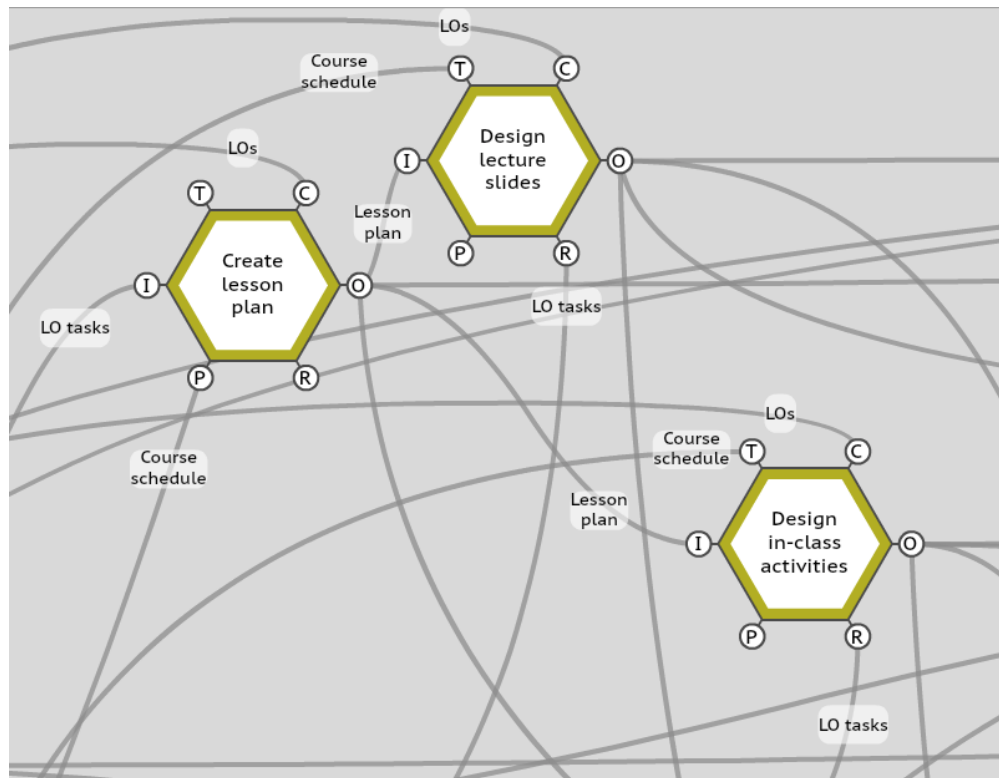


Figure 4.6. Visualization of lesson planning functions.

As summarized in Table 4.17, there are a couple ways the lesson planning functions can vary. For each function, there are two general ways the output can vary that could affect downstream functions: alignment with expected content and clarity or thoroughness. If the lesson plan is not aligned with the content that needs to be delivered, it will likely cause both the designed lectures slides and designed activities to also fail to align with the content appropriately. Poor alignment with the lecture slides and designed activities will result in more variable treatment by the different teaching teams that have to use those outputs, where less experienced instructors may be more likely to use the poorly aligned slides and activities and more experienced instructors may have a better handle on how to adapt it to better teach the content. A similar set of outcomes can occur if the initial lesson plan or subsequently designed slides or activities are unclear or insufficiently thorough to properly cover the learning objectives. As the outputs of these functions are passed to the teaching teams, unclear or insufficiently thorough outputs will increase the likelihood that instructors will individually change the presentation of content, which increases the variability throughout the system.

Table 4.17. Potential variability of lesson planning functions

IST	Generalized function: Design lesson plans		
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Create lesson plan	Alignment	<u>Misaligned</u> : Lesson plan does not appropriately address the intended learning objectives and content [V↑]	Unlikely
		<u>Aligned</u> : Lesson plan aligns with learning objectives and content [V↔]	Typical
	Clarity/Thoroughness	<u>Unclear/Not thorough</u> : Lesson plan is not communicated clearly or sufficient depth to cover content [V↑]	Unlikely
		<u>Adequate</u> : Lesson plan is sufficiently clear and thorough to appropriately cover content [V↔]	Typical
Design lecture slides	Alignment	<u>Misaligned</u> : Content in lecture slides do not cover the appropriate content, leading to greater likelihood of different presentations across sections [V↑]	Possible, unlikely
		<u>Aligned</u> : Lectures slides are well aligned with the content and likely to be presented consistently across sections [V↔]	Typical
	Clarity/Thoroughness	<u>Unclear/Not thorough</u> : Lecture slides are difficult to understand or supplements to communicate content, leading instructors to potentially make unique alterations [V↑]	Possible, unlikely
		<u>Adequate</u> : Lecture slides are clear and thorough, likely to be presented consistently across sections [V↓]	Typical
Design in-class activities	Alignment	<u>Misaligned</u> : Designed activities do not support learning of the content, leading to greater likelihood of different activities across sections [V↑]	Possible, unlikely
		<u>Aligned</u> : Activities strongly support learning of content, likely to be consistent across sections [V↓]	Typical
	Clarity	<u>Unclear</u> : Designed activities are difficult to understand, leading instructors to use modified or different activities [V↑]	Possible, unlikely
		<u>Adequate</u> : Activities are easy to understand, likely to be consistent across sections [V↓]	Typical

Cognitive functions involved in designing assignment task

There are five cognitive functions utilized by the IST as assignment tasks are developed (see visualization in Figure 4.7). After generating a set of possible performance tasks when developing the learning objectives, the first task in creating an assignment is to select the specific tasks that will be assigned (see ‘Select appropriate task’ in Table 4.18). While backward design does dictate the design of assessments before pedagogical planning, it might be helpful for the design of teaching materials (i.e., lecture slides and activities) and the selection and design of assigned tasks to occur in an iterative fashion (Wiggins & McTighe, 2005). In other words, to ensure alignment (see Table 4.19 for types of variability), the assigned task should inform how the students are taught; however, the task should also fit within what can realistically be supported through the teaching materials. In addition to alignment, the selected task may also vary in terms of its open-endedness, where more open-ended tasks lead to more variable student work.

Table 4.18. Cognitive level IST functions involved in designing assignment tasks

Function	Select appropriate task	Develop task context	Write task instructions	Assign task	Develop task solution
Description	Select an appropriate task that will elicit performance of one or more LOs	Situate the selected task within a meaningful engineering-related context	Create a set of instructions to guide students through performance of the task	Assemble the task context and instructions into a cohesive task within the overall assignment	Develop a model solution to the assigned task
Input	<ul style="list-style-type: none"> • LO-based tasks 	<ul style="list-style-type: none"> • Selected task 	<ul style="list-style-type: none"> • Selected task 	<ul style="list-style-type: none"> • Task context • Task instructions 	<ul style="list-style-type: none"> • Assigned task
Output	<ul style="list-style-type: none"> • Selected task 	<ul style="list-style-type: none"> • Task context 	<ul style="list-style-type: none"> • Task instructions 	<ul style="list-style-type: none"> • Assigned task 	<ul style="list-style-type: none"> • Official task solution
Precondition	---	---	---	---	---
Resource/ E.C.	---	---	<ul style="list-style-type: none"> • Designed slides • Designed activities 	---	---
Control	<ul style="list-style-type: none"> • Learning objectives • Designed slides • Designed activities 	---	<ul style="list-style-type: none"> • Task context • Learning objectives 	---	<ul style="list-style-type: none"> • Learning objectives
Time	---	---	---	<ul style="list-style-type: none"> • Assignment deadlines 	<ul style="list-style-type: none"> • Grading deadlines

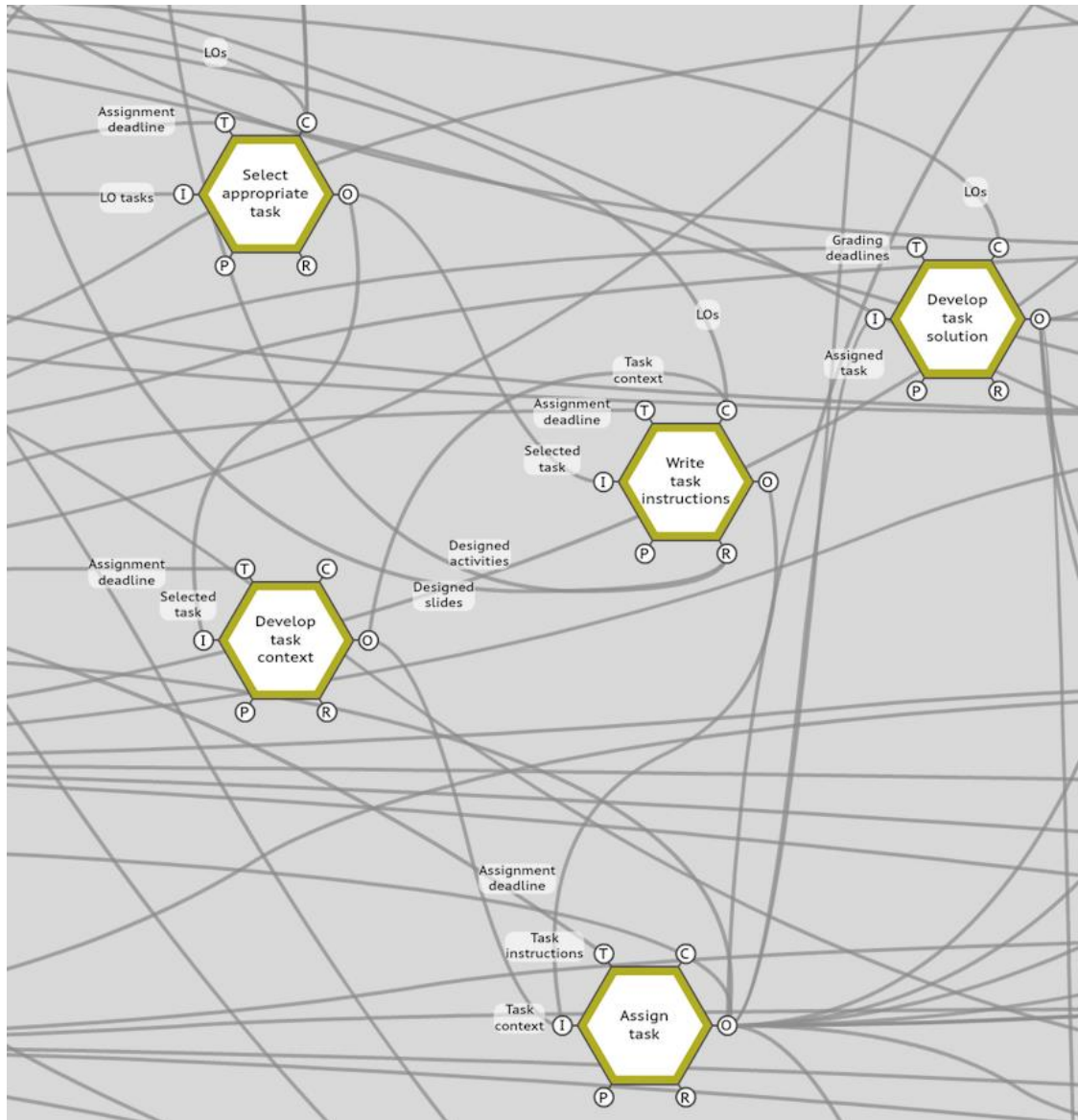


Figure 4.7. Visualization of assessment task design functions.

Table 4.19. Potential variability of assessment task design functions

IST		Generalized function: Design assessment task	
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Select appropriate task	Alignment	<u>Misaligned</u> : The selected task being more misaligned with the content will cause the assigned task to be more difficult for students, increasing variability of task performance [V↑]	Unlikely
		<u>Aligned</u> : The selected task aligns with the content, allowing the task to be completed by the students as expected [V↔]	Typical
	Open-endedness	<u>Closed-ended</u> : The task has significant constraints, limiting the variability in approaches that might be taken by students and making grading more straightforward [V↓]	Likely
		<u>Open-ended</u> : The task is less constrained, allowing more greater variation in the way students approach the task and making grading more subjective [V↑]	Likely
Develop task context	Understandability	<u>Difficult to understand</u> : The difficulty of interpreting the context contributes to additional variability of the students to understand what is expected of them within the task, increasing variability of task performance and grader understanding [V↑]	Possible, unlikely
		<u>Easy to understand</u> : The context is easy to understand, most students and graders are likely to interpret it similarly. The context does not contribute to variability of task performance or evaluation [V↔]	Possible, unlikely
Write task instructions	Clarity	<u>Unclear</u> : Instructions that are unclear will be harder for students to follow and know what is expected, leading to more variable responses [V↑]	Possible, unlikely
		<u>Clear</u> : Instructions that are very clearly written should be easily and consistently interpreted by students who will understand expectations [V↓]	Possible, likely
	Scaffolding	<u>Low</u> : Students are given very little guidance with the task and may not know what to do, likely increasing the number of approaches used [V↑]	Possible, likely
		<u>Moderate</u> : Students are given enough guidance on the task to recognize a general approach to use [V↔]	Typical
Assign task	Difficulty	<u>High</u> : Students are explicitly given every step along the way to perform the task, leaving little room for interpretation or variability [V↓]	Possible, unlikely
		<u>Difficult to perform task</u> : Students will produce more variable responses to tasks that are harder [V↑]	Possible, likely
Develop model response	Comprehensiveness	<u>Easy to perform task</u> : Students will produce more consistent, high-level performance on easier tasks [V↓]	Possible, unlikely
		<u>Narrow</u> : The solution only represents a narrow range of possibly acceptable responses, leaving more decision making in the hands of more variable graders [V↑]	Possible
	Accuracy	<u>Broad</u> : The solution fully encapsulates and represents all possible responses, improving graders' decisions [V↓]	Possible
		<u>Inaccurate</u> : The solution has errors, leading graders to make incorrect grading decisions [V↑]	Possible, unlikely
		<u>Accurate</u> : The solution is accurate, and graders make appropriate grading decisions [V↔]	Typical

As the learning objectives for this course are generally related to programming tasks, they tend to be context-free. To make the assigned tasks more authentic engineering tasks, rather than simply assigning students a context-free task identified in the previous function, the IST then engages in the ‘Develop task context’ function. As Table 4.18 shows, this function is primarily affected by the personal knowledge and experience of the person writing the task. There is an important way that the output, the written task context, can vary (see Table 4.19)—its understandability. Understandability is multidimensional in this context, representing how likely the context is to be familiar to the students, how abstract it is, how much it depends on assumed prerequisite knowledge, and how clearly it is written. These are all lumped together because they collectively affect whether or not the students will be able to easily interpret the context of the problem or whether it will contribute context-irrelevant variance to performing the task. If the context is not easily understandable, it can be expected that student responses will be more variable. Similarly, as the grader has to have an understanding of the context to properly interpret some students’ responses and the grader little more schooling than the students, less understandable contexts can lead to variability in their understanding, as well.

The selected task can also be used to ‘Write task instructions’ (see Table 4.18). Writing the instructions can be guided by the context that was developed along with the learning objectives that are expected to be demonstrated. It is important to keep the learning objective in mind while writing the task instructions to ensure that if the students will be expected to perform a task in a certain way that they the instructions effectively communicate that expectation. It is also helpful in developing the task instructions to make sure that there is alignment between what the students will see on the assignment and what they are shown in class through slides and activities. The task instructions can vary in terms of their clarity and the extent to which the instructions scaffold the task (see Table 4.19). If the instructions are unclear, it is likely that students’ interpretations of the task will vary and will approach the problem in more variable ways. Similarly, the task can be presented by breaking down every partial step along the way (i.e., with a lot of instructional scaffolding) or can be written more broadly. Naturally, the fewer explicit directions the students are given, the more likely they will be to generate a wider variety of responses to the task.

Next, the IST assembles all the pieces into a cohesive assignment (‘Assign task’ shown in Table 4.18). Collectively, the outputs of the previous three functions contribute to the overall difficulty of the task. That is, the combination of how well aligned with the content taught, the

open-endedness, the understandability of the context, the clarity of the instructions, and the scaffolding all mix to put the task on a spectrum of difficulty. More difficult tasks will tend to produce more variable answers from students and will be, in general, more challenging to grade (Suto & Nádas, 2010).

While the preceding functions mostly feed into how the teaching team will teach the content and the students' performance of the task, the last task-related function, 'Develop a model response,' most directly affects the graders. In this function, the IST must make a "solution" to show the graders that will help the graders develop a sense of what constitutes an acceptable response to the assigned task. When the task is less constrained, it is likely that there will be a wider range of possible acceptable responses. As a result, the IST's solution may represent just one of many possible solutions. The more effectively the solution can communicate the range of acceptable responses, the easier it will be for the graders to develop an appropriate mental model while grading. It has also happened, on rare occasion, that the solution produced by the IST has an error, which can cause the graders to develop an incorrect solution model and grade incorrectly. Hopefully when errors occur, they are caught, corrected, and spread across sections, but there is a small possibility that does not occur.

Cognitive functions associated with designing grading guidelines

The IST has three cognitive level functions associated with the generalized function of designing guidelines for grading, as shown in Table 4.20. These functions—'Anticipate range of performance,' 'Differentiate between performance levels,' and 'Assemble rubric elements'—should occur in a sequential manner, so that the output of one can be used to initiate or guide the next. The outputs of the first two functions, as can be seen in Figure 4.8, are almost entirely internal to the generalized function of designing grading guidelines, whereas the output of the final function, the rubric, is what is used by the students and graders.

Table 4.20. Cognitive level IST functions associated with designing grading guidelines

Function	Anticipate range of performance	Differentiate between performance levels	Assemble rubric elements
Description	Anticipate the variable levels of performance of the LO	Distinguish between proficiency levels of variable task performance	Assemble all the components of a rubric into a cohesive tool
Input	<ul style="list-style-type: none"> • LO-based tasks 	<ul style="list-style-type: none"> • Anticipated performances 	<ul style="list-style-type: none"> • Learning objectives • Evidence items • Official task solution • Assigned task
Output	<ul style="list-style-type: none"> • Anticipated performances 	<ul style="list-style-type: none"> • Performance levels 	<ul style="list-style-type: none"> • Rubric
Precondition	---	---	---
Resource/ E.C.	<ul style="list-style-type: none"> • Learning objectives • Evidence items 	<ul style="list-style-type: none"> • Learning objectives • Evidence items 	<ul style="list-style-type: none"> • Performance levels
Control	---	---	---
Time	---	---	<ul style="list-style-type: none"> • Training deadline

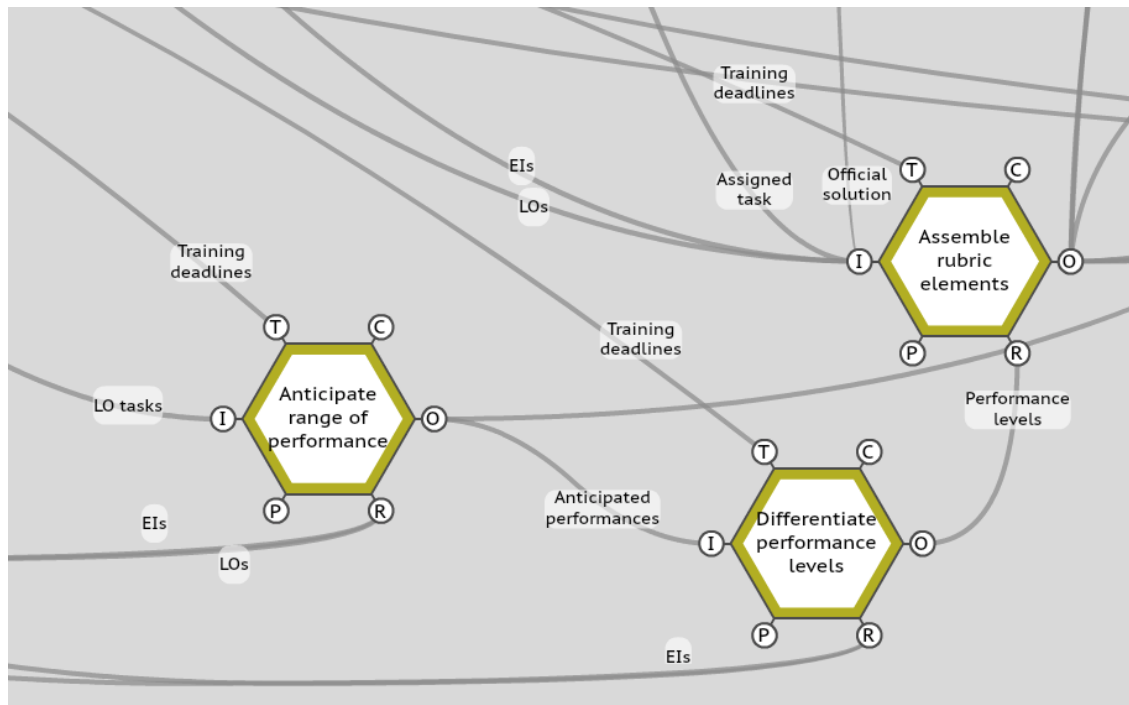


Figure 4.8. Visualization of grading guideline design functions.

The first function, ‘Anticipate range of performance,’ is an extension beyond what must be performed during the ‘Develop model response’ function that was discussed previously. The previous function’s best performance includes identifying the full range of possible solutions for a particular selected task. Following Popham’s (1997) suggestion that rubrics should be free from specific context and focus on the construct at hand, this function’s ideal output is the full set of possible performances of the learning objective in a more general sense. While it is unlikely to be able to fully anticipate every possible way the students will respond to a task, the more comprehensively this function is performed, the more robust the rubric can be with handling variable student responses (see summary of variability in Table 4.21).

Table 4.21. Potential variability of grading guideline design functions

IST	Generalized function: Design grading guidelines		
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Anticipate range of performance	Comprehensiveness	<u>Narrow</u> : The range identified fails to capture ways students will respond, causing rubric to not address how to handle some student work [V↑]	Possible, likely
		<u>Comprehensive</u> : The complete range of possible responses is identified, allowing the rubric to address how to handle all cases [V↓]	Possible, unlikely
Differentiate between performance levels	Discriminability	<u>Weak</u> : Either too many or too few performance levels are considered, or the performance levels are stuck within a rigid structure encouraging the ‘Halo effect’, causing graders to struggle to find a grade that properly reflects student performance [V↑]	Possible, unlikely
		<u>Strong</u> : An appropriate number of levels are identified for all possible performances [V↓]	Possible
Assemble rubric elements	Usability	<u>Low</u> : The rubric is difficult to use, either due to confusing or unnatural layout of content or because too much information is present, causing usage patterns to vary across graders [V↑]	Possible, unlikely
		<u>High</u> : The rubric is easy to use and likely to be used consistently by all users [V↓]	Possible, unlikely
	Specificity	<u>High/Overlapping</u> : The portions specified for grading represent a small portion of the overall answer and/or overlap with portions designated throughout, leading to the possibility of students being penalized repeatedly for individual errors and causing graders to make variable decisions [V↑]	Typical
		<u>Low/Unique</u> : The portions specified by the rubric are unique or reflect overall task performance, so grading decisions feel appropriately representative [V↔]	Possible, unlikely
	Robustness	<u>Weak</u> : The rubric fails to address student performances, leaving graders uncertain how to grade the response [V↑]	Possible, likely
		<u>Strong</u> : All possible approaches used by students are covered by the rubric, allowing graders to know what decisions to make in all cases [V↓]	Possible, likely

The next function, ‘Differentiate between performance levels,’ considers the range of possible performances of the task to properly differentiate discrete “levels” of performance. The proper number of performance levels and the ideal cutoffs between performance levels can be difficult to determine but is an important function (Goldberg, 2014; Moskal, 2003; Popham, 1997; Tierney & Simon, 2004). If there are too many performance levels, it may be hard for the grader to identify which level is most appropriate for a student’s response. On the other hand, if there are too few levels, the grader may feel conflicted that a student’s work should lie between two performance levels, causing different graders to make variable decisions. The approach used in the

class for this study is based on the number of evidence items identified, which simplifies the discriminability, but can still lead to conflicted emotions for the grader if the number of evidence items required for a given performance level does not feel representative of a student's performance.

The final cognitive function within designing grading guidelines is the 'Assemble rubric elements' function. In this function, the IST member brings together the learning objectives, the evidence items, information about what aspects of the students' responses will be graded, additional information specific to the assigned task, and cutoffs for each performance level. The output is the rubric, which is one of the most important documents in the entire grading system. While there is variability associated with each of the elements that comprise the rubric, there are also unique aspects of the assembled rubric that can contribute to variability in the system. First, the total amount and arrangement of information on a given rubric can affect the rubric's overall "usability" (i.e., how easily the grader can find the information they need and how likely they are to search for it). Popham (1997) and Tierney and Simon (2004) emphasized that concision is important in rubrics, as more text generally leads to information being ignored. This may relate to excessive cognitive load when there are too many pieces of information that grader must consider simultaneously (Sweller, 1994). Another way the rubric can vary is how it specifies portions of responses to be graded. When the area to be graded is a small fraction of the student's entire response or when the same portions are graded repeatedly, there is a chance that a single error in that portion of work can cause the chosen grade to feel misaligned with the quality of the work, leading to variable decision making in graders. Finally, based on the output of the 'Anticipate range of performances' function, the rubric may vary in its robustness (i.e., how well it handles variable performances from the students). If the rubric is not sufficiently robust, there can be considerable variability in how graders evaluate a response that is not appropriately addressed by the rubric.

Cognitive functions associated with designing grader training

The final IST generalized function of designing grader training consists of the five cognitive functions summarized in Table 4.22. As Figure 4.9 demonstrates, there is a bit of an appropriate sequence for four of the five functions to be performed properly. Communicating the

grading expectations is related to training graders but, as a background function, is a broader action that should begin before the semester and be repeated throughout the semester.

Table 4.22. Cognitive level IST functions associated with designing grader training

Function	Communicate grading expectations	Create sample problem	Create or select example cases	Create quiz feedback	Create grading quizzes
Description	Communicate the purpose of training and the underlying philosophy of grading in the course	Write an example task intended to elicit performance of the LO	Select or generate sample responses to the sample problem to demonstrate how to properly apply the rubric	Create the feedback that will be given to the graders upon completion of the quiz	Create a quiz for the graders to report their scoring decisions for the sample responses to the sample problem
Input	---	<ul style="list-style-type: none"> • LO-based tasks 	<ul style="list-style-type: none"> • Sample problem 	<ul style="list-style-type: none"> • Sample problem • Example responses 	<ul style="list-style-type: none"> • Rubric • Sample problem • Example responses • Designed feedback
Output	<ul style="list-style-type: none"> • Grading expectations 	<ul style="list-style-type: none"> • Sample problem 	<ul style="list-style-type: none"> • Example responses 	<ul style="list-style-type: none"> • Designed feedback 	<ul style="list-style-type: none"> • Grading quiz
Precondition	---	<ul style="list-style-type: none"> • Assigned task 	---	---	---
Resource/ E.C.	---	---	<ul style="list-style-type: none"> • Official task solution • Anticipated performances 	<ul style="list-style-type: none"> • Rubric • Official task solution 	---
Control	---	---	<ul style="list-style-type: none"> • Rubric 	---	---
Time	---	---	---	<ul style="list-style-type: none"> • Training deadline 	<ul style="list-style-type: none"> • Training deadline

The ‘Communicate grading expectations’ function is important throughout the system for many of the graders’ functions. This function consists of explaining to the graders the general philosophy behind grading in the course (i.e., to grade based on the performance of specific learning objectives), the expectation that graders should make every effort to produce accurate grades by training and making an effort to thoroughly understand the tasks, the solutions, and the way to apply rubrics, and the purpose of training as a means to calibrate grading decisions rather than to “teach” how to grade, as many participants expressed during their interviews. Table 4.23 shows how the expectations communicated can vary in terms of clarity and forcefulness, affecting how strictly the graders adhere to the expectations.

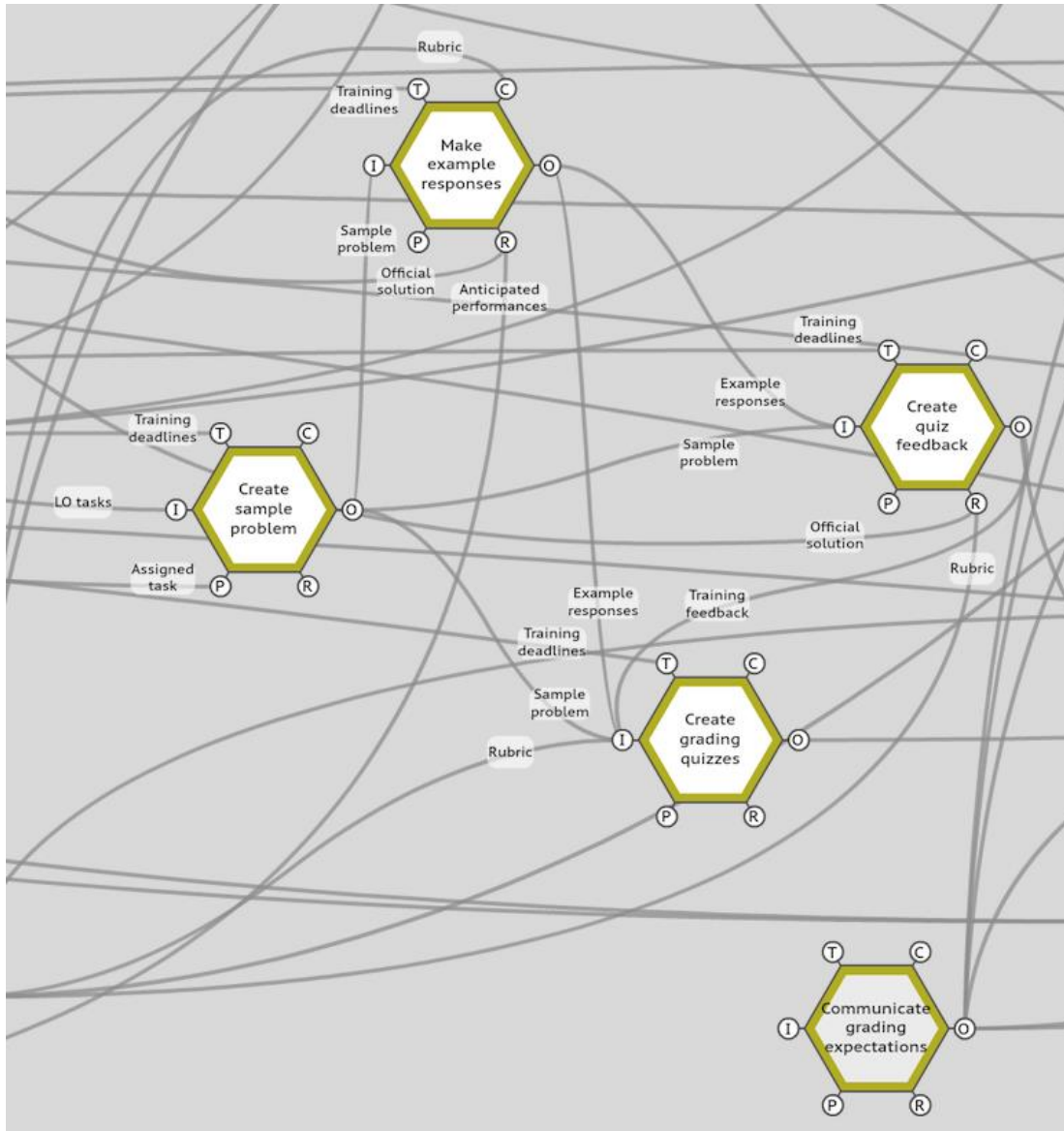


Figure 4.9. Visualization of training design functions.

Table 4.23. Potential variability of training design functions

IST			
Generalized function: Design grader training			
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Communicate grading expectations	Clarity	<u>Unclear</u> : Expectations are not clearly communicated; graders do not have a good grasp of how they should be approaching training or grading [V↑] <u>Clear</u> : Expectations are clear; graders understand appropriate philosophy and approaches [V↓]	Possible, likely Possible, unlikely
	Forcefulness	<u>Weak</u> : Expectations are not reiterated or strictly enforced; graders are less likely to adhere [V↑] <u>Strong</u> : Expectations are reiterated and enforced; graders adhere to expectations [V↓]	Possible, likely Possible, unlikely
Create sample problem	Alignment	<u>Misaligned</u> : The sample task does not align with the assigned task that will be graded and graders struggle to generalize training to grading [V↔] <u>Aligned</u> : The sample task is identical to the assigned task and the graders easily can directly apply learning from training to grading [V↓]	Possible, likely Possible, likely
		<u>Not at all representative</u> : The sample cases bear no resemblance to actual student work; the training is ineffective at preparing and calibrating graders [V↔] <u>Partially representative</u> : The sample cases demonstrate some of what graders will see from students, but likely miss possible responses [V↔] <u>Completely representative</u> : The sample cases fully address all possible responses, allowing graders to perfectly calibrate [V↓]	Unlikely Typical Unlikely
Create quiz feedback	Accuracy	<u>Inaccurate</u> : Feedback regarding the “definitive” grading decisions are incorrect with respect to the sample, giving graders false information [V↑] <u>Accurate</u> : Feedback is accurate; graders can learn how they are expected to evaluate work similar to the samples they are shown [V↓]	Possible, unlikely Typical
	Specificity	<u>Not specific to training needs</u> : Feedback lacks detail to explain the rationale behind the ‘definitive’ decisions, making calibration difficult [V↔] <u>Specific to training needs</u> : Feedback is detailed enough that graders are able to understand and learn from examples of ‘definitive’ decisions [V↓]	Possible, likely Possible, unlikely
Create grading quizzes	Usefulness	<u>Not useful</u> : The sample problem, responses, and/or feedback do not help graders to understand how to make grading decisions or the format of the quizzes limit attentional completion of quizzes [V↔] <u>Useful</u> : The samples and feedback expose the graders to learning objectives and evidence items and demonstrate applications to assist future grading decisions [V↓]	Unlikely Likely

The first function the IST must perform to create the training for each learning objective is to ‘Create a sample problem.’ This should at least align strongly with the assigned task, to best facilitate grading. That said, even if the training uses the identical task for training as the assigned task, it is likely that the same learning objective will be applied again to a different task, inherently making the training less aligned with subsequent uses. Ideally, the graders would be able to generalize what they learn about applying a given learning objective rubric to any use of that learning objective, regardless of the assigned task; however, interviews with graders revealed that the graders struggled, or perceived struggle, with this generalization when the tasks were not identical. It should also be noted that a different task is often utilized for training if the training modules are created before the assignment is fully completed, which occasionally does occur.

Following the sample problem, the IST must ‘Create or select example cases’ or responses representing student work. The sample responses need to be answers to the sample problem, which means that if the sample problem was not used in the class in previous semester, artificial student responses need to be generated. In either case, it is important to consider the full range of possible student responses to make the training samples as representative of what the graders will see while grading as possible to be maximally instructive, noting that Crisp (2010) recommends exemplars for each performance level with a rubric.

Before the quiz is administered to graders for training, the IST should ‘Create quiz feedback,’ that will help the graders to learn from the sample cases used in the training modules. Table 4.22 shows that, ideally, that means the graders should be informed of the “definitive marks” for each grading decision and should be given a specific explanation as to why it is the “definitive” decision for that particular sample case based on the rubric’s specifications and what constitute acceptable responses. As Table 4.23 shows, this means that the feedback may vary in terms of whether the feedback is fully accurate and gives adequately specific feedback for graders to learn from their incorrect decisions during training. During the interviews with graders, several suggested that there were occasional errors in the training feedback and that the feedback was not sufficiently specific for them to understand why their grading decisions were incorrect, limiting their ability to calibrate their grading decisions.

When the IST ‘Creates grading quizzes,’ the elements of the three functions before it are combined along with the rubric and delivered to the graders using an online platform. The variability of each of the inputs to this function lead to the overall usefulness of the quizzes for

training and calibrating grading decisions. The alignment of the sample problem, the representativeness of the sample responses, and the accuracy and specificity of the feedback all contribute to the quality of the training experience. Additionally, however, the format, timing, and frequency of quizzes, along with the effectiveness of the expectations communicated to the graders also affect the extent to which the graders appropriately utilize and make the most of the training. As was communicated in the interviews with graders, the number of separate documents made training unwieldy and the number of quizzes they had to complete made training feel overwhelming, leading many graders to frankly admit that they either did not complete all of the training modules or did not take the training seriously.

4.3.2 Teaching team cognitive functions

There are seven teaching team functions at the cognitive level of abstraction, most of which directly utilize outputs from the IST cognitive functions and either indirectly or directly affect the student functions (see Figures 4.10 and 4.11). The teaching team is the intermediary between the IST and the students. While the IST functions occur anywhere from once every few semesters to multiple times throughout the semester, the teaching team functions occur either before or during every class session. Variability of the teaching team functions directly affects the variability in the quality of students' performances on the assigned tasks.

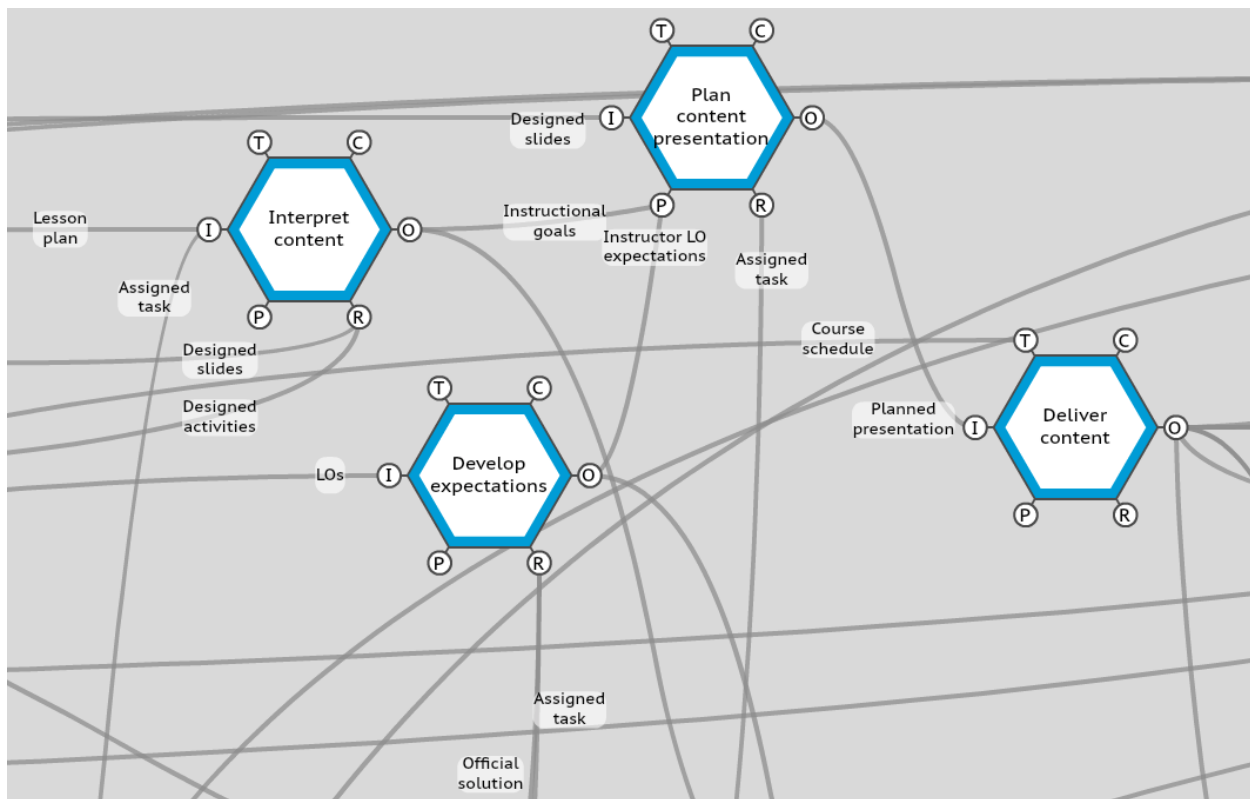


Figure 4.10. Visualization of course content delivery functions.

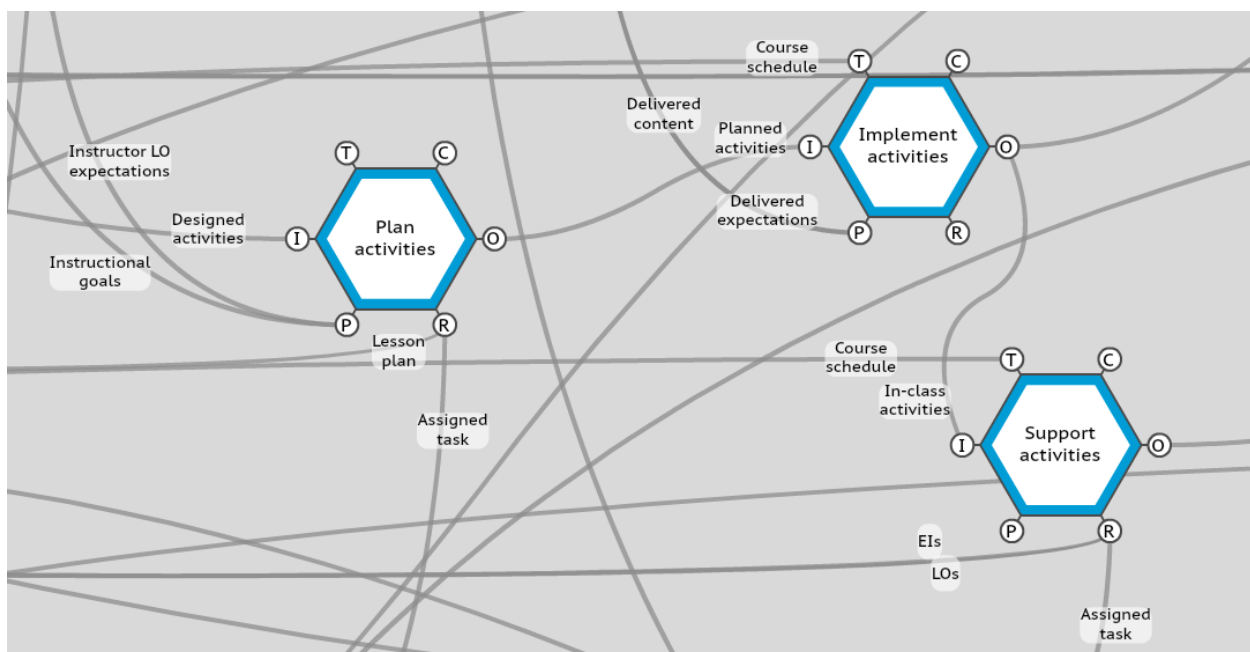


Figure 4.11. Visualization of activity guidance functions.

Cognitive functions associated with delivering course content

The generalized teaching team function of delivering course content consists of four internal cognitive functions, as shown in Table 4.24. While the teaching team consists of an instructor, a GTA, four in-class UTAs, and two out-of-class UTAs, these first four functions are primarily the responsibility of the instructor; however, some instructors may involve their teaching assistants with planning and delivering content.

Table 4.24. Cognitive teaching team functions associated with delivering course content

Function	Interpret content	Develop expectations	Plan presentation of content	Deliver content
Description	Interpret instructional goals for class session	Develop expectations for student performance of the learning objectives	Plan for how to present content in class based on interpretation of instructional goals, individual knowledge, lesson plan, slides	Give in-class presentation of content to the students
Input	<ul style="list-style-type: none"> • Lesson plan • Assigned task 	<ul style="list-style-type: none"> • Learning objectives 	<ul style="list-style-type: none"> • Designed slides 	<ul style="list-style-type: none"> • Planned presentation • Delivered course content • Delivered LO expectations
Output	<ul style="list-style-type: none"> • Instructional goals 	<ul style="list-style-type: none"> • Instructor LO expectations 	<ul style="list-style-type: none"> • Planned presentation 	
Precondition	---	---	<ul style="list-style-type: none"> • Instructional goals • Instructor LO expectations 	---
Resource	<ul style="list-style-type: none"> • Designed slides • Designed activities 	<ul style="list-style-type: none"> • Assigned task • Official task solution 	<ul style="list-style-type: none"> • Assigned task 	---
Control	---	---	---	---
Time	---	---	---	<ul style="list-style-type: none"> • Course schedule

The first function the teaching team (or, more specifically, the instructor) must engage in is to ‘Interpret content.’ The IST presents a general lesson plan to the instructors and GTAs in a weekly meeting, along with lecture slides and embedded in-class activities. The instructor and GTA should review the information presented to develop an understanding of the instructional goals for the students in each lesson. Analyzing the assigned task may also help the teaching team members in developing their understanding of the instructional goals. As this process most likely involves a person interpreting the product of a different person, there is some potential that the

instructional goals recognized by the teaching team may not fully align with the learning goals initially identified by the IST members. Further, teaching teams of different sections may differ in their interpretations. This may be a more likely outcome for instructors who have less experiences teaching the course. Table 4.25 summarizes the potential variability of this function, and the other cognitive functions related to content delivery.

Table 4.25. Potential variability of course content delivery functions

Teaching Team	Generalized function: Deliver course content		
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Interpret content	Alignment	<u>Misaligned:</u> Teaching team's instructional goals do not align with IST's learning goals; content presentation may vary from IST's intentions [V↑]	Possible, unlikely
		<u>Aligned:</u> Teaching team's instructional goals align with IST's learning goals; content likely to be presented as intended [V↔]	Typical
Develop expectations	Alignment	<u>Misaligned:</u> Teaching team's LO performance expectations do not align with IST's; expectations communicated may diverge from IST's intentions [V↑]	Possible, unlikely
		<u>Aligned:</u> Teaching team's LO performance expectations align with IST's; expectations communicated closely represent those intended by IST [V↔]	Typical
Plan presentation of content	Effectiveness	<u>Ineffective:</u> Content, including performance expectations, are not accurately portrayed by the presentation [V↑]	Possible
		<u>Effective:</u> Content effectively communicates the intended content [V↔]	Typical
Deliver content	Alignment	<u>Misaligned:</u> Content presented deviates from planned presentation; students do not receive content as intended [V↑]	Possible, unlikely
		<u>Aligned:</u> Content is presented as planned [V↔]	Typical

Based on the instructional goals identified, the teaching team must 'Develop expectations' for performance of learning objectives. This process is facilitated by the assigned task and approach demonstrated by the official solution. Like the previous function, the primary way performance expectations may vary is with respect to their alignment with the LO expectations intended by the IST.

Based on their understanding of the instructional goals and LO performance expectations of the teaching team, along with the background knowledge and experience the teaching team members bring to the process, they can then employ the 'Plan presentation of content' function to develop the slides they plan to present. The IST provides a slide deck, which may be delivered

exactly as is, or may be revised by the teaching team as they see fit. This may include removing information they believe to be confusing or too time intensive or adding content they believe will help illustrate the concepts. Misalignment in either the instructional goals or the LO performance expectations may cause the planned presentation to vary in its effectiveness at communicating the content to the students.

On the day of each class session, the teaching team then ‘Delivers content’ to students using the planned presentation. Typically, the presentation goes as planned; however, there are times when circumstances deviate from expectations and the content is ultimately delivered differently than was planned. For example, the instructor could get a flat tire and depend on the GTA to teach the class session, who may have a different interpretation of the content or not understand the purpose of added slides or content. Alternatively, an activity may take an unexpectedly long amount of time and the instructor may need to rush through content they originally intended to spend more time developing. In any case, what was presented to the students may vary from what was planned.

Cognitive functions involved in guiding student practice

There are three cognitive functions that constitute the generalized function of guiding student practice. Like the last set of teaching team functions, the instructor is the primary actor for the first two functions. The third function in this group, however, is strongly impacted by the GTA and in-class UTAs, which can cause significant variability. That said, Figure 4.11 shows that these functions, like the previous set of functions, only directly affect a few downstream functions—namely, the students’ cognitive functions. Table 4.26 summarizes the three functions involved in guiding student practice.

Table 4.26. Cognitive teaching team functions involved in guiding student practice

Function	Plan activities	Implement activities	Support activities
Description	Plan the activities to be used in class to support student learning of content and development of LO proficiency	Implement the planned activities during the class session	Provide guidance and support to students during the in-class activities
Input	• Designed activities	• Planned activities	• In-class activities
Output	• Planned activities	• In-class activities	• Activity support
Precondition	• Instructional goals • Instructor LO expectations	• Delivered course content • Delivered LO expectations	---
Resource	• Lesson plan • Assigned task	---	• Assigned task • Learning objectives • Evidence items
Control	---	---	---
Time	---	• Course schedule	• Course schedule

Before the students can be guided, the teaching team must ‘Plan activities.’ This process consists of the teaching team appraising the activities designed by the IST and deciding whether those activities should be implemented as planned by IST, with some amount of modification, thrown away completely, or replaced with alternative activities. Ultimately, the same instructional goals and LO expectations need to be communicated, and the students need to gain practice performing the learning objectives with the benefit of the guidance of a more knowledgeable other before they should be evaluated on doing so. As instructors have the authority to modify the activities, the planned activities may vary in their alignment with the designed activities and, consequently, how appropriately they support learning of the intended content. Table 4.27 summarizes the output variability.

Table 4.27. Potential variability of in-class activity guidance functions

Teaching Team	Generalized function: Guide in-class activities and practice		
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Plan activities	Alignment	Misaligned: Planned activities fail to support the learning outcomes expected from the original designed activities [V↑]	Unlikely
		Aligned: Planned activities are the same or support the intended learning outcomes effectively [V↔]	Typical
Implement activities	Alignment	Misaligned: In-class activities deviate from envisioned activities; students lack intended learning opportunity [V↑]	Possible, unlikely
		Aligned: In-class activities align with envisioned activities; students experience intended learning opportunity [V↔]	Possibly, likely
Support activities	Quality	Low: Students receive guidance that is, at best, not helpful, or, at worst, contradicts interpretations of other members of the teaching team [V↔ or V↑]	Possible, likely
		High: Students are given high quality guidance to facilitate learning the content and practice performing LOs [V↓]	Possible, likely

After planning the activities, along with the lesson plans, the teaching team must ‘Implement the activities’ during the class session. As with giving the presentation, various circumstances can cause the planned activities to deviate from expectations, such as having the planned time for the activity cut short, which can reduce the learning opportunities for the students. Along similar lines, the activities are now being relayed to a class of up to 120 students, and any number of circumstances could limit the transmission of the information or intended goals of the activity from the instructor to the students. As such, the implementation of the activity may not be aligned perfectly with the activity that was envisioned by the instructor.

The final teaching team function is a bit more complex than the others. With each of the teaching team functions up to this point, the instructor was likely the primary actor (although some instructors may involve their TAs in instructional tasks more than other). However, the other members of the teaching team still engage in the cognitive functions of interpreting content and developing LO performance expectations. The possibility of variability in those interpretations and expectations comes to the fore in the final teaching team function of ‘Supporting activities,’ because the primary actor is no longer the instructor. In fact, the instructor outnumbered in the goal of supporting the in-class activities, as there are five TAs and only one instructor. As a result, even though each member of the teaching team should give feedback and guidance to students during in-class activities based on the explicit learning objectives and evidence items to support performance on the assigned activities, it is possible that different members of the teaching team

have differing interpretations of those materials. Further, it is likely that the different members of the teaching team will vary in their ability to support student learning. This means that the support and guidance provided to students varies significantly in quality from student to student. To make matters worse, students vary in their tendency to seek help during the in-class activities and not all members of the teaching team are equally skilled at identifying students in need and intervening when necessary.

4.3.3 Student cognitive functions

There are four student functions at the cognitive level of abstraction, which are most directly affected by the teaching team's functions. While the system grows substantially when the teaching teams enter the system, as there are well over a dozen teaching teams, the system grows more than exponentially with the entrance of the students. As a result, even though the students engage in so few functions, the sheer number of students involved and the probability of any given student moving down a different potential path vastly increases the overall system's variability. The students must engage in these functions with every class session and/or assignment. As shown in Figure 4.12, the students produce one external output: their performance of the assigned task, which is the very object that the graders must ultimately evaluate.

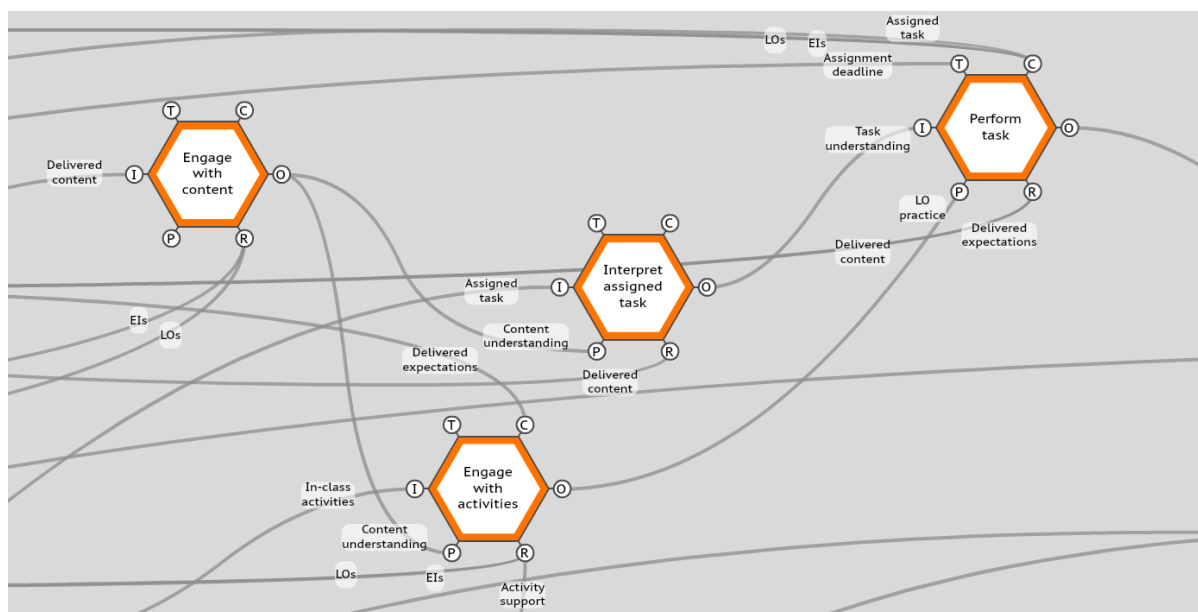


Figure 4.12. Visualization of content learning and task performance functions.

Cognitive functions associated with learning course content

The students engage in two cognitive functions associated with the generalized function of learning course content, as shown in Table 4.28. The first function is to engage with the content that is presented to them. As with the teaching team having to interpret the learning goals of the IST, introducing potential variations in the content being delivered, the student's engagement with the content represents an additional opportunity for misunderstanding and varied output. While official documentation of learning objective and evidence items are available, the students must develop their own understanding of the content based on course content that is delivered to them by the teaching team. Further, even though the LOs and EIs are available to the students, it does not guarantee the students will utilize them to facilitate their learning. Thus, the extent to which students develop content understanding is highly variable based on many, primarily internal, factors, such as background knowledge and experience, studiousness, and attention to detail. Further, external factors such as ability to filter out distractions in class and pay attention to the instructor, as well as the alignment of the content presented to the student by the teaching team can affect the way the students learn. Table 4.29 summarizes this variability.

Table 4.28. Cognitive student functions associated with learning course content

Function	Engage with content	Engage in practice activities
Description	Read lecture slides and listen to lecture with an effort to learn the content	Participate in in-class activities as directed with an effort to learn and practice content
Input	<ul style="list-style-type: none">• Delivered course content	<ul style="list-style-type: none">• In-class activities
Output	<ul style="list-style-type: none">• Content understanding	<ul style="list-style-type: none">• Experience with LOs
Precondition	---	<ul style="list-style-type: none">• Content understanding
Resource	<ul style="list-style-type: none">• Learning objectives• Evidence items	<ul style="list-style-type: none">• Activity support• Learning objectives• Evidence items
Control	---	<ul style="list-style-type: none">• Delivered LO expectations
Time	---	---

Table 4.29. Potential variability of content learning functions

Students	Generalized function: Learn content		
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Engage with content	Alignment	<u>Misaligned</u> : Student develops an understanding of the content that differs from the instructor's content interpretation, likely contributing to more variable task performance [V↑]	Possible, likely
		<u>Aligned</u> : Student develops an understanding of the content consistent with the instructor's, allowing them to more easily perform the task as the instructor envisions [V↓]	Possible, likely
	Quality	<u>Low</u> : Student struggles to engage with content and, even if aligned with instructor, lacks sufficient understanding to know what is expected of them, affecting their ability to perform the assigned task [V↑]	Possible, likely
		<u>High</u> : Student engages deeply with content and learns it well enough to fully understand expectations and can likely perform the task as expected [V↓]	Possible, likely
Engage in practice activities	Quality	<u>Low</u> : Student does not engage intently in practice or does not receive guidance or support from peers or teaching team to gain quality experience performing the LO, thus unlikely to do so on the assigned task [V↑]	Possible, likely
		<u>High</u> : Student gains high quality experience practicing the task in class and is likely to perform the task as expected in the assignment [V↓]	Possible, likely

The students should also spend the allocated time in class to 'Engage in practice activities.' Differences in personalities across students will affect the way the students approach the in-class activities; however, the ability of the students to effectively engage with the practice activities may also be greatly affected by the students around them, who may significantly improve or detract from their learning experiences and the support they get from peers and the teaching team. As a result, the output varies primarily in terms of the quality of the experience they obtain while practicing performance of the learning objectives. The large number of students ensures that some students will likely engage deeply with the content and practice while others will not. Thus, variability of these functions is likely inevitable, if reducible. It is also important to emphasize that in some cases, students may very well try to engage, but still struggle to develop understanding or to gain high quality experience performing the LOs—these are the students the teaching team needs to support the most.

Cognitive functions involved in performing the assigned task

The second two cognitive functions of the students are to 'Interpret the assigned task,' and to 'Perform the task,' as summarized in Table 4.30. These functions are both dependent upon the

quality to which the students engaged in the previous two cognitive functions. Students who developed a strong content understanding and experience with the learning objectives will be more likely to perform the assigned task well. That said, the additional sources of variability intrude. For example, even if a student has a strong content understanding, it is possible that they struggle to understand the context or instructions provided in the assigned task. As a result, their understanding of the task may not align with the actual context or goals of the task—this is almost directly a function of the clarity of the task context and instructions. If they then utilize a faulty understanding of the task, it does not matter whether they utilize the LO expectations and content delivered to them and refer to the task instructions, LOs, and EIs while performing the task, they will still likely perform the task incorrectly. Thus, the task performance will vary in terms of its overall quality. The task performance will also vary in terms of how easy it is for another person, namely the grader, to understand (i.e., how legible is the work and how clear is it what the student was trying to do). Further, the task performance can vary by the conventionality of the solution, where unconventional responses are less likely to be aligned with the rubric, challenging graders' grading decisions.

Table 4.30. Cognitive student functions associated with performing the assigned task

Function	Interpret assigned task	Perform task
Description	Interpret the goals of the assigned task based on the context and instructions	Attempt to perform the task based on understanding of content, LO expectations, and task
Input	<ul style="list-style-type: none"> • Assigned task 	<ul style="list-style-type: none"> • Task understanding
Output	<ul style="list-style-type: none"> • Task understanding 	<ul style="list-style-type: none"> • Task performance
Precondition	<ul style="list-style-type: none"> • Content understanding 	<ul style="list-style-type: none"> • Experience with LOs
Resource	<ul style="list-style-type: none"> • Delivered course content 	<ul style="list-style-type: none"> • Delivered LO expectations • Delivered course content • Task instructions
Control	---	<ul style="list-style-type: none"> • Learning objectives • Evidence items
Time	---	<ul style="list-style-type: none"> • Assignment deadlines

Table 4.31. Potential variability of task performance functions

Students	Generalized function: Perform assigned tasks to demonstrate learning		
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Interpret assigned task	Alignment	<u>Misaligned</u> : Students interpretation of the task deviates from the intended interpretation, likely leading to a different performance than expected [V↑]	Possible, unlikely
		<u>Aligned</u> : Student interprets the task as intended, understands what is involved in performing the task [V↔]	Possible, likely
Perform task	Quality	<u>Low</u> : Student performance is very poor and does not represent adequate performance of the LO, making it easy to grade consistently [V↓]	Possible, likely
		<u>Moderate</u> : Student performance is of middling quality with some aspects of the LO achieve and others not, or the student has somewhat achieved the goal of the task through unexpected means, making consistent grading difficult [V↑]	Typical
		<u>High</u> : Student performance of task is strong, and student clearly demonstrates achievement of the LO, making it easy to grade consistently [V↓]	Possible, likely
	Typicality	<u>Atypical</u> : Student performance is unlike that which might have been anticipated in advance; may align poorly with rubric; grading decisions are difficult and inconsistent [V↑]	Possible, unlikely
		<u>Typical</u> : Student performance fits within typically expected responses; may align well with rubric; grading decisions are easy to make and consistent [V↓]	Likely
	Clarity	<u>Unclear</u> : Student performance is difficult to understand, either due to limited legibility or student's struggle to communicate their work effectively; graders more likely to have variable interpretations [V↑]	Possible, likely
		<u>Clear</u> : Student work is straightforward and easy to interpret; it would be difficult for graders to interpret incorrectly [V↓]	Likely

4.3.4 Graders' cognitive functions

There are thirteen core functions at the cognitive level of abstraction that graders consistently utilize, one of which comprises up to six separate sub-functions that will vary greatly depending on the specific context of the assigned task, learning objective being evaluated, and the student's work. The graders are always be expected to train to calibrate their grading decisions and to take appropriate steps at the start of their grading to orient themselves toward the specific tasks they are evaluating. Beyond that, the functions the graders perform in the process of evaluating the work are dependent directly on the consequences of the earlier functions' output variables. The complexity of the evidence items directly influences the level of processing the grader is expected to do to determine if the student's work demonstrates that item sufficiently. For example, determining that the first selection structure used is an 'if' statement only requires the grader to

locate the selection structure and check that it is, in fact, an ‘if.’ On the other hand, determining that the input values of a test case involving several variables will lead to the claimed output requires at least a rudimentary calculation to verify. Further, even if the evidence item dictates a certain approach, a particular student’s response may render the intended evaluation method moot if that portion of the response is missing or severely off-base. The variability in use and outputs of these functions will directly contribute to the potential variability of the scoring of the students work.

Cognitive functions associated with training to calibrate grading decisions

Table 4.32 summarizes the five cognitive functions that the grader should engage in while training to calibrate their grading decisions. As discussed in the Methods chapter, the training modules consist of a set of documents, including a sample problem, a solution to the sample problem, a rubric, and two examples of student work. The graders are then expected to attempt to grade the two samples and input their grading decisions into the software in which the quizzes are given (Blackboard, at the time of this study). The graders are then given feedback about their performance on the quizzes, which they are expected to use to revise their understanding of how to properly apply the rubrics and to get a general sense of what types of responses they can anticipate from students. Figure 4.13 shows a visual representation of how these functions interact.

Table 4.32. Cognitive grader functions associated with training to calibrate grading decisions

Function	Interpret LO	Interpret EI	Complete quiz	Calibrate grading decisions	Anticipate student responses
Description	Read and develop an interpretation of the learning objective	Read and develop interpretation of the evidence items in the learning objective	Complete the quiz to practice grading and receive calibration feedback	Use the results of the training quiz and provided feedback to calibrate understanding of applying the rubric	Develop a mental model of the range of performances expected for the LO
Input	<ul style="list-style-type: none"> • Learning objectives 	<ul style="list-style-type: none"> • Evidence items 	<ul style="list-style-type: none"> • Grading quiz 	<ul style="list-style-type: none"> • Quiz results 	<ul style="list-style-type: none"> • Rubric understanding
Output	<ul style="list-style-type: none"> • LO interpretation 	<ul style="list-style-type: none"> • EI interpretation 	<ul style="list-style-type: none"> • Quiz results • Exposure to examples • Training feedback 	<ul style="list-style-type: none"> • Rubric understanding 	<ul style="list-style-type: none"> • Performance expectations
Precondition	---	---	<ul style="list-style-type: none"> • Grading expectations 	<ul style="list-style-type: none"> • Grading expectations 	---
Resource/ E.C.	---	---	<ul style="list-style-type: none"> • LO interpretation • EI interpretation 	<ul style="list-style-type: none"> • Training feedback • Learning objectives • Evidence items 	<ul style="list-style-type: none"> • Exposure to examples • Delivered LO expectations
Control	---	---	<ul style="list-style-type: none"> • Rubric 	---	---
Time	---	---	<ul style="list-style-type: none"> • Training deadlines 	<ul style="list-style-type: none"> • Grading deadlines 	<ul style="list-style-type: none"> • Grading deadlines

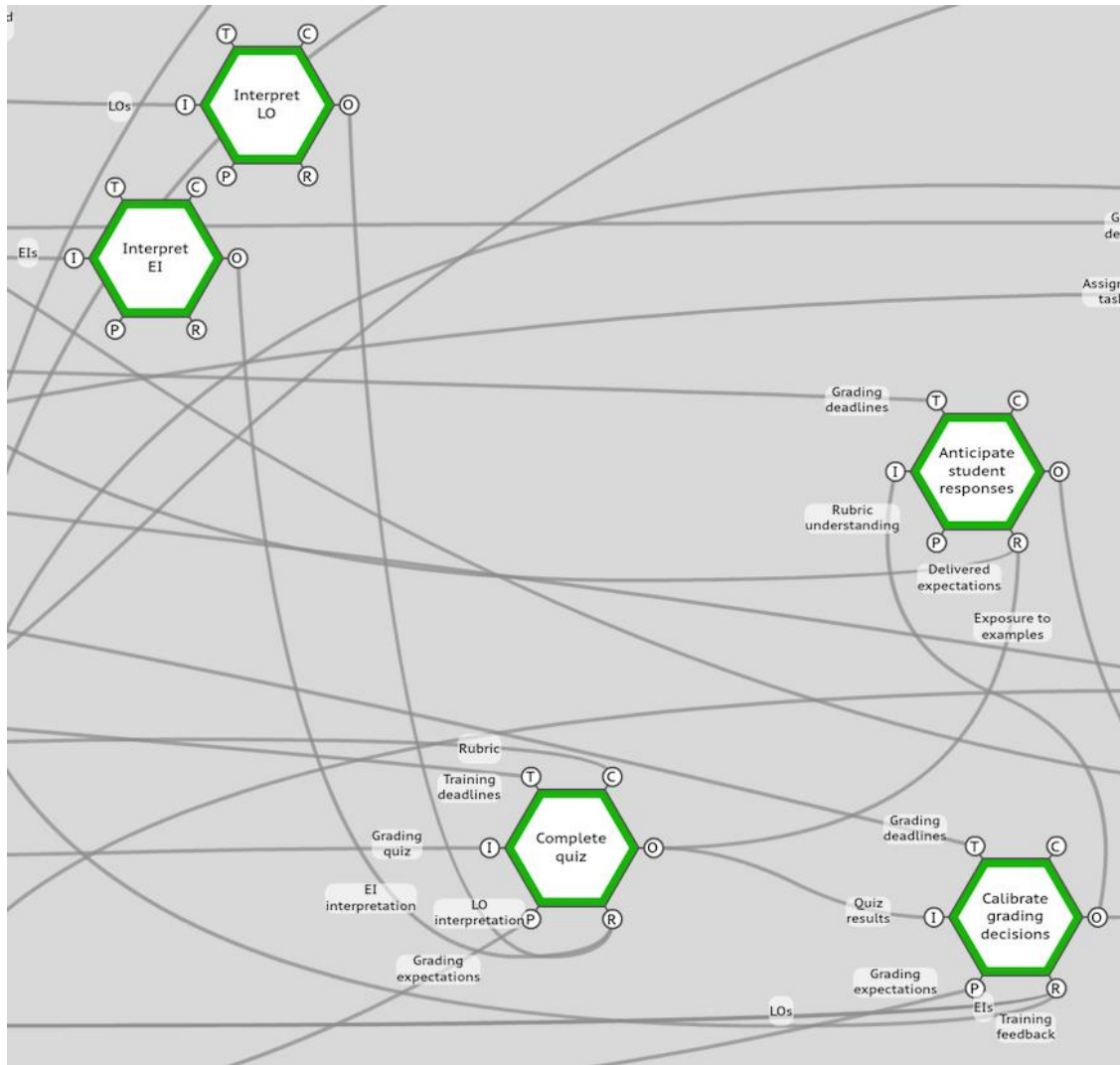


Figure 4.13. Visualization of grader training functions.

The first cognitive function employed by graders during training is to ‘Interpret the LO,’ followed by the need to ‘Interpret EIs.’ These likely occur at, more-or-less, the same time. They are both simple functions taking one input (i.e., the LO or EI) and one output (i.e., the corresponding interpretation). For both of these functions, the outputs vary in terms of whether or not the interpretations developed by the grader aligns with the interpretations expected and intended by the IST. Notably, however, each LO has a variable number of EIs associated with it. As such, the grader may need to apply the ‘Interpret EI’ function repeatedly to develop a full interpretation. When only two EIs are present for a given LO, it is more likely that the grader will read and interpret properly each EI. On the other hand, an LO with 12 EIs may incentivize a more

cursory reading of the EIs with less thought process devoted toward interpretation. This could be the result of an efficiency-thoroughness tradeoff decision of the grader or due to overloaded cognitive demand associated with interpreting all of the EIs at once. Either way, along with the alignment, clarity, precision, coverage, and complexity, the number of EIs (as related to precision and coverage) may correspond to increasingly variable interpretations across graders. The variabilities of these and other training functions are summarized in Table 4.33.

Table 4.33. Potential variability of grader training functions

Graders		Generalized function: Train to calibrate grading decisions	
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Interpret LO	Alignment	<u>Misaligned</u> : Grader's interpretation of the LO is not aligned with the IST's intended interpretation, possibly due to lack of clarity of the articulated LO, making training quiz harder [V↑]	Possible, unlikely
		<u>Aligned</u> : Grader's interpretation aligns with the IST's intended interpretation; quizzes are likely to be easier to do perform well [V↓]	Typical
Interpret EI	Alignment	<u>Misaligned</u> : Due, potentially, to many factors associated with the input, grader interpretation of the EI may not align with the IST's intended interpretation, causing graders to evaluate different things than intended [V↑]	Possible, likely
		<u>Aligned</u> : The grader's interpretation of the EI aligns with the IST's intended interpretation, making training easier [V↓]	Possible, likely
Complete quiz	Internalization (of examples)	<u>Superficial</u> : Graders do not pay close attention to the examples; ability to develop an understanding of the rubric and expectations of student performance is limited, leading to a weaker grasp of what constitutes acceptable responses [V↑]	Possible, likely
		<u>Deep</u> : Graders pay close attention to the examples; gain a stronger ability to understand the rubric and expectations of student performance and, ultimately, a strong mental solution model [V↓]	Possible, unlikely
	Value (of results)	<u>Low</u> : Grader earns a low score on the quiz; depending on the grader's personality and the effort they put into the quiz, this could either motivate or discourage efforts to improve rubric understanding [V↑ or V↓]	Possible, likely
		<u>High</u> : Grader earns a high score on the quiz; depending on the grader's personality and the effort they put into the quiz, this could reinforce understanding (appropriately or not) or could lead to dismissal of the provided feedback [V↑ or V↓]	Possible, likely
	Relevance (of feedback)	<u>Irrelevant</u> : Feedback does not provide information to related to the way the grader interpreted or graded the sample, limiting effectiveness as supporting training and calibration [V↔]	Possible, likely
		<u>Strongly relevant</u> : Feedback directly addresses an error in the grader's understanding of how to apply the rubric [V↓]	Possible, likely

Figure 4.33 continued

Graders		Generalized function: Train to calibrate grading decisions	
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Calibrate grading decisions	Alignment	<u>Misaligned</u> : Grader's understanding of how to apply the rubric does not align with the IST's perception of definitive marking; grader is more likely to make inappropriate grading decisions [V↑]	Possible, likely
		<u>Aligned</u> : Grader's understanding of how to apply the rubric aligns with the IST's perception of definitive marking; grader is more likely to make appropriate grading decisions [V↓]	Possible, likely
	Internal: Timing	<u>Distant</u> : Understanding was developed (i.e., training) many weeks prior and may not be fresh in the grader's mind when grading; grader may forget important ideas and may be less consistent [V↑]	Possible, likely
		<u>Recent</u> : Understanding was developed right before grading; grader has important ideas fresh in their mind and more likely to be consistent [V↓]	Likely
	Internal: Experience	<u>None</u> : Grader has no previous experience evaluating the learning objective; less likely to have a strong grasp and apply as expected [V↑]	Possible, likely
		<u>High</u> : Grader has a lot of experience evaluating the learning objective; may have a strong grasp on how to evaluate [V↓] <i>or</i> may be overly confident in ability and pay less attention to specifics [V↑]	Possible, less likely
Anticipate range of student performances	Comprehensiveness	<u>Narrow</u> : Grader does not develop a strong sense of how students will reply, cannot consistently identify how to evaluate common performance patterns [V↑]	Possible, likely
		<u>Broad</u> : Grader considers all likely task performances and understands how they should be marked with the rubric in advance, facilitating more consistent grading decisions [V↓]	Possible, unlikely

Next, the graders use their interpretations along with the sample problem text, sample solution, and rubric to evaluate the two student samples to 'Complete the quiz.' The graders select whether or not each evidence item is satisfied by the sample and the overall learning objective proficiency level in one quiz and just the proficiency level for the second quiz. Completing the quiz will give the graders an exposure to examples of student work and will get results and training feedback based on their grading decisions. As long as the graders complete the quiz, the exposure to examples of work will be consistent; however, the graders may vary in terms the extent to which they internalized the response patterns demonstrated by each sample. The results will vary based on the accuracy of their scoring selections with respect to the definitive scoring selections, which can range from making all incorrect selections to making all correct selections. Finally, the training feedback will vary with respect to how relevant it is based on the particular grading decisions they

made during the training quizzes. Notably, when this study was conducted, the feedback provided was singularly designed for all graders regardless of how they graded. As a result, the feedback may not have addressed the specific way the grader interpreted the sample response or made grading decisions.

After completing the quiz, the grader should interpret the results, with guidance from the feedback provided and with reference to the documents, to develop a calibrated understanding of how to apply the rubric. Likely, if they did well on the quiz, they will interpret the results as meaning they are on the right track and do not need to adjust their understanding of the rubric; although, the grading expectations should suggest that they read the feedback even if they perform well to ensure their understanding is appropriate, as they could have done well on the quiz despite having a flawed understanding. It is also expected that a poor performance on the quiz should prompt the graders to reflect on their performance, using the feedback to guide their focus on the evidence items they inappropriately selected or failed to select. Regardless, this process leads to an understanding of the rubric that may vary in terms of its alignment with definite marking practices (i.e., its calibration). However, rubric understanding may also vary based on the recency of the training, whereby some learning objectives may be trained early in the semester and assessed again later in the semester without training in between and most are trained the week before using. Similarly, rubric understanding may vary based on how many times the grader has applied the rubric, either in the current semester or past semesters. These latter two variabilities are more internal to the process of developing a calibration and affect the output of alignment of understanding of the rubric.

The graders should next use their understanding of the rubric, along with the example student responses from the training and the expectations for LO performance delivered by their instructor to ‘Anticipate a range of performances’ they are likely to see from the students while grading. Developing this range of expected student responses can help to establish an understanding of how the different types of responses the grader might see would fit within the rubric. Doing this before the start of grading is ideal, so any misunderstandings or misalignments identified between the rubric and expected work can either help to guide the students in class to prevent those errors, improve understanding of how the error is addressed by the rubric so it can be applied consistently, or if the rubric is truly limited or flawed, can lead to revisions of the rubric.

Failing to anticipate expected performances, or doing so inadequately, can limit the grader's ability to consistently handle different infrequent but recurring response patterns from students.

Cognitive functions associated with preparing to evaluate task performance

The generalized function of preparing to evaluate task performance consists of three cognitive functions, which may need to occur to different extents depending on the alignment between the actual assigned task and the training materials and the recency of training. Table 4.34 shows that before the grader grades the first student task performance, they should review the problem, both in terms of the task and the instructions, assigned to the students, develop a mental model of what constitutes an acceptable response, and review the rubric, which is likely to have specifying text that was not present in the training as it is contextualized based on the specific problem. This ensures that the grader is properly oriented to evaluating the task at hand. It is also likely that the grader will return to these functions between or while evaluating student responses and more so when just starting to evaluate a set of responses, as the information may start to be internalized after repeated reviews. Figure 4.14 visualizes these functions, showing their inputs and interactions. The outputs mostly directly influence the evaluation functions.

Table 4.34. Cognitive grader functions associated with preparing to evaluate task performance

Function	Interpret problem	Develop model of acceptable solution	Review rubric
Description	Interpret the assigned task to understand context and directions of the problem	Develop a mental model of what constitutes an acceptable response to the assigned task	Use the rubric to understand the construct being graded in the context of the assigned task
Input	<ul style="list-style-type: none"> Assigned task 	<ul style="list-style-type: none"> Official task solution 	<ul style="list-style-type: none"> Rubric 'What to grade' interpretation 'What to grade' in WM LO in WM EI in WM
Output	<ul style="list-style-type: none"> Problem interpretation Problem in WM 	<ul style="list-style-type: none"> Solution model interpretation Solution in WM 	
Precondition	---	---	---
Resource/ E.C.	---	<ul style="list-style-type: none"> Problem interpretation Performance expectations 	---
Control	<ul style="list-style-type: none"> Grading expectations 	<ul style="list-style-type: none"> Grading expectations 	<ul style="list-style-type: none"> Grading expectations
Time	<ul style="list-style-type: none"> Grading deadlines 	<ul style="list-style-type: none"> Grading deadlines 	<ul style="list-style-type: none"> Grading deadlines

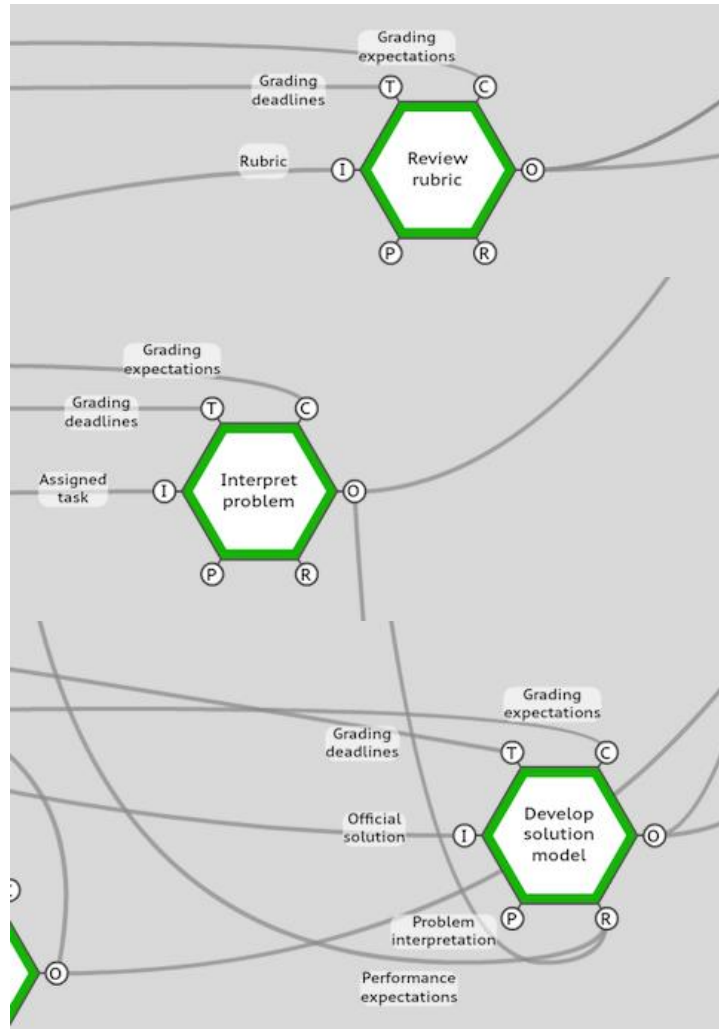


Figure 4.14. Visualization of evaluation preparation functions. Note: this image has had empty space removed to conserve space.

The first cognitive function is to ‘Interpret the problem.’ That is, the grader should read through the problem context and problem instructions to develop a good understanding of what the students were being expected to do. This can help the students to know if a particular task performance is reasonable given what was presented to the students. Understanding the problem should be communicated as an expectation of grading. The result of performing this action is not only developing an understanding or interpretation of the problem but bringing relevant details about the problem to the working memory to facilitate interpretation of the students’ responses. The problem interpretation may vary in alignment with the problem’s interpretation as intended by the IST, which could lead graders to make inappropriate interpretations and decisions about student work. The problem being in the working memory may also vary in terms of how well the

information is actually present in the working memory, which may degrade over time or depend on repeated exposure. The variability of these outputs and those of the other two functions are summarized in Table 4.35.

Table 4.35. Potential variability of preparing to evaluate task performance functions

Graders			
Generalized function: Prepare to evaluate task performance			
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
All three functions	Alignment (of interpretations of the problem, solution, and what to grade)	<u>Misaligned</u> : Grader's interpretation of the corresponding information is misaligned from IST's intended interpretation; decisions are more likely to disagree with definitive marking [V↑]	Possible, likely
		<u>Aligned</u> : Grader's interpretation aligns with IST's intended interpretation; decisions are more likely to agree with definitive marking [V↓]	Possible, likely
All three functions	Presence in WM (problem, solution, what to grade, LO, and EI)	<u>Absent</u> : Grader is missing information crucial to evaluating student task performance from their WM; if not corrected, grading decisions may vary [V↑]	Likely (over time)
		<u>Present</u> : Grader has all information necessary for evaluating student task performance in WM; likely to make appropriate grading decisions [V↓]	Likely

The graders should then use their interpretation of the problem to review the problem solution provided by the IST. They should also incorporate their anticipated range of possible task performances to identify the full range of what they will consider to be acceptable or unacceptable aspects of a student's performance. Note that, as discussed in the IST cognitive functions, the assigned task can vary with respect to open-endedness and the provided solution can vary with respect to how well it communicates the range of acceptable task performances. Taken together, the open-endedness of the task and comprehensiveness of the provided solution and rubric will dictate the extent to which the graders will need to make inferences about what should be considered an acceptable response to award achievement of an evidence item or learning objective. Reviewing the solution will provide the graders with a mental model of acceptable solutions and will bring that model into working memory. These outputs vary like the previous function: the grader's mental solution model may be more or less aligned with those of a definitive marker and the presence of the solution model in the working memory may vary.

The last step the graders need to take before they begin grading is to 'Review the rubric' for the assigned task, which likely has additional context-specific information that was not present in the rubric during training. This information directs the grader to look at specific portions of the

students' task performance and may provide other guiding details related to the specific problem (e.g., for an evidence item of having a test case table that lists all necessary test cases, the test cases that must be present will be different for different problems and the additional text gives this information). In addition to helping the graders know what part of the task performance they need to focus on, which can vary in alignment, the function also brings other key details (i.e., the LO, the EIs, and what to grade) into working memory.

Cognitive functions associated with evaluating task performance

Figure 4.15 shows the visualization of the last two major functions of the grading system. Both of these functions, however, each consist of several sub-functions that are highly variable depending on the contexts established by the earlier functions. The first of these is the 'Evaluate performance' function, which consists of up to six sub-functions. The functions associated with evaluation have been broken down into two sets: those generally using System 1 processing (summarized in Table 4.36) and those generally using System 2 processing (summarized in Table 4.37) (Kahneman & Frederick, 2002; Suto & Greator, 2008).

All of the previous functions discussed are expected to absolutely occur, whether they occurred once months earlier or right before the evaluation happened. The following functions will be highly dependent on the learning objective, evidence items, and task performance, so there is no general work-as-imagined instantiation to illustrate as has been the case with all the previous functions. There are work-as-imagined instantiations for ideal evaluation when a specific context is considered; however, even within these contexts, the work-as-completed instantiations may result in not only variable output of expected functions, but unexpected functions being used or expected functions not being used (see Chapter 5 for more detail). The direct cause of this has to be inferred but is likely due to variability of the 'Prepare to evaluate' cognitive functions or the 'Training' functions leading to misinterpretations, not having the right information in working memory, or attempting to be more efficient, apply heuristics, or grade holistically.

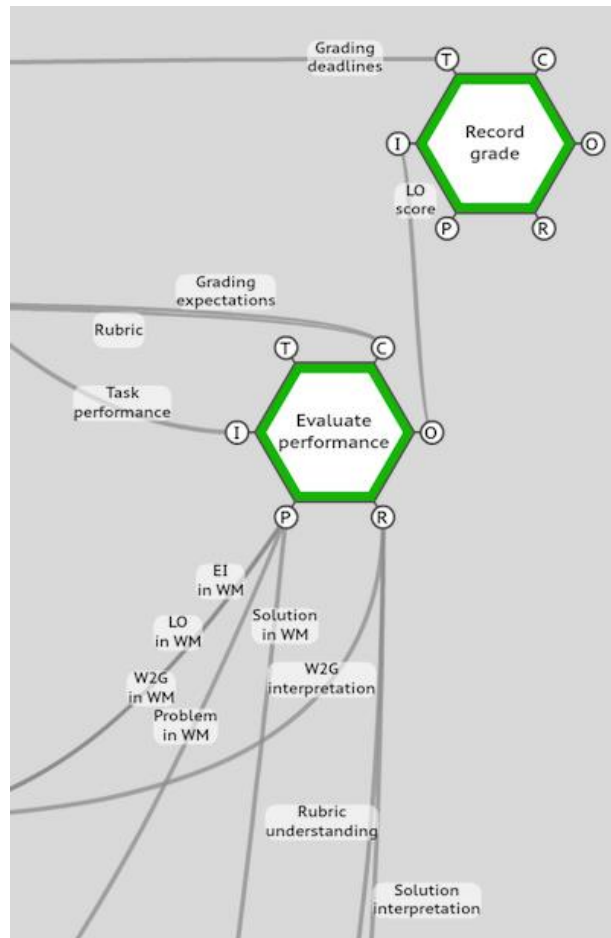


Figure 4.15. Visualization of evaluation and scoring functions.

Table 4.36. System 1 cognitive grader functions associated with evaluating task performance

Function	Scan for aspect of response	Check for exact match	Check for effective match
Description	Look through the student's task performance to find a specific aspect, either to check for presence or to analyze further	Compare the student's performance of the task exactly to the expected solution	Compare the student's performance of task to a range of acceptable solutions
Input	<ul style="list-style-type: none"> • Task performance • <i>Determination that aspect is present</i> 	<ul style="list-style-type: none"> • Located aspect • <i>Determination that aspect matches exactly</i> 	<ul style="list-style-type: none"> • Interpretation of response • <i>Determination that response effectively matches</i>
Output	<ul style="list-style-type: none"> • <i>Determination that aspect is absent</i> • <i>Located aspect</i> 	<ul style="list-style-type: none"> • <i>Determination that aspect does not match exactly</i> 	<ul style="list-style-type: none"> • <i>Determination that response does not effectively match</i>
Precondition	<ul style="list-style-type: none"> • 'What to grade' in WM 	<ul style="list-style-type: none"> • 'What to grade' in WM • Solution in WM 	<ul style="list-style-type: none"> • 'What to grade' in WM • Solution in WM • Problem in WM
Resource/ E.C.	<ul style="list-style-type: none"> • Solution model interpretation • 'What to grade' interpretation 	<ul style="list-style-type: none"> • Solution model interpretation • 'What to grade' interpretation 	<ul style="list-style-type: none"> • Solution model interpretation • 'What to grade' interpretation
Control	---	---	---
Time	---	---	---

Note: The red, italicized inputs and outputs are mutually exclusive—only one occurs in an instance.

Table 4.37. System 2 cognitive grader functions associated with evaluating task performance

Function	Evaluate meaning of response	Scrutinize response to infer student knowledge	Judge criterion satisfaction
Description	Process the student's performance semantically, structurally, or logically to interpret their work	Scrutinize aspects of the student's performance to ascertain the extent to which the learning objective is demonstrated	Make a judgment to, or not to, award credit based on the outcome of the previous analysis
Input	<ul style="list-style-type: none"> • <i>Task performance</i> • <i>Located aspect</i> • <i>Determination that aspect does not match exactly</i> 	<ul style="list-style-type: none"> • <i>Interpretation uncertainty</i> • <i>Determination that response does not meet expectations</i> 	Determination that... <ul style="list-style-type: none"> • <i>aspect is present</i> • <i>aspect is absent</i> • <i>aspect matches exactly</i> • <i>aspect does not match exactly</i> • <i>response effectively matches</i> • <i>response does not effectively match</i> • <i>response meets expectations</i> • <i>response is acceptable</i> • <i>response is unacceptable</i>
Output	<ul style="list-style-type: none"> • <i>Interpretation of response</i> • <i>Determination that response meets expectations</i> • <i>Determination that response does not meet expectations</i> • <i>Interpretation uncertainty</i> 	<ul style="list-style-type: none"> • <i>Determination that response is acceptable</i> • <i>Determination that response is not acceptable</i> 	<ul style="list-style-type: none"> • <i>Met criterion</i> • <i>Unmet criterion</i>
Precondition	<ul style="list-style-type: none"> • 'What to grade' in WM • Solution in WM 	<ul style="list-style-type: none"> • Solution in WM • LO in WM • EI in WM • Problem in WM 	<ul style="list-style-type: none"> • LO in WM • EI in WM
Resource/ E.C.	<ul style="list-style-type: none"> • Solution model interpretation • 'What to grade' interpretation 	<ul style="list-style-type: none"> • Solution model interpretation • Rubric understanding 	<ul style="list-style-type: none"> • Rubric understanding • Task performance • Solution model interpretation
Control	---	---	<ul style="list-style-type: none"> • Rubric • Grading expectations
Time	---	---	---

Note: The red, italicized inputs and outputs are mutually exclusive—only one occurs in an instance. Underlining indicates sets of inputs and outputs that are expected to align.

The first System 1 evaluation function, which occurred for almost all assigned tasks because of the “what to grade” specification is to ‘Scan for aspect of response.’ That is, the rubric specifies a particular portion of the student’s task performance to grade, so the grader must use their familiarity of their mental solution model to scan through the student’s performance to find the specified portion. This may be easier in some cases than others. For closed-ended work, locating the specified portion may be difficult for incorrect student work because it looks different than might have been expected—part of why anticipating a range of student responses prior to grading is helpful. If the task performance is acceptable for a closed-ended task, location is easy. If the task is very open-ended, locating the specific portion can be harder.

Another reason the functions are more complex during evaluation is that there is no longer a certain output that will vary. Instead, the output itself may vary, depending on the contextual factors. For the ‘Scan’ function, some evidence items state that an aspect must be present in the student’s response to achieve credit. In this case, the purpose of scanning is purely to determine if that aspect is present. However, the appropriate output will depend on each specific student’s performance. If the aspect is absent, the correct output will be a ‘determination that the aspect is absent.’ On the other hand, if the aspect is present, the correct output will be a ‘determination that the aspect is present.’ In some cases, the evidence item requires the grader to perform evaluation beyond simple detection of presence but there is still a specified aspect to evaluate. In those cases, scanning is still necessary, but the expected output, given the presence of the aspect, is the ‘located aspect.’ Due to these differences, the means of function variability are slightly different than with other functions—when the choice is between two outputs under one context, the variability is whether the correct output is produced, while locating the aspect may vary if the wrong aspect is located (see Table 4.38 for summary of variability).

Table 4.38. Potential variability of evaluation and scoring functions

Graders	Generalized function: Evaluate task performance		
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Scan for aspect of response	Accuracy (detection)	<u>Inaccurate:</u> Grader believes the specified aspect is present when it is absent or vice versa, likely leading to an incorrect criterion satisfaction judgment [V↑]	Uncommon
		<u>Accurate:</u> Grader accurately determines the presence of the specified aspect, likely leading to the correct criterion satisfaction judgment [V↓]	Common
	Accuracy (location)	<u>Incorrect aspect:</u> The wrong aspect is located; evaluation will be based on the wrong portion of the response [V↑]	Uncommon
		<u>Correct aspect:</u> The correct aspect is located; evaluation will be based on the correct portion of the response [V↓]	Common
Check for exact match	Accuracy (determination)	<u>Inaccurate:</u> Grader believes there is an exact match when there is not or vice versa, likely leading to an incorrect criterion satisfaction judgment [V↑]	Uncommon
		<u>Accurate:</u> Grader accurately determines if the specified aspect matches exactly, likely leading to the correct criterion satisfaction judgment [V↓]	Common
	Accuracy (of output)	<u>Inaccurate:</u> Grader interprets response as meeting expectations when it does not or vice versa, likely leading to an incorrect criterion satisfaction judgment [V↑]	Uncommon
		<u>Accurate:</u> Grader interprets response's meeting of expectations or the need to pursue deeper analysis appropriately, likely leading to the correct criterion satisfaction judgment [V↓]	Common
Evaluate meaning of response	Accuracy (of interpretation)	<u>Incorrect interpretation:</u> Grader misinterprets the student's response; determination of effective matching is less likely to be correct [V↑]	Uncommon
		<u>Correct interpretation:</u> Grader interprets the student's response accurately; determination of effective matching is more likely to be correct [V↓]	Common
	Accuracy (determination)	<u>Inaccurate:</u> Grader believes there is an effective match when there is not or vice versa, likely leading to an incorrect criterion satisfaction judgment [V↑]	Uncommon
		<u>Accurate:</u> Grader accurately determines if the specified aspect effectively matches, likely leading to the correct criterion satisfaction judgment [V↓]	Common
Scrutinize response to infer student knowledge	Alignment (determination)	<u>Misaligned:</u> Grader's determination of acceptability would not align that the determination made by a definitive marker, likely leading to an incorrect criterion satisfaction judgment [V↑]	Uncommon
		<u>Aligned:</u> Grader's determination of acceptability aligns with that of a definitive marker, likely leading to the correct criterion satisfaction judgment [V↓]	Common
Judge criterion satisfaction	Alignment (judgment)	<u>Misaligned:</u> Grader judges a criterion of being satisfied despite negative analysis or vice versa; overall LO score may vary [V↑]	Uncommon
		<u>Aligned:</u> Grader's judgment of criterion satisfaction aligns with their analysis [V↓]	Common

The next System 1 function that may occur, depending on the context of the evidence item, will occur once the aspect to be graded is located. If the evidence item is looking to see if the student's response is a specified value, the grader will 'Check for exact match' (see Table 4.36). That is, does the student's performance of the task look exactly like the solution or expected performance? In some cases, an exact match can be expected, and a very simple verification of a match can occur. As a result, there are two possible outcomes of this function (see Table 4.38). If the student's response matches, then an affirmative 'determination that aspect matches exactly' should be the function output. If the response does not match, the output should be the negative 'determination that aspect does not match exactly' should be the function output. Once again, the variance of this function is whether or not the correct determination, or function output, is made.

Sometimes an exact match cannot be expected, either because of the open-endedness of complexity of the task. Evaluation of the evidence item may still inspire the grader to check to see if the student's response effectively matches their solution model. For example, a student's solution may use different variable names or may construct a statement with relational or logical operators differently, but the student's work may still do what is expected or desired. Before the grader can decide on a match, they must carefully read through and think about the student's response to understand what the student did. Thus, the next function is actually the first System 2 function, 'Evaluate meaning of response' (see the first function in Table 4.37).

There are three possible inputs for the 'Evaluate meaning' function. Sometimes the evidence item demands a deeper analysis of a response from the start. In such a case, either the located aspect of the response is evaluated (i.e., an input of the 'located aspect') or the entire response should be evaluated (i.e., an input of the 'task performance'). In other cases, the grader may be looking for a match but noticed while checking for an exact match that the student's response did not match exactly (i.e., an input of a 'determination the aspect does not match exactly'). Comparing the student's performance to their model of an acceptable solution, the function may lead to one of four outputs. If the goal was to ultimately determine if the portion effectively matched, the output is the grader's interpretation of the student's response. If the evidence item required a deeper evaluation from the start, the output will be either that the response meets expectations or does not. Lastly, if the grader's attempt to understand the student's work at a semantic, structural, or logical level does not lead to a clear understanding, the function will output interpretation uncertainty. Once again, function variability relates to the appropriate output

being produced; however, if the ‘interpretation of response’ is the output, it may also vary with respect to the accuracy of that interpretation (see Table 4.38).

If the evidence item was such that matching was an appropriate strategy, but openness of complexity forced the use of the ‘Evaluate meaning’ function, the next is to ‘Check for effective match’ (see the last function in Table 4.36). This requires the interpretation of the response output by the ‘Evaluate meaning’ function because the response will be inherently different from the solution model and will require a degree of analysis to determine if the effect of the student’s performance is equivalent. The grader then uses their interpretation of the student’s performance to determine if the interpreted work does or does not effectively match the solution model. That is, whether or not the differences between the student’s response and the solution model superficial. The only potential variability of this output is whether or not it produces the correct output (see Table 4.38).

If the ‘Evaluate meaning’ function was performed but produced either the ‘interpretation uncertainty’ or ‘determination that response does not meet expectations’ outputs occur, the grader should next engage in the ‘Scrutinize response to infer student knowledge’ function (see Table 4.37). This function is used when the student’s task performance is difficult to understand or is imperfect but not immediately obviously incorrect. This function involves a deeper exploration of the task performance and may involve some inference about what a student may have been intending (for instance, when an English-language learning student writes a text-based response, their weak language skills may force the grader to infer their intended meaning). There is considerable room for internal variability of this function because the act of judgment has inherent subjectivity. One grader may be more willing to give a student the benefit of the doubt than others. This function requires the deepest level of processing and the grader needs to consider the specific evidence item, the overall learning objective, the context and directions of the problem, the solution model, and possibly even knowledge about the student (although, avoiding bias if this information is needed). If the demonstration of the evidence item is partial in some way, the grader ultimately has to make the judgment of whether or not the response is acceptable. Like the other evaluation functions, the variability is based on the alignment of this decision with respect to the decision that would be made by the definitive marker (see Table 4.38).

‘Judge criterion satisfaction’ (summarized in Table 4.37) is the last evaluation function that should be utilized and takes as input the output of the last function used before it. This assumes

that the correct decision was made by the grader as to the furthest extent they needed to evaluate response. For instance, the grader may have interpreted the evidence item to believe they only needed to check for the presence of something. This would result in an output of ‘Determination that aspect is present/absent,’ which could be an appropriate penultimate evaluation function if the evidence item dictates as such; however, the grader may have misinterpreted the evidence item. As such, there are many different inputs that this function could take, but if the input is negative (i.e., not present, no match, not acceptable) the output should be that the criterion is unmet while a positive input should result in an output that the criterion is met. The output could vary in terms of whether the ultimate judgment of criterion achievement is aligned with that of a definitive marker (see Table 4.38). Additionally, while, at the individual level, judgment may vary in terms of alignment, collections of graders’ judgments can vary in terms of consistency or agreement.

Cognitive functions associated with scoring decisions

There are four cognitive functions associated with scoring decisions. In practice, the preparation for evaluation functions might occur at the start of grading, with occasional review when needed, and the evaluation functions should occur, as contextually appropriate, for each evidence item. Once the judgment of criterion satisfaction is determined for a given evidence item, that should trigger the first of the scoring decision functions (shown in Table 4.39). The second and third functions should then only occur once all of the evidence items have been evaluated and a final grade can be assigned and recorded.

Table 4.39. Cognitive grader functions associated with scoring decisions

Function	Document criterion satisfaction	Aggregate criteria	Decide overall score	Record grade
Description	Document the achievement of each criterion	Add up the number of unmet evidence items	Decide on the overall score based on analysis of student performance of task	Document overall grade by clicking the correct button in the grading software
Input	<ul style="list-style-type: none"> • <i>Met criterion</i> • <u><i>Unmet criterion</i></u> 	<ul style="list-style-type: none"> • Documentation of unmet criteria 	<ul style="list-style-type: none"> • Set of unmet criteria 	<ul style="list-style-type: none"> • <i>LO score</i> • <i>Modified LO score</i>
Output	<ul style="list-style-type: none"> • <i>Documentation of met criterion</i> • <u><i>Documentation of unmet criterion</i></u> 	<ul style="list-style-type: none"> • Set of unmet criteria 	<ul style="list-style-type: none"> • LO Score 	---
Precondition	---	---	---	---
Resource/ E.C.	---	---	---	---
Control	---	<ul style="list-style-type: none"> • Grading expectations 	<ul style="list-style-type: none"> • Rubric 	---
Time	---	---	---	<ul style="list-style-type: none"> • Grading deadlines

Note: The red, italicized inputs and outputs are mutually exclusive—only one occurs in an instance. Underlining indicates sets of inputs and outputs that are expected to align.

The first scoring decision function is to ‘Document criterion satisfaction.’ That is, this function makes note of whether or not the output of the last evaluation function is that the criterion was met or unmet. The documentation should align with the determination. However, this function was observed in interviews to vary in the way it was conducted. Some graders will physical write a “check” or an “x” next to each evidence item when they made a judgment while others will simply make a mental note. Thus, the output can vary in terms of whether it is documented physically or mentally, as well as whether or not an error occurred in documentation (e.g., the grader thought the criterion was met, but then marked an “x”) (see Table 4.40).

Table 4.40. Potential variability of evaluation and scoring functions

Graders	Generalized function: Record score		
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood
Document criterion satisfaction	Format	<u>Mental</u> : Grader makes mental note of evidence item achievement; more likely to commit an error when recalling details to determine the overall LO score [V↑]	Likely
		<u>Physical</u> : Grader physically documents the achievement of each individual evidence item; less likely to commit an error when determining overall LO score [V↓]	Likely
	Accuracy (documentation)	<u>Inaccurate</u> : Grader's documentation of evidence item achievement conflicts with their judgment, leading to likely error in overall LO score [V↑]	Unlikely
		<u>Accurate</u> : Grader's documentation of evidence item achievement is the same as their judgment [V↓]	Typical
Aggregate criteria	Accuracy	<u>Inaccurate</u> : Grader incorrectly counts the number of achieved evidence items, likely leading to error in overall LO score [V↑]	Possible, likely
		<u>Accurate</u> : Grader correctly counts the number of achieved evidence items [V↓]	Possible
Decide overall score	Severity	<u>Harsh</u> : Grader assigns a lower LO score than would have been assigned by a definitive marker [V↑]	Possible, likely
		<u>Moderate</u> : Grader assigns the same LO score as would have been assigned by a definitive marker [V↓]	Possible
		<u>Lenient</u> : Grader assigns a higher LO score than would have been assigned by a definitive grader [V↑]	Possible, likely

Once all of the evidence items have been evaluated, the grader needs to decide an overall LO score. To do that, they first need to determine the total number of evidence items that were not achieved (as the rubrics are written based on number of evidence items that were not achieved, which can be seen in the sample rubric in §3.2.3). This results in the set, or number, or criteria that were not met. The grader may accurately or inaccurately count the number of unmet criteria (see Table 4.40).

After the grader has determined the number of unmet evidence items, they can look at the performance levels indicated on the rubric to ‘Decide the overall score’ for the learning objective. The overall score can vary in severity relative to the definitive mark, such that their score is too harsh or too lenient (see Table 4.40). This can be the result of any errors that may have occurred throughout the evaluation functions, an error in the determination of the number of unmet evidence items, or an error in reading the performance levels of the rubric. It could also be the result of an intentional decision to overrule the rubric, as will be discussed in the next section. Note that the

extent to which a scoring decision is harsh or lenient can vary in magnitude (for instance, if the score should have been ‘Insufficient Evidence’ and was given a ‘Proficient’ versus given a ‘Developing’). Further, like the individual criterion judgments, when pooled with other graders, the overall score decisions can vary in terms of overall consistency.

The final function in the system is the ‘Record grade’ function. This is taking the grade that was decided upon in the previous function and clicking the button in the software to document the overall score. Note that Table 4.39 indicates that the input to the function is either ‘LO score’ or ‘Modified LO score.’ This is because there are additional functions that were observed to occur sporadically throughout the interviews, one of which modifies the score. These other functions are discussed in the next section. Also, there is no output to this function as it is the final function in the system and the grade has already been determined. That said, graders do occasionally accidentally skip recording an LO score in the software or click the button for the incorrect score. Fortunately, this can be fixed relatively easily when a student brings it to a TA’s attention, particularly in the case of a score not being selected at all.

Additional cognitive functions observed during grading

The last sets of functions do not occur with any sort of regularity and cannot be predicted in any exact sense. These functions were directly observed in the interviews or were inferred based on behavior in the interviews. That said, these functions all align with Sharit’s (2006) model of human erroneous actions related to human fallibility and attempts to employ barriers to minimize error; as a result, while specific patterns of occurrence may be difficult to anticipate, they can all be expected to occur from time to time based on the nature of human action. Tracking the occurrence of these functions may also indicate when the system is overburdening the graders.

The first two functions related to the limitations of attention and working memory and the demands of the grading task. As the grading task requires the grader to maintain many pieces of information (i.e., details about the problem being graded, the specific portion of the response that is being graded, a general sense of what constitutes an acceptable response, the overall learning objective that is being evaluated, and the individual evidence items that constitute the learning objective), it is natural for graders to occasionally ‘Forget [item]’s of information they need to properly evaluate student work. While content expertise and experience with grading a given learning objective allow for these details to be chunked more cohesively in working memory, it

can be expected that details will be forgotten, particularly as the grader is just beginning a grading session. As the first function in Table 4.41 shows, the function does not have a visible initiator. However, in line with Schön's (1983) description of a reflective practitioner, the grader likely encounters something unexpected that triggers recognition of a lack the information. The forgetting function is versatile, acting on any item the grader may need. Forgetting most likely varies from grader to grader as a result of personal differences, such as experience, background knowledge, and age, but could vary for an individual grader based on factors like fatigue.

Table 4.41. Cognitive grader functions associated with working memory

Function	Forget [item]	Bring [item] to working memory
Description	Recognize a need to refer to a document in order to refresh details	Revisit a document to or item within a document to bring it back within working memory
Input	---	<ul style="list-style-type: none"> • Gap of [item] in WM
Output	<ul style="list-style-type: none"> • Gap of [item] in WM 	<ul style="list-style-type: none"> • [Item] in WM
Precondition	---	---
Resource/ E.C.	---	<ul style="list-style-type: none"> • [Item's source]
Control	---	<ul style="list-style-type: none"> • Grading expectations
Time	---	---

Note: Brackets indicate that different objects within the system (e.g., learning objectives, evidence items, solution models) can be used interchangeably within the function.

Table 4.41 also shows the function that should follow forgetting a piece of information: 'Bringing [items] to working memory.' In this study, when a grader lost important information from their working memory, it was typically inferred based on the observation of the follow-up information retrieval task; however, it should be noted that the greatest potential for variability is the grader losing pertinent information from working memory but failing to recognize it. If this error were to occur, it could propagate to all decisions made with respect to that missing or degraded information for one or multiple students. This may explain why some graders were observed to look at an incorrect response and make the conscious decision that the response was acceptable. In the event that the grader does recognize their faulty memory, graders who were guided by the expectation of producing an accurate grade were observed to go back to the document where the needed information was originally presented to refresh that information (that is, to transform the "gap of [item in WM]" into the item being in the WM).

The next set of functions all relate to developing an understanding of meaning. Like the previously discussed functions, the functions shown in Table 4.42 can each apply to any of the

pieces of information the graders need to know (i.e., the learning objective, the evidence items, what they should be grading, the context of the problem, or what acceptable answers should look like). The first function, ‘Question meaning,’ can be observed by any utterance from the grader expressing confusion about how they should interpret a piece of information. There is variability from grader to grader in the likelihood of this function, as background knowledge and experience may allow them to more easily interpret the information quickly while uncertainty or self-doubt may cause some graders to be more prone to question their understanding. Still, if they are following the grading expectation of making their best effort to grade accurately, they should read the information critically and monitor their understanding. On the other hand, the output of the function could vary as to whether the grader’s confusion stems from inability to understand versus potential disagreement with the perceived intent of the item.

Table 4.42. Cognitive grader functions associated with the interpretation process

Function	Question meaning	Translate to support understanding	Overrule interpretation
Description	Express confusion, implicitly or explicitly, about the meaning or purpose of an item	Assist the development of understanding of the problem, LO, EI, W2G, or solution by translating into own words	Intentionally imposing own perspective about the meaning of LO, EI, W2G, or solution over what was intended
Input	• [Item] interpretation	• Confusion about [item]	• Confusion about [item]
Output	• Confusion about [item]	• Revised [item] interpretation	• Revised [item] interpretation
Precondition	---	• [Item] interpretation	• [Item] interpretation
Resource/ E.C.	---	• [Item’s source]	• [Item’s source]
Control	• Grading expectations	• Grading expectations	---
Time	---	---	---

Note: Brackets indicate that different objects within the system (e.g., learning objectives, evidence items, solution models) can be used interchangeably within the function.

The second and third functions presented in Table 4.42 are both responses to the confusion produced by the questioning meaning function but tend to have opposite effects. The first function, ‘Translate to support understanding,’ is a strategy to facilitate understanding that likely dampens variability of interpretation when employed. The variable aspect of this function is whether or not the grader has sufficiently translated the concept into their own language to accurately grasp the intended meaning before settling, which is likely a function of the grader’s background knowledge, experience, and ability to self-regulate learning. On the other hand, the ‘Overrule interpretation’

function occurs when a grader believes they have properly interpreted an item but disagree with their perception of the intent and actively choose to dissent. Such a reaction is likely the result of perceived injustice or unfairness or could be an emotional response (Forsythe et al., 2015; Lerner et al., 2015). As such, personal and circumstantial differences could drive variability but in either event, variability will increase in the system when the function is employed. Still, even if a disagreement is identified, likelihood of actively defying grading expectations is highly dependent on personal tendencies of the grader (though defiant behavior like this was noted by Sharit (2006)). Table 4.43 summarizes expected variability of each of the previously discussed functions.

Table 4.43. Potential variability of extra evaluation functions

Graders		Generalized function: Not applicable – can occur at multiple times	
Cognitive function	Form of variability	Possible effects on downstream functions	Likelihood
Forget [item]	Internal: likelihood of occurrence	Should trigger subsequent function [V↑]	Possible, likely
Bring [item] to working memory	Internal: likelihood of occurrence	Lack of occurrence when needed: Missing information in subsequent functions [V↑]	May depend on nature of the task
Question meaning	Internal: likelihood of occurrence	Should trigger a subsequent function [V↑]	Possible, likely
	Output: root cause (understanding vs. disagreement)	Root cause of output should affect which function is triggered	May depend on nature of the task
Translate to support understanding	Output: precision	Imprecise: retain limited understanding [V↑]	Possible, likely
		Acceptable: understanding is improved [V↓]	Typical
Overrule interpretation	Internal: likelihood of occurrence	Incorrect interpretation [V↑]	Possible, unlikely

The last two functions, shown in Table 4.44, are both behaviors that were observed during the interviews that help to dampen system variability for graders following the expectation of producing accurate grades. The first function, ‘Reassure self about score,’ seems to be the grader double checking to make absolutely sure the score they assigned aligns with the score they believe fits the student’s performance. As such, the function can either result in the grader feeling more confident about their scoring decision or identifying concern with their decision. If they determine that they are confident with their decision, the instantiation is complete. On the other hand, if they are concerned with the score, they will perform the ‘Modify score’ function, which may encompass

repeating several of the earlier evaluation functions. If the grader detects an error while reviewing the previous functions, they will modify the score and feel more confident about their decision. In both cases, the double-checking behavior of these functions both serve as barriers to prevent output errors from exiting the system. The impact on the variability of the system is summarized in Table 4.45.

Table 4.44. Cognitive grader functions associated with doubt

Function	Reassure self about score	Modify score
Description	Reassure ones' self about the accuracy of the score assigned	Change an original score after identifying issue with original score
Input	<ul style="list-style-type: none"> • LO score 	<ul style="list-style-type: none"> • Concern with score
Output	<ul style="list-style-type: none"> • <i>Confidence in score</i> • <i>Concern with score</i> 	<ul style="list-style-type: none"> • Modified LO score • Confidence in score
Precondition	---	---
Resource/ E.C.	<ul style="list-style-type: none"> • Task performance • Solution model interpretation 	<ul style="list-style-type: none"> • LO score • Task performance • Solution model interpretation
Control	<ul style="list-style-type: none"> • Rubric • Grading expectations 	<ul style="list-style-type: none"> • Rubric • Grading expectations
Time	---	---

Note: The red, italicized inputs and outputs are mutually exclusive—only one occurs in an instance.

Table 4.45. Potential variability of extra evaluation functions

Graders	Generalized function: Not applicable – can occur at multiple times		
Cognitive function	Form of output variability	Possible effects on downstream functions	Likelihood of variability
Reassure self about score	Internal: likelihood of occurrence	Tendency to occur dependent on grader [V↓]	Likely
	Output: accuracy	Inaccurate: come to the wrong conclusions about accuracy of score [V↑]	Possible, unlikely
		Accurate: come to the right conclusion about accuracy of score [V↔ or V↓]	Typical
Modify score	Output: precision	Imprecise: grading error retained [V↔] Acceptable: grading error removed [V↓]	Possible, likely Typical

5. MODEL INSTANTIATIONS

Chapter 4 addressed Stage 1 of this study by providing a lengthy description of the overall grading system model and indicated how the model could potentially vary, disregarding specific examples of that variance. Chapter 4 also presented the “work-as-imagined” instantiation of the set of static functions (that is, keeping the evaluation function abstracted). This chapter addresses Stage 2 of this study by exploring observed variations in the background functions, and how that affects expected and observed instantiations. Thus, this chapter is devoted to presenting the results of Stage 2. In doing so, it will address, at least partially, the following abbreviated research questions:

RQ 2: How do model instantiations vary?

RQ 2a: How does context affect the work-as-imagined instantiations of the system?

RQ 2b: How do work-as-completed and work-as-imagined instantiations differ?

RQ 2c: Which contextual factors contribute most to observed variability?

RQ 2d: How resilient is the system outcome to internal variability?

This chapter will present results associated with these questions and the next chapter will tie some of the results in with the literature to more fully answer the questions.

Some of the background functions, such as the schedule setting, teaching, or learning functions, were not observed through the collected data and, while they certainly impact the system, their outcomes and impact on the system cannot be ascertained in the scope of this research. Other background functions, such as designing the learning objectives, evidence items, assignments, grading materials, and training materials, were not directly observed, but their outputs were used in the study and can therefore be described based on their variability with respect to one another. The foreground grading functions were directly observed through the think-aloud interviews.

This chapter is organized to align with the organization of Chapter 4. First, the IST’s functions at the deepest level of abstraction are analyzed, describing the ways the outputs are observed to vary and some of the downstream observations. This is followed by a brief discussion of the teaching team and students’ functions and the observable variance. Next, the graders’ use of functions and the variabilities of their outputs are presented with respect to the ideal “work-as-imagined” model instantiations. Finally, the system will be reviewed at a higher level to explore, broadly, how variability aggregates throughout the system.

5.1 Guidance for Interpreting Findings

There is one extremely important caveat for reading this chapter. The way function outputs are presented as varying within this chapter are described *relative to each other*. That is, if a particular output is referred to as unclear, it is unclear *relative to* the other outputs of the same function. It *does not* mean that the output in question is objectively unclear in any absolute sense. All of these materials were designed by highly knowledgeable and experienced educators and educational support staff. Thus, even aspects that are described with terms that may be considered disparaging (e.g., “weak,” or “insufficient”) are classified as such only with respect to the other elements within these self-contained sets. They may still be of considerably higher quality than materials designed by less experienced educators and should not be viewed negatively.

It should also be noted that the sections in this chapter present tables of the different dimensions in which outputs can vary, based on the function descriptions in Chapter 4. These tables report where each observed output fell within the dimensions’ variability spectra. When possible, more objective quantitative measures were used and are reported to communicate positionality on the spectra. However, a qualitative placement on each spectrum is also included to assist interpretability. The tables also indicate how the model suggests the placement on the spectrum should influence downstream functions. The text supporting each table discusses whether or not those variabilities were supported through observations of the system in the think-aloud interviews. That said, it is recognized that the ecological validity of the data collection was imperfect, so observation or lack of observation of the instance does not guarantee its presence or absence in practice.

Before any analysis can begin, it will be important to have an understanding of the assignment that was used to gather all of the data for this study. The assignment consisted of three problems. These problems are described briefly in Section 5.2 before exploring each of the functions observed in relation to the assignment in Sections 5.3 through 5.6. Also note that specific aspects of the rubrics are discussed throughout this chapter, so all of the rubrics are included in Appendix G.

5.2 Problem Descriptions

The first problem on the assignment initially taught students about using a quantity called the Reynolds number to predict the way fluid will flow through pipes, based on density, velocity, pipe diameter, and fluid viscosity. The Reynolds number can be calculated easily from these four variables. Students were given cutoff values below which flow is laminar, between which flow is transitional, and above which flow is turbulent. The students were also given a table of minimum and maximum valid values for each of the four variables. Finally, the students were given a flowchart describing code that would compute the Reynolds number and classify flow type for valid inputs or produce an error for invalid inputs. The students were asked to create and communicate a set of test cases and to convert the provided flowchart into MATLAB code. The students' performances of each of these tasks were evaluated using three learning objectives.

The second problem on the assignment taught students about five atmospheric layers and how each layer's temperature profile varies with altitude. The students were provided the US Standard Atmosphere 1976 model to predict the temperature at altitudes ranging from 0 to 86 kilometers above sea level. The model requires different constants based on the atmospheric layer. The students were provided a table of with the corresponding constants and the altitude ranges for each atmospheric layer. The students were then asked to create a flowchart to plan a function that would identify the atmospheric layer and temperature at any valid altitude, to create and communicate a set of test cases to evaluate their flowchart, and to convert their flowchart to MATLAB code. The students' performances of these tasks were evaluated using five learning objectives, two of which were also used to evaluate performance of problem 1.

The last problem on the assignment revisited a problem completed in a previous assignment where the students had to create a user-defined function to determine the acceptability of a contact lens design based on a set of parameters. In the previous assignment, the students were given an access-restricted function into which they could put a set of input parameters to output a logical decision. They were asked to call the access-restricted function within their own executive function. In the new problem, the students were asked to create their own version of the access-restricted function, given that they were unable to see the actual code within the access-restricted version. The students' functions and their corresponding executive functions were evaluated based on two learning objectives.

5.3 IST Functions

There are two sets of IST cognitive functions that will not be explored here: the setting course schedules functions and the designing lesson plans functions. There are a few reasons for this. First, specific documents related to the course schedules and the lesson plans were not collected through the course of this study. As such, there are no artifacts upon which to base any inferences about the potential outputs of the corresponding functions or to assume any associated impacts on the system. Second, for the course schedules, it is assumed that, while tight deadlines would have a considerable impact on functions throughout the system, for a given semester, the deadlines will affect all sections and all students in the same way. Thus, any understanding of the effects of scheduling functions would likely require a comparison across several semesters. Third, for the lesson plans, the autonomy to implement content according to the professional judgments of the instructors ensures that variability likely exists across all sections in a way that could not reasonably be observed through this study. Again, it is expected that those differences in instruction very possibly lead to differences in student learning, the autonomy afforded in the downstream teaching team functions renders moot any analysis of lesson planning documents in this study.

5.3.1 Content creation

There were 10 learning objectives (LOs) evaluated across the entire assignment used in the think-aloud interviews. Based on the model presented in Chapter 4, the LOs produced by the LO articulation process can vary in terms of their clarity and their breadth. In order to minimize subjectivity of comparison, these two dimensions were operationalized. The clarity of each learning objective was scaled based on an average standardized score of a wide range of lexical, semantic, and structural measures using textual analysis software (see §3.4.3). As the LOs are operationalized through a set of observable evidence items (EIs), the breadth of each LO was measured by the number of EIs the LO spans.

Most LOs fell within the middle ranges of the measures for clarity and breadth (see Table 5.1). The clarity measure was determined by averaging and scaling the outputs of a web-based lexical complexity analyzer (Ai & Lu, 2010; Lu, 2012) and a syntactic complexity analyzer (Ai & Lu, 2013; Lu, 2010; Lu, 2011; Lu & Ai, 2015). Using this measure, one LO was significantly

clearer than the others and three were relatively unclear. Based on the measure, the clearest LO was, “Code a selection structure” while the least clear was, “Create a user-defined function that adheres to programming standards.” Meanwhile, “Convert between these selection structure representations: English, a flowchart, and code,” and “Construct a flowchart for a selection structure using standard symbols and pseudocode,” were also relatively unclear. These “unclear” phrases use complex and loaded words that might be challenging for students with weaker grasps of the English language. However, while it is expected that the LO that is being evaluated be at the forefront of a grader’s mind while evaluating task performance, the observed instances suggest graders pay little attention to the LO text, itself. As such, this measure likely makes little difference to downstream functions.

Table 5.1. Observed variabilities of content creation function outputs

IST	Generalized function	Create content	
Cognitive function	Articulate LOs		Articulate EIs
Output variability	Clarity (scale score)	Breadth (#EIs)	Coverage
LO 1 (Problem 1)	Adequate (1.93) [V↔]	Moderate (4) [V↔]	Sufficient [V↔]
LO 2 (Problem 1)	Adequate (1.62) [V↔]	Narrow (2) [V↓]	Sufficient [V↔]
LO 3 (Problem 1)	Unclear (0.33) [V↑]	Moderate (6) [V↔]	Insufficient [V↑]
LO 4 (Problem 2)	Adequate (1.25) [V↔]	Broad (11) [V↑]	Sufficient [V↔]
LO 5 (Problem 2)	Unclear (0.71) [V↑]	Moderate (6) [V↔]	Mostly [V↔]
LO 6 (Problem 2)	Adequate (1.93) [V↔]	Moderate (4) [V↔]	Sufficient [V↔]
LO 7 (Problem 2)	Adequate (1.62) [V↔]	Narrow (2) [V↓]	Sufficient [V↔]
LO 8 (Problem 2)	Clear (3.00) [V↓]	Broad (10) [V↑]	Sufficient [V↔]
LO 9 (Problem 3)	Adequate (1.11) [V↔]	Moderate (5) [V↔]	Sufficient [V↔]
LO 10 (Problem 3)	Unclear (0.00) [V↑]	Broad (10) [V↑]	Mostly [V↔]

The ten LOs had considerable differences in breadth (see Table 5.1). The narrowest LO spanned only two EIs (LO 2) while the broadest spanned 11 (LO 4). Five of the observed LOs, two of which were duplicates of two others, fell into an intermediate range of four to six EIs. Meanwhile, three had a large number of EIs, spanning 10 or 11. As will be shown later, the propensity of graders to approach grading more holistically leads to a greater likelihood of disregarded EIs as the number of EIs increases.

Almost contrasting breadth is the coverage of each LO based on the EIs that are identified within. The coverage represents how fully the LO is described by the EIs. As shown in Table 5.1, the majority of the LOs are well represented by the EIs—that is, deep thought about the LO did

not lead to the identification of behaviors the students should demonstrate in order to be considered proficient at the LO (see Tables 5.2 through 5.4 for all articulated EIs). However, two (LOs 5 and 10) seem to be missing a small component and one (LO 3) is missing some bigger pieces. LO 5 relates to coding selection structures but does not contain an EI about the structure including all necessary paths. LO 10 could or should reasonably include EIs related to formatting and suppression of output. More significantly, LO 3 has EIs that relate directly to the conversion of a flowchart to code; however, the LO text suggests conversion between English, flowcharts, and code, leaving a gap in all other forms of conversion. While these are ways the EIs could be improved, there were no instances observed in the interviews that would suggest these issues had any bearing on the results of grading.

Table 5.2. Observed variability of evidence item articulation for problem 1 learning objectives

Learning Objective	Evidence Item	Alignment	Clarity (scale score)	Precision	Complexity*
LO 1 (Prob. 1) Create test cases to evaluate a flowchart	1) Creates thorough set of test cases to test all possible outcomes in the flowchart	Aligned [V↓]	Moderate (1.91) [V↔]	Adequate [V↔]	System 1 [V↓]
	2) Use English to describe each test and how the information moves through the flowchart for that test	Partial [V↔]	Moderate (1.43) [V↔]	Imprecise [V↑]	System 2 [V↑]
	3) Lists input arguments in a valid format	Partial [V↔]	Clear (2.96) [V↓]	Imprecise [V↑]	System 1 [V↓]
	4) Test values are consistent with the test description	Aligned [V↓]	Clear (2.74) [V↓]	Precise [V↓]	System 2 [V↑]
LO 2 (Prob. 1) Track a flowchart with a selection structure	1) Identify correct path given the test value(s)	Aligned [V↓]	Clear (2.68) [V↓]	Imprecise [V↑]	System 2 [V↑]
	2) Describe the outcome(s) in English with resulting values when appropriate (not code results)	Partial [V↔]	Moderate (1.78) [V↔]	Imprecise [V↑]	System 1 [V↓]
LO 3 (Prob. 1) Convert between these selection structure representations: English, a flowchart, and code	1) Recognize that a diamond structure with one input arrow and two output arrows (labeled Yes/No or True/False) translates to an if or elseif statement	Aligned [V↓]	Unclear (0.02) [V↑]	Precise [V↓]	System 1 [V↓]
	2) The number of diamonds in the flowchart translates exactly to the number if and elseif statements	Aligned [V↓]	Moderate (1.43) [V↔]	Adequate [V↔]	System 1 [V↓]
	3) Recognize that the first 1-in/2-out diamond in a flowchart (or first following other non-decision instructions or the first on a Yes path following a decision) is an if statement	Aligned [V↓]	Unclear (0.04) [V↑]	Adequate [V↔]	System 1 [V↓]
	4) Recognize that all immediately following 1-in/2-out diamonds on the No or False path are elseif statements	Aligned [V↓]	Moderate (1.24) [V↔]	Adequate [V↔]	System 1 [V↓]
	5) Recognize an else statement is implied if there are operations between the only or last diamond and the convergence of the flowchart connecting lines	Aligned [V↓]	Unclear (0.16) [V↑]	Precise [V↓]	System 1 [V↓]
	6) Recognize that a convergence of the entire No or False path with the entire Yes or True path translates to an end statement	Aligned [V↓]	Unclear (0.53) [V↑]	Precise [V↓]	System 1 [V↓]

Table 5.3. Observed variability of evidence item articulation for problem 2 learning objectives

Learning Objective	Evidence Item	Alignment	Clarity (scale score)	Precision	Complexity*
LO 4 (Prob. 2) Construct a flowchart using standard symbols and pseudocode	1) Flowchart symbols: Start and stop for the overall flowchart are represented by ovals	Aligned [V↓]	Moderate (1.73) [V↔]	Precise [V↓]	System 1 [V↓]
	2) Flowchart symbols: Inputs and outputs are represented by parallelograms	Aligned [V↓]	Clear (2.33) [V↓]	Adequate [V↔]	System 1 [V↓]
	3) Flowchart symbols: Decisions are represented by diamonds	Misaligned * [V↑]	Clear (2.63) [V↓]	Precise [V↓]	System 1 [V↓]
	4) Flowchart symbols: Processes, such as calculations, are represented by rectangles	Aligned [V↓]	Clear (2.13) [V↓]	Adequate [V↔]	System 1 [V↓]
	5) Flowchart symbols: Operations are connected with arrows with points at one end to indicate flow	Aligned [V↓]	Moderate (1.98) [V↔]	Precise [V↓]	System 1 [V↓]
	6) Arrows must connect all flowchart elements and indicate a continuous flow from start to stop	Aligned [V↓]	Moderate (1.28) [V↔]	Imprecise [V↑]	System 2 [V↑]
	7) Arrows must converge prior to stop so that there is only one arrow into the stop	Aligned [V↓]	Moderate (1.13) [V↔]	Precise [V↓]	System 1 [V↓]
	8) Flowchart process ends in one place (cannot end in multiple places)	Aligned [V↓]	Clear (2.15) [V↓]	Imprecise [V↑]	System 1 [V↓]
	9) Text within the symbols is in concise English (not code or only math) that conveys the purpose of the step	Aligned [V↓]	Unclear (0.0) [V↑]	Imprecise [V↑]	System 2 [V↑]
	10) Decisions are accompanied by Yes/No or True/False text on the appropriate arrows	Aligned [V↓]	Moderate (1.73) [V↔]	Imprecise [V↑]	System 1 [V↓]
	11) Flowchart represents all possible outcomes required by the problem	Aligned [V↓]	Clear (2.32) [V↓]	Precise [V↓]	System 2 [V↑]
LO 5 (Prob. 2) Construct a flowchart for a selection structure using standard symbols and pseudocode	1) Decisions that are part of a selection structure are represented with a diamond filled with a condition	Aligned [V↓]	Moderate (1.32) [V↔]	Imprecise [V↑]	System 1 [V↓]
	2) Decisions have one input arrow and two output arrows (one for Yes/True and one for No/False)	Aligned [V↓]	Moderate (1.92) [V↔]	Imprecise [V↑]	System 1 [V↓]
	3) There are operations on the Yes/True path	Aligned [V↓]	Clear (2.60) [V↓]	Adequate [V↔]	System 1 [V↓]
	4) For multiple related selections (i.e., if-elseif-else), there are no operations between the decisions along the No/False path	Aligned [V↓]	Moderate (1.27) [V↔]	Adequate [V↔]	System 1 [V↓]
	5) For multiple related selections (i.e., if-elseif-else), the Yes/True and No/False path arrows converge after all related decisions and (optionally) the operations for the else path	Aligned [V↓]	Unclear (0.76) [V↑]	Precise [V↓]	System 2 [V↑]
	6) Operations are included in the selection structure as required by the problem	Aligned [V↓]	Clear (2.24) [V↓]	Precise [V↓]	System 2 [V↑]

Table 5.3. continued

Learning Objective	Evidence Item	Alignment	Clarity (scale score)	Precision	Complexity*
LO 6 (Prob. 2) is the same as LO 1 (Prob. 1)					
LO 7 (Prob. 2) is the same as LO 2 (Prob. 1)					
LO 8 (Prob. 2) Code a selection structure	1) Begin a selection structure with an if	Aligned [V↓]	Clear (2.77) [V↓]	Precise [V↓]	System 1 [V↓]
	2) The if is accompanied by a condition for which a true result corresponds to code that immediately follows	Aligned [V↓]	Unclear (0.09) [V↑]	Precise [V↓]	System 2 [V↑]
	3) elseif is used for a series of related conditions	Aligned [V↓]	Clear (2.24) [V↓]	Precise [V↓]	System 1 [V↓]
	4) Each elseif is accompanied by a condition which a true result corresponds to code that immediately follows	Aligned [V↓]	Unclear (0.39) [V↑]	Imprecise [V↑]	System 2 [V↑]
	5) elseif is a single word - there is no space between else and if	Aligned [V↓]	Mod. (1.82) [V↔]	Imprecise [V↑]	System 1 [V↓]
	6) An else is used to handle any condition(s) not addressed in the earlier parts of the selection structure and not used if no code is needed before the end	Aligned [V↓]	Unclear (0.08) [V↑]	Precise [V↓]	System 2 [V↑]
	7) An else is not accompanied by a condition	Aligned [V↓]	Clear (2.38) [V↓]	Imprecise [V↑]	System 1 [V↓]
	8) end is used to terminate the selection structure	Aligned [V↓]	Clear (2.43) [V↓]	Precise [V↓]	System 1 [V↓]
	9) Statements between the if, elseif, else, and end are indented	Partial [V↔]	Clear (2.09) [V↓]	Precise [V↓]	System 1 [V↓]
	10) A selection structure addresses all necessary paths for a given problem	Partial [V↔]	Clear (2.22) [V↓]	Precise [V↓]	System 2 [V↑]

Table 5.4. Observed variability of evidence item articulation for problem 3 learning objectives

Learning Objective	Evidence Item	Alignment	Clarity (scale score)	Precision	Complexity*
LO 9 (Prob. 3) Coordinate the passing of information between functions	1) Call to a user-defined function occurs in the proper function or script	Aligned [V↓]	Moderate (1.81) [V↔]	Precise [V↓]	System 1 [V↓]
	2) Variables passed into a user-defined function are defined prior to calling the user-defined function	Partial [V↔]	Moderate (1.69) [V↔]	Precise [V↓]	System 2 [V↑]
	3) Variables passed into a user defined function are defined prior to calling the user-defined function	Partial [V↔]	Moderate (1.69) [V↔]	Imprecise [V↑]	System 2 [V↑]
	4) User-defined functions are called in the order necessary to complete the coding task	Aligned [V↓]	Moderate (1.54) [V↔]	Precise [V↓]	System 2 [V↑]
	5) No use of global variables (to circumvent proper passing of information through function calls)	Partial [V↔]	Moderate (1.55) [V↔]	Precise [V↓]	System 1 [V↓]
LO 10 (Prob. 3) Create a user-defined function that adheres to programming standards	1) Help lines contain input and output argument definitions, with units as appropriate	Aligned [V↓]	Moderate (1.83) [V↔]	Precise [V↓]	System 2 [V↑]
	2) Help lines contain concise description of the program	Aligned [V↓]	Clear (2.51) [V↓]	Imprecise [V↑]	System 2 [V↑]
	3) Help lines show the call to the function	Aligned [V↓]	Clear (3.00) [V↓]	Precise [V↓]	System 1 [V↓]
	4) Complete programmer and contributor information in the header (names and emails)	Aligned [V↓]	Moderate (1.72) [V↔]	Precise [V↓]	System 1 [V↓]
	5) Complete problem details including assignment number, problem number	Partial [V↔]	Moderate (1.97) [V↔]	Precise [V↓]	System 1 [V↓]
	6) Code items are in the correct section	Aligned [V↓]	Moderate (1.91) [V↔]	Precise [V↓]	System 2 [V↑]
	7) Computed values are assigned to variables	Aligned [V↓]	Clear (2.60) [V↓]	Precise [V↓]	System 1 [V↓]
	8) Code blocks have explanatory comments	Aligned [V↓]	Clear (2.70) [V↓]	Imprecise [V↑]	System 2 [V↑]
	9) Variables have commented definitions and units	Aligned [V↓]	Clear (2.59) [V↓]	Precise [V↓]	System 1 [V↓]
	10) Minimal use of hardcoding	Aligned [V↓]	Clear (2.59) [V↓]	Imprecise [V↑]	System 1 [V↓]

Articulation of evidence items can vary in four other ways beyond their overall coverage of the LOs: alignment with the LOs, clarity, precision (i.e., whether they represent a single, distinct construct), and complexity (i.e., whether system 1 processing or system 2 processing is likely needed to evaluate them). Based on subjective analysis, nearly all (45 out of 54) of the EIs are well aligned with their LOs. The EIs that are marked as “partial,” or partially aligned, are mostly the result of being nitpicky. For instance, EI 3 in LO 1 is, “Lists input arguments in a valid format” for the LO “Create test cases to evaluate a flowchart.” Arguably, the format and listing of the arguments is not directly related to the *creation of the test cases*, but perhaps with the creation of a test case table or the communication of test cases. The one EI that is marked as “misaligned” is mostly considered misaligned because it overlaps heavily with another LO. With an LO for constructing a flowchart and an LO for constructing a flowchart of a selection structure, one EI in each LO effectively covers the same construct—either could have been marked as misaligned. A similar, though less significant, overlap happens with EI 1 in LO 2 and EI 4 in LO 1. When these overlaps occur in conjunction with a student failing to demonstrate the particular EI, graders made differing interpretations and expressed annoyance with students being double penalized.

Clarity of EIs was evaluated in the same way as the clarity for the LOs, except scaled across the 54 distinct EIs rather than the eight distinct LOs. Of the 54 EIs, 22 were relatively clear, nine were unclear, and 23 fell in between. The unclear EIs were generally more structurally complex statements with several clauses and some more advanced words. Compare, for instance, the relatively clear, “Help lines show the call to the function,” versus the relatively unclear, “An else is used to handle any condition(s) not addressed in the earlier parts of the selection structure and not used if no code is needed before the end.” This difference in clarity is not insignificant. Fifty-eight percent of instances where graders expressed confusion with an EI occurred with the unclear EIs and 60% of instances where graders had to re-read the rubric related to the unclear EIs. Further, the most unclear EI led to evaluation functions producing inappropriate outputs 27.5% of the time compared to 8.3% of the time for the clearest EI.

The determination of “precision” of the EIs was based on whether or not the EIs were unidimensional and clearly distinct from one another. EIs that were multidimensional or overlapped with others were considered either “adequate” or “imprecise,” depending on the severity. It should be noted that in most of these cases, the issue was identified based on attempting to apply the rubric to flawed student work—generally, one would likely not have expected an

issue. An instance of multidimensionality occurs in EI 1 of LO5: “Decisions in a selection structure are represented by diamonds *filled with a condition*.” This requires students to know both that diamonds represent selection structures (which is also EI 3 in LO 4) and that a condition must be included.

The vast majority of the “Imprecise” designations are due to overlaps. For instance, EI 5 in LO 4 is primarily associated with using arrows in a flowchart. However, the wordings of EIs 6, 7, and 10 all include “arrows.” As a result, a literal interpretation of the EIs would require the flowchart to have arrows to achieve any of the EIs, even though the essential components of those EIs are related to features other than the arrows, themselves. This was observed repeatedly in interviews to lead to confusion and overruling in graders. Imprecision can also be a result of using terminology that is not sufficiently descriptive for graders to understand how to interpret. For instance, EI 3 of LO 1 (“Lists input arguments in a valid form”) is clearly written and superficially easy to understand, but “valid form” is too vague for graders to know what is or is not valid. This one EI led to confusion for five of the 17 graders observed.

Finally, the EIs can vary in terms of whether or not they elicit system 1 or system 2 processing to evaluate. Less complex EIs require only system 1 thinking. For example, EI 2 of LO 3 (“The number of diamonds in the flowchart translates exactly to the number if and elseif statements”) translates to the grader as, “check that there are x diamonds in the flowchart.” Evaluation of this EI only requires counting and determining if the two numbers are equal. Compare this to EI 2 of LO 1 (“Use English to describe each test and how the information moves through the flowchart for that test”), which translates to the grader as, “read the description of the test case to make sure it represents an appropriate path through the flowchart.” This inherently requires a greater level of cognitive processing to read, interpret, and evaluate the case with respect to the context and, along with particular features of student responses, is the primary driver of different work-as-imagined instantiations of the evaluation functions. All of the EIs that were within the lowest level of agreement (less than 40%) with the definitive mark required system 2 processing (i.e., using the functions to evaluate meaning or scrutinize work). Overall, graders were 14% less likely ($p < 0.0001$) to agree with the definitive mark for system 2 EIs. It should be noted, however, that while an EI might only indicate the need for system 1 processing, evaluation of any EI can switch to system 2 if the work being evaluated is sufficiently unclear that the grader needs to scrutinize the work.

5.3.2 Assessment task design

Designing an assessment task has four primary outputs: the task itself, the task's context, the task's instructions, and a model solution. These outputs can each vary in several different dimensions. The task itself can vary with respect to its alignment with the learning objectives used to evaluate its performance and its open-endedness. The problem context can vary in terms of its overall understandability to the students (i.e., how familiar the context is, how much prerequisite knowledge it requires, and level of language used to describe it). The instructions can vary with respect to their clarity and the extent to which the instructions break the task down into smaller steps. Finally, the model response can vary by its accuracy and how comprehensively it represents the entire set of acceptable solutions. The variabilities of these functions observed through the course materials used in the think-aloud interviews are summarized in Table 5.5.

Table 5.5. Observed variability of assessment task design function outputs

IST	Generalized function Design assessment task						
Cognitive function	Select appropriate task		Develop task context	Write task instructions		Develop model response	
Output variability	Alignment	Open-endedness	Understand-ability (scale score)	Clarity (scale score)	Scaffold-ing	Accuracy	Comprehen-siveness
LO 1 (Problem 1)	Aligned [V↔]	Moderate [V↔]	Moderate (2.0) [V↔]	Moderate (1.38) [V↔]	Low [V↑]	Acceptable [V↔]	Narrow [V↑]
LO 2 (Problem 1)	Aligned [V↔]	Low [V↓]		Moderate (1.10) [V↔]	Low [V↑]	Accurate [V↓]	Moderate [V↔]
LO 3 (Problem 1)	Aligned [V↔]	Closed [V↓]		Unclear (0.96) [V↑]	High [V↓]	Accurate [V↓]	Broad [V↓]
LO 4 (Problem 2)	Aligned [V↔]	Open [V↑]		Unclear (0.98) [V↑]	Moderate [V↔]	Accurate [V↓]	Narrow [V↑]
LO 5 (Problem 2)	Aligned [V↔]		Moderate- (2.25) [V↓]	Clear (2.74) [V↓]	Low [V↑]	Acceptable [V↔]	Moderate [V↔]
LO 6 (Problem 2)	Aligned [V↔]	Moderate [V↔]	Moderate- (1.75) [V↑]	Clear (3.00) [V↓]	Low [V↑]	Accurate [V↓]	Moderate [V↔]
LO 7 (Problem 2)	Aligned [V↔]	Low [V↓]		Clear (2.44) [V↓]	Moderate [V↔]	Accurate [V↓]	Moderate [V↔]
LO 8 (Problem 2)	Partial [V↑]	High [V↑]		Moderate (1.09) [V↔]	Moderate [V↔]	Accurate [V↓]	Narrow [V↑]
LO 9 (Problem 3)	Aligned [V↔]	Open [V↑]		None [V↑]	No solution [V↑]		
LO 10 (Problem 3)	Aligned [V↔]						

The tasks for the assignment all seemed appropriate. Each task should have reasonably elicited demonstration of the selected learning objectives in their completion. As such, the tasks were all well aligned with the LOs; however, one very minor point of contention would be that because the students were instructed to create a flowchart and translate the flowchart to code, LO 8 (“Code a selection structure”) might have more appropriately been LO 3 (“Convert between selection structure representations...”). That said, it makes sense to try to assess a broader spectrum of LOs in an assignment rather than the same ones repeatedly. Interview observations did not noticeably indicate that this factor had an impact on the system; although, misalignment might have been perceptible.

The tasks were also variably open-ended. On the most closed-ended side, in problem 1, the students were asked to directly translate a provided flowchart into code. The LO associated with this task requires that the translation be exact to achieve credit, meaning there really only was one possible solution. On the other end of the spectrum, in problem 2 the students were asked to create their own flowcharts. Although the elements that needed to be included were clearly laid out in the problem description, there was still a large amount of potential freedom in how the students designed their flowchart. For instance, the exact order or placement of some operations were up to the students’ interpretations: it would have been equally acceptable to calculate the temperature inside the selection structure as it would have been to assign the boundary layer inside the selection structure and perform the calculation outside the selection structure. The more open-ended tasks create greater difficulty for the grader because it requires them to have a broader understanding of what constitutes acceptable responses and limits the extent to which they can be dependent upon the provided solution and rubric or, alternatively, requires a more comprehensive solution or rubric.

The understandability of the context is a combination of the familiarity of the content, the need for prerequisite knowledge, and the clarity of the context description text. The scaled score is an average of these values. While the overall understandability of the three problems are not terribly different, problem 2 is the easiest to understand and problem 3 the most difficult. For problem 2, while the text is the most difficult structurally and semantically, it is likely that students have some familiarity with atmospheric science concepts from the grade school science curriculum and the knowledge and skills necessary to approach the problem are limited to interpreting a simple mathematical expression involving simple variables. Problem 1, on the other hand, has the easiest

textual description to understand, but the students are likely to be unfamiliar with fluid mechanics and problem requires some understanding of physical quantities like velocity and viscosity. Problem 3 is the hardest, with a mid-level of textual complexity, dealing with a likely to be unfamiliar topic of medical devices (i.e., contact lenses) and requiring an understanding of the earlier version of the problem and its solution. The differences in understandability likely affects the quality of student task performances, which was not observed in this study. On the other hand, while it could affect graders' understanding, as well, it was not observed to contribute to differences in grading—the graders paid little direct attention to the problem contexts across the interviews.

The task instructions varied in terms of clarity and the amount the task was explicitly broken down into simpler sub-tasks for the students—that is, its scaffolding. The clarity was determined using the same process as with the LOs and EIs. Two task instructions were relatively unclear (though close to moderate) and three were quite clear. The instructions for translating the flowchart into MATLAB code and for creating the flowchart for the atmospheric layer code were both considered unclear; however, this appears to be influenced by the relatively large amount of scaffolding provided in each, which included more sophisticated language than some of the other instructions, increasing the structural and lexical complexity. The clear instructions were likely to be more easily interpreted by students because they were direct, single line statements such as, “Select a series of test cases to thoroughly test all possible paths on your flowchart,” and “Record the atmospheric layer or error for each test case.” As a result, the amount of scaffolding provided seemed to correlate negatively with the clarity. The directions to translate the flowchart to code in problem 1 were highly detailed and walked the student through each step. On the other hand, test case table instructions were very limited in supporting students through the process. At the extreme, the third problem was evaluated based on adherence to programming standards, but the instructions did not provide any reminders to the students to adhere to these standards (which is okay, given that the students should have had two months of experience with this). Like the context understandability, while these variabilities likely affect the system, their impact on the system was not observable through this study.

The model response's accuracy may vary. All but two of the model responses were completely accurate, which is typically the case; however, an occasional error may slip through, particularly when the problem is adapted from a previous semester by changing numbers. In the

case of the task associated with LO 1, a number was changed in the problem that was not changed in the model response. The same basic issue occurred with the test case table for LO 6, where the test case descriptions should have said “layer 3” and “layer 4,” but instead said “layer 4” and “layer 5.” These errors did not lead to any confusion or incorrect decisions for graders during the interviews but could reasonably have caused a grader to make incorrect judgments in practice if they were not paying close enough attention. One other case to note is that because the final LO was concerned with adherence to programming standard, there was no model response provided. As will be discussed in the grader section, a majority of graders did not pay close enough attention and graded the wrong code for the LO that did not have a corresponding model response. While the lack of model response was not the most likely cause of the common error, if a model response provided may have drawn their attention to the proper details.

In addition to accuracy, model responses can vary in terms of how comprehensively they address a range of acceptable solutions. It can reasonably be expected that as a task becomes increasingly open-ended, it becomes increasingly difficult to capture the full spectrum of possible student responses and the model response would represent a narrower portion of acceptable responses. This is observed in the data, where the two open-ended tasks narrowly represent possible acceptable solutions while the closed-ended task broadly addresses possible solutions. Unfortunately, it does not seem that the variability presented by the samples of student work were sufficient to identify if this would have been an issue for graders in the interviews. With the samples used, the comprehensiveness of the model solution was not observed to directly cause any issues in the grading process.

5.3.3 Grading guideline design

There are four primary ways that grading guidelines can vary that could be expected to affect subsequent functions: the discriminability between performance levels, the degree to which the rubric specifies portions of a student’s response to grade, the robustness of the rubric to handle a wide range of student responses, and the overall usability of the rubric. The observed variabilities of these functions across the rubrics used in the think-aloud interviews are summarized in Table 5.6.

Table 5.6. Observed variabilities of grading guideline design function outputs

IST	Generalized function	Design grading guidelines		
Cognitive function	Differentiate performance levels	Assemble rubric elements		
Output variability	Discriminability	Specificity	Robustness	Usability
LO 1 (Problem 1)	Weak [V↑]	High [V↑]	Strong [V↓]	Moderate [V↔]
LO 2 (Problem 1)	Weak [V↑]	High [V↑]	Adequate [V↔]	High [V↓]
LO 3 (Problem 1)	Strong [V↓]	Moderate [V↔]	Strong [V↓]	Moderate [V↔]
LO 4 (Problem 2)	Acceptable [V↔]	High [V↑]	Strong [V↓]	Low [V↑]
LO 5 (Problem 2)	Strong [V↓]	High [V↑]	Strong [V↓]	Moderate [V↔]
LO 6 (Problem 2)	Weak [V↑]	High [V↑]	Strong [V↓]	Moderate [V↔]
LO 7 (Problem 2)	Weak [V↑]	High [V↑]	Adequate [V↔]	High [V↓]
LO 8 (Problem 2)	Acceptable [V↔]	High [V↑]	Adequate [V↔]	Moderate [V↔]
LO 9 (Problem 3)	Strong [V↓]	High [V↑]	Adequate [V↔]	Moderate [V↔]
LO 10 (Problem 3)	Acceptable [V↔]	Moderate [V↔]	Strong [V↓]	Moderate [V↔]

Some grading guidelines are able to differentiate between levels of performance more effectively than others. In the case of how the rubrics in this study are designed, this primarily refers to how well the discrete overall performance levels reflect differences in achievement of evidence items. Two of the eight unique learning objectives were relatively weak in their discriminability while three were relatively strong. There are two main issues that contributed to weakness in discriminability. For LO 1 (which is also LO 6), the rubric specified two separate test cases to evaluate for the achievement of each evidence item. This became problematic for many graders when a sample response achieved the evidence item for one of the cases and excluded the other case. This caused different graders to make three different grading decisions: (1) no credit given for the evidence items based on a strict interpretation of the rubric; (2) full credit given for the evidence items based on a lax interpretation of the rubric and argument that one demonstration of achievement is sufficient; and (3) half credit given for the evidence item. Some graders made the executive decision to split the difference between the two extreme interpretations because the student's performance did not feel appropriately represented by the rubric.

LO 2 (which is also LO 7) had the same issue with clustering of requirements within an evidence item but added a second challenge to discriminability. The presence of only two evidence items naturally leads to the exclusion of one of the four default performance levels—either both are achieved, one is achieved, or none are achieved. However, when coupled with the clustered requirements for achievement, graders feel compelled to award partial achievement of evidence

items when one of the two requirements is achieved. When there is no intermediate proficiency score between zero and one or between one and two evidence items achieved, the rubric again fails to adequately discriminate between performance levels and produces variable grading decisions. This is similar to, but slightly distinct from, the issue found in the LOs deemed “acceptable” (LOs 4, 8, and 10) where the number of evidence items more than doubles the number of performance levels. For example, in LO 4, achievement of six of the 11 evidence items results in “insufficient evidence,” or 0 points. Graders express concern when a student who has demonstrated no evidence items receives the same credit as a student who has demonstrated more than half (i.e., 6/11 or 55%) of the evidence items. When there are many evidence items, the necessary ranges clumped into discrete performance levels fails to discriminate between various levels of performance fairly.

The rubrics in this course incorporated various specifications throughout to reduce the amount of each student’s work that the grader would have to look at and evaluate and to minimize the need for broader judgments. Besides LO 3, which evaluates an entire flowchart, and LO 10, which evaluates an entire code for adherence to programming standards, all of the other LOs specify small portions of a response to evaluate. These interviews demonstrated, however, that this practice can lead to a number of different variabilities within the grading process: (1) overlooking or misinterpreting the specifying text, (2) emotional response to misrepresentation of performance, and (3) an over-reliance on specification.

The first variability associated with specification is that some graders may not read the rubric carefully enough and may overlook or misinterpret the text that specifies what portion to grade. There are several examples of where this occurred. The rubric for LO 3 specified that graders only grade the selection structure for selecting flow type in code that also used a selection structure to identify errors. Two of the graders did not see the specifying text and graded all of the selection structures or, in the case of the example without the flow type selection structure, only graded the error selection structure, leading to vastly different grades than those who followed the specification (i.e., scoring the response as ‘proficient’ versus ‘no attempt’). LO 9 specified that only the lines of code associated with Lens ID ‘LM 17’ should be evaluated, but the majority of graders overlooked this specification, evaluating all of the cases presented. This led to vastly different grades for the samples that did not include the ‘LM 17’ lens for the graders who read the specification versus those who did not. Lastly, LO 10 specified evaluating the decision function rather than the executive function, but the majority of graders evaluated the wrong function. In this

case, because the LO was associated with programming standards, which students tend to apply in a consistent fashion in all their programs, the differences in grading decisions were not dramatic, but still led to minor variations.

The second variability associated with specification is when the specification results in an overall score that misrepresents the overall quality of the work. One example was in the case of LO 9's specification of just lens 'LM 17.' Two of the sample responses failed to modify the lens IDs from the dataset provided in the previous assignment's version of the problem to the new dataset used for this assignment and, as a result, did not have the right lens ID. A few graders identified that the correct lens ID was missing but chose to overrule the rubric when they recognized that the student had properly demonstrated the EIs and LO. Graders noted that this can occur in both directions. For instance, while both of these cases did not occur in the interview samples, considering the test case table for LO 1, a student could theoretically be awarded three of the four EIs if they only had the specified two of the seven test cases in their table while a student with five of the seven cases, but missing the two specified cases, might be awarded zero or one of the four EIs. On the other hand, in an attempt to limit the amount graders had to evaluate, the same cases were often specified in multiple EIs (e.g., EIs 2 and 3 in LO 1) or even multiple LOs (e.g., LOs 1 and 2). When the student has an incorrect or missing response for the specified portion, the graders feel that the students are overly penalized for individual errors. Any of these variations of misrepresentation of performance due to specificity result in some graders overruling the rubric—40 of the 42 (95.2%) observed instances of overruling occurred in the context of over-specified EIs or LOs.

The third form of variability related to specificity is a potential overreliance on specifications. There are two primary examples. First, LOs 1/6, and 9 provides specifying text for three of the four EIs and four of the five EIs, respectively. While none of the EIs with specifying text were overlooked, the EIs without specifying text were overlooked 17 times. An alternative version of this occurred with EI 1 in LO 1/6. In these cases, while the EI stated, “creates *thorough* set of test cases *to test all possible outcomes* in the flowchart,” the specifying text only stated the number of cases that should be present. As a result, only two of the graders actually paid attention to the content of each test case to verify that each of the expected outcomes was present and all others only counted the number of rows. In other words, it would seem that when the blue specifying text is present, the EI itself is more-often-than-not disregarded.

The rubrics were subjectively appraised as either adequately or strongly robust. There were no glaringly weak rubrics that only narrowly addressed possible student solutions. It is important to note that while this dimension of the rubric is closely aligned with the comprehensiveness of the model response, it can differ in a not-directly-correlational sense. Ideally, both the model response and the rubric would be fully comprehensive and robust, but the robustness of the rubric could be related to how the EIs interact to encompass variable student responses or could be improved through additional text in the “what to grade” portion of the rubric. For instance, while the model response for LO 1 only had specific values in the input arguments, the rubric included extra text that specifies the whole range of acceptable values that could be used, helping the grader more easily evaluate work. The few LOs that were marked as adequate only had minor issues. For instance, the very short “Error: invalid viscosity” and “Laminar flow” outputs listed in the rubric do not give a thorough indication to the grader about what differing responses should qualify as “in English” and “not code results.” The provided outputs could be exactly the resulting outputs of the code; meanwhile, the response provided by sample 1 (i.e., “Print ‘flow is turbulent’”) feels like pseudocode and left multiple graders wondering how to handle. Alternatively, LOs 8 was created on the assumptions that students solved an open-ended task in a specific way—using a typical “if-elseif-else” selection structure. Two of the three samples had students write slightly different structure that were still fully functional, leading to a similar issue where graders felt the score produced by the rubric with many EIs based on the use of the traditional structure did not represent the student’s production of a structure that would still always produce the right output.

Overall rubric usability was based on the visual design of the rubric and operationalized through a few factors: the amount of information included within the rubric and the ability of the rubric to draw the graders’ attention to pertinent details, (operationalized by whether graders missed important information and amount of information included). In this way, the measure relates to dimensions such as the breadth of the LO the rubric specificity. It should be noted that there is a complex, somewhat contradictory relationship occurring within the usability dimension. When text is added to help provide guidance to the grader, the grader’s limited attention, time, or capacity of working memory may cause them to skim past or overlook that guidance. The additional information is helpful with decisions, but multiple graders noted that the more information is packed into the rubric, the more likely they are to ignore it. The usability of the rubrics observed ranged from low to high, but six of the eight unique rubrics were somewhere in

between. LO 4 was considered low usability because of the large number of evidence items (11) and the fact that all the text in the “what to grade” portion was blue and all the text in the EIs was black, causing some important information to not stand out to graders. On the other hand, previously discussed issues notwithstanding, LO 2/7 is brief enough that most graders appeared to look through a larger portion of the information and the used color effectively to highlight important information.

5.3.4 Grader training design

The grader training can vary with respect to the alignment between the sample problem used in the training and the problem graded in the assignment, the representativeness of the examples of student work relative to real student work, the accuracy of the “definitive scores” the graders are shown upon completion, and the specificity of the feedback to their learning needs. It may also vary in how useful the training is to the graders, but this is based on the overall output of the training, which is more needed as a structural component for the model and is simply an amalgamation of the other training outputs.

Unfortunately, training documents for three of the LOs were not collected when the data was available: LO 5 was missed due to a collection error while LOs 9 and 10 were initially trained prior to the assignment in this study and the training modules for previous assignments were not recognized as needed when they were available. Of the available training modules, Table 5.7 shows very little observed variability across the LOs. All of the collected LO modules were nearly exactly aligned with the assignment—they were each slightly more complex versions of the same problem (e.g., problem 2 but using nine atmospheric layers rather than five); however, because LOs 6 and 7 are the same as 1 and 2 but for a different problem, they were slightly less aligned.

Table 5.7. Observed variabilities of grader training design function outputs

IST	Generalized function	Design grader training		
Cognitive function	Create sample problem	Create or select example cases	Create quiz feedback	
Output variability	Alignment	Representativeness	Accuracy	Specificity
LO 1 (Problem 1)	Aligned [V↓]	Partial [V↔]	Moderate [V↔]	Moderate [V↔]
LO 2 (Problem 1)	Aligned [V↓]	Partial [V↔]	Moderate [V↔]	Moderate [V↔]
LO 3 (Problem 1)	Aligned [V↓]	Partial [V↔]	Accurate [V↓]	Moderate [V↔]
LO 4 (Problem 2)	Aligned [V↓]	Partial [V↔]	Accurate [V↓]	Moderate [V↔]
LO 5 (Problem 2)	Training documents not available			
LO 6 (Problem 2)	Partial [V↓]	Partial [V↔]	Moderate [V↔]	Moderate [V↔]
LO 7 (Problem 2)	Partial [V↓]	Partial [V↔]	Accurate [V↓]	Moderate [V↔]
LO 8 (Problem 2)	Aligned [V↓]	Partial [V↔]	Accurate [V↓]	Moderate [V↔]
LO 9 (Problem 3)	Training documents not available			
LO 10 (Problem 3)	Training documents not available			

All training modules were partially representative of student work. The two examples covered some errors that could be expected from students, but not all. This is understandable, as it would be difficult to capture a full spectrum of student responses with only two examples of work. Similarly, all of the modules were similar in the specificity of their feedback. The feedback was based on what feedback should be given to the student demonstrated in the sample, which is helpful for calibrating decision making to an extent but does not provide specific feedback to the grader if they identified an EI as being achieve or not achieved incorrectly. For instance, if they thought an EI was not achieved that the definitive mark considered achieved, the only feedback they would receive would be that they should have considered it achieved but not why and what about their thinking was faulty.

Lastly, there were three LOs that had “moderate” accuracy. This is because there was one EI that was marked incorrectly in the “definitive score” for one of the quizzes in LO 2, which may have misled some graders. Additionally, the two examples provided for LO 1 seemed to communicate differing expectations for performance of EI 2 compared to one another. This could also have led to confusion.

It is important to note that none of the variabilities observed across these functions could be possibly linked to any observations of the graders’ grading behaviors. For one, while the training should have occurred right before actual grading, the interviews occurred over one month after they would have graded the assignment. As a result, the training and application of the rubrics

would likely not have been fresh in their minds. Further, there would have been no way to know how each grader engaged with the training, and previous analyses indicate many graders did not take the training seriously.

5.4 Teaching Team and Student Functions

The teaching team and students are grouped together here because both agent groups were largely unobserved. The samples of student work can only serve as a proxy to get a sense of how performance of tasks may vary, with the recognition that considerable additional variance likely occurred in practice. As such, both teaching team generalized functions (i.e., delivering course content and guiding student activity and practice) and one student generalized function (i.e., learning content) are not explored here. It should be emphasized that these functions almost certainly contribute significantly to the variability of system performance, as outlined in chapter 4.

The variability of student responses collected cannot be used to infer anything about the excluded functions. For the teaching teams, all sections likely have a wide range of student performance. The samples used in the think-aloud interviews were purposefully selected to present a range of performances for graders to evaluate. One could potentially make conclusions about the teaching team, but it would require analysis of the entire section's performances across the whole semester, along with artifacts of course slides and field observations of class sessions. Similarly, it is difficult to make absolute inferences about students' learning of content based on individual samples of student work. Although, making conclusions about student knowledge is, in itself, the purpose of assessment, the potential influence of external factors, such as physical and emotional wellbeing and competing obligations, mean that task performance does not inherently produce a perfect reflection of knowledge or ability. It is important to recognize that course grades are compilations of many samples of student work and that single samples may misrepresent student learning.

5.4.1 Task performance

The first cognitive function associated with performing the task is interpreting the task; however, as with the other teaching team and student functions, there is no way to ascertain performance of that task based on the available samples. Although, it is expected that, as the

assessment task design section explains, variability in understandability of context, clarity of instructions, and scaffolding all likely affect variability in student interpretation of the task. Instead, it will be assumed that the samples provide a glimpse of the variability of performing the task. This section will make general reference to a few select responses to demonstrate extremes. More specific details of individual responses will be discussed in the next section to highlight the variability of the use and output of grading functions. All relevant samples of student work used in the think-aloud interviews are included in Appendix H for reference.

The overall variability of the three dimensions of student task performance are quality, typicality, and clarity (summarized in Table 5.8). The classification of quality was based on the overall fraction of achieved EIs for each LO, compared to the definitive mark, rather than based on the LO score, which discretizes the data too much to allow for proper discrimination for this analysis. Typicality was based on alignment between the student's response and anticipated responses, measured through perceived alignment with the rubric. Clarity was based on comparative subjective analyses from the samples used in the interviews. What is considered clear or unclear for a student response could vary greatly with consideration of additional examples that might shift overall expectations for student performances.

Table 5.8. Observed variabilities in task performance function outputs

Student	Generalized function		Perform assigned task	
	Cognitive function		Perform task	
Output variability		Quality (fraction of EIs achieved)	Typicality	Clarity
LO 1 (Prob. 1)	Sample 1	Moderate+ (0.75) [V↑]	Typical [V↓]	Clear [V↓]
	Sample 2	Moderate (0.50) [V↑]	Atypical [V↑]	Clear [V↓]
	Sample 3	Moderate- (0.25) [V↑]	Atypical [V↑]	Moderate [V↔]
LO 2 (Prob. 1)	Sample 1	Moderate (0.50) [V↑]	Atypical [V↑]	Clear [V↓]
	Sample 2	Moderate (0.50) [V↑]	Moderate [V↔]	Clear [V↓]
	Sample 3	High (1.00) [V↓]	Typical [V↓]	Clear [V↓]
LO 3 (Prob. 1)	Sample 1	Moderate+ (0.67) [V↑]	Moderate [V↔]	Moderate [V↔]
	Sample 2	Low (0.00) [V↓]	Moderate [V↔]	Clear [V↓]
	Sample 3	High (1.00) [V↓]	Typical [V↓]	Clear [V↓]
LO 4 (Prob. 2)	Sample 1	Moderate (0.64) [V↑]	Atypical [V↑]	Unclear [V↑]
	Sample 2	Moderate (0.55) [V↑]	Atypical [V↑]	Unclear [V↑]
	Sample 3	High (1.00) [V↓]	Typical [V↓]	Moderate [V↔]
LO 5 (Prob. 2)	Sample 1	Moderate+ (0.67) [V↑]	Moderate [V↔]	Unclear [V↑]
	Sample 2	Moderate+ (0.67) [V↑]	Atypical [V↑]	Unclear [V↑]
	Sample 3	Moderate+ (0.67) [V↑]	Typical [V↓]	Moderate [V↔]
LO 6 (Prob. 2)	Sample 1	Moderate (0.50) [V↑]	Atypical [V↑]	Clear [V↓]
	Sample 2	High (1.00) [V↓]	Typical [V↓]	Clear [V↓]
	Sample 3	Moderate- (0.25) [V↑]	Atypical [V↑]	Moderate [V↔]
LO 7 (Prob. 2)	Sample 1	Moderate (0.50) [V↑]	Moderate [V↔]	Clear [V↓]
	Sample 2	High (1.00) [V↓]	Typical [V↓]	Clear [V↓]
	Sample 3	High (1.00) [V↓]	Moderate [V↔]	Moderate [V↔]
LO 8 (Prob. 2)	Sample 1	Moderate (0.50) [V↑]	Atypical [V↑]	Moderate [V↔]
	Sample 2	High (0.80) [V↓]	Moderate [V↔]	Moderate [V↔]
	Sample 3	High (0.80) [V↓]	Typical [V↓]	Moderate [V↔]
LO 9 (Prob. 3)	Sample 1	High (0.80) [V↓]	Typical [V↓]	Clear [V↓]
	Sample 2	Moderate- (0.40) [V↑]	Moderate [V↔]	Clear [V↓]
	Sample 3	Moderate- (0.40) [V↑]	Moderate [V↔]	Clear [V↓]
LO 10 (Prob. 3)	Sample 1	High (0.80) [V↓]	Typical [V↓]	Moderate [V↔]
	Sample 2	High (0.90) [V↓]	Typical [V↓]	Clear [V↓]
	Sample 3	Moderate+ (0.75) [V↑]	Typical [V↓]	Clear [V↓]

The overall quality of students' responses ranged from demonstrating 0% to 100% of the EIs, according to the definitive marking. Eleven of the 30 samples were "high" quality and only one was a "No attempt," "low" quality response (in hindsight, there should have been more low-quality and fewer high-quality responses). The vast majority of work fell somewhere in between. As the definitive marking applied a relatively strict interpretation of the rubric, it is possible that there is more of a "true" quality value that better represents the overall quality of the work for some of the samples. For example, LO 3 is about converting between selection structure

representations and specifies one of two selection structures in the code to evaluate. Sample 2 did convert one selection structure but not the one that was specified. If the specified structure were changed, that same response might be classified as high quality. A more flexible application of the rubric, which was observed in the behaviors of the faculty and staff applying the rubric, might produce different scores that more accurately evaluate the student's abilities.

The typicality was based on the alignment between student response and anticipated approaches, as communicated by the model solution and the rubric. Typical responses mostly followed an expected approach and imposed no need for unique interpretation of the rubric. Moderate responses diverged slightly from the expected approach and conflicted with the rubric but could be reasonably reconciled. Meanwhile, atypical responses were unanticipated that they effectively broke the rubric. For some LOs (LOs 3, 7, and 9), no responses strayed far enough from the model or rubric to be considered "atypical," but were different enough from the model response that they could (and were observed to) cause minor confusion for the grader. For instance, for LO 8, students had to code a selection structure for atmospheric layers. Sample 3 is not perfect because it does not handle negative altitudes, earning it no credit for EIs 6 and 10; however, the student used a standard "if-elseif-else" structure that enabled straightforward evaluation. Sample 2 first checks that the altitude does not exceed the upper limit and then enters a nested "if-elseif-else" selection structure. The structure is a bit unexpected and requires momentary scrutiny, but ultimately can be handled by the rubric (also failing to account for negative altitudes). On the other hand, sample 1, uses an "if-if-if" selection structure. The result is that there are no "elseif" or "else" statements in the code. This makes it difficult to evaluate EIs such as, "Each 'elseif' is accompanied by a condition which a true result corresponds to code that immediately follows" because while the statement is not true, it is also not false—there is no "elseif" not followed by a condition. This "atypical" response breaks the implicit assumptions of the rubric. While none of the LO 8 samples were graded horribly inconsistently across graders, the atypical sample 1 was graded less consistently than the moderately typical sample 2, which was less consistent than the typical sample 3.

The clarity of each response was a subjective interpretation of how difficult it is to read or understand the student's response. Most were at least moderately clear or clear, even if the responses were lower in quality. On the other hand, even higher quality responses could be unclear. For example, Sample 1 achieves two-thirds of the EIs for LO 5; however, the handwritten response

is a little more difficult to read and interpret, and graders were observed to read some of the writing differently than others. Sample 3 of LO 10 was also considered unclear because the code provided, though given a proper function name and header, seemed to include code for the other function they were expected to write. This made it difficult to figure out what was going on with the code. The other samples that were classified as moderate clarity were mostly due to issues like limited commenting to describe lines or blocks of code (e.g., LO 10, sample 1) or limited inclusion of helpful information (e.g., the laminar flow case being unlabeled in LO 1, sample 3). These issues caused graders to have to take a slightly closer look, but they generally were able to come to common interpretations without much difficulty.

5.5 Grader Functions

Graders are the final agent to act in this system. As a result, upstream function variability generally leads to differences in grader behavior. The evaluative task imposed by each evidence item dictates that different types of evaluative functions will be necessary to properly score a given response. Further, the quality of each sample response drives greater differentiation in work-as-imagined evaluation functions for that sample. Discussion of the previous functions and their observed outputs have been mostly viewed through the subsequent behaviors taken by graders. As such, this section will briefly discuss work-as-imagined instantiations stemming from the particular task-sample pairings, then will discuss how actual work-as-completed grading diverged from the work-as-imagined instantiations, all after discussing training and preparation to evaluate.

5.5.1 Training

Most of the training functions, like the teaching team and student functions, were not included in the data collected for this study and would be difficult to infer. There is data available about the quiz scores for the training modules for the semester in which the interviews occurred, but the participants and quiz data have all been de-identified and can no longer be integrated to provide deeper information about the graders. What is known is the recency of training for each LO and the number of times each LO had been used in assignments prior to the assignment used in this study. It is expected that the more recently a grader has trained on an LO, the fresher the training will be in their mind. Table 5.9 shows that for all but LOs 9 and 10, the training occurred

right before grading the assignment. However, because the interviews happened over the span of a few weeks at the end of the semester, the recency was variable across participating graders. That said, there was no evidence in the interviews that this factor had a noticeable effect.

Table 5.9. Observed variabilities of training function outputs

Grader	Generalized function	Train to calibrate
	Cognitive function	Calibrate decisions
Output variability	Timing (weeks since)	Experience (# uses)
LO 1 (Problem 1)	Recent (0) [V↓]	None (0) [V↑]
LO 2 (Problem 1)	Recent (0) [V↓]	None (0) [V↑]
LO 3 (Problem 1)	Recent (0) [V↓]	None (0) [V↑]
LO 4 (Problem 2)	Recent (0) [V↓]	None (0) [V↑]
LO 5 (Problem 2)	Recent (0) [V↓]	None (0) [V↑]
LO 6 (Problem 2)	Recent (0) [V↓]	None (0) [V↑]
LO 7 (Problem 2)	Recent (0) [V↓]	None (0) [V↑]
LO 8 (Problem 2)	Recent (0) [V↓]	None (0) [V↑]
LO 9 (Problem 3)	Distant (4) [V↑]	Some (1) [V↓]
LO 10 (Problem 3)	Distant (5) [V↑]	Some (2) [V↓]

It is expected that a potentially more complex relationship exists with experience, whereby the more times an LO has been graded, the more familiar the graders should be with it; while this may lead to a more consistent interpretation, it may also lead to overconfidence in some graders believing they knew the LO without needing to review it thoroughly. This behavior was observed with LO 10, which had been used twice before this particular assignment. Many of the graders applied a more holistic approach to grading this LO, not looking closely at each EI.

There are a couple other important notes regarding grading. First, there is no way to know from any available information how authentically the graders in this study engaged with training. Further, some graders were far more experienced than others—some were sophomores in first semesters as teaching assistants and some were fifth-year seniors who had been teaching assistants since they were sophomores. This information was mentioned in conversation but not formally collected or documented. It is quite possible that those differences in experience affected the way they performed or benefited from training, but it is impossible to make such comparisons.

5.5.2 Evaluation preparation

All of the graders engaged in some amount of evaluation preparation (reading the problem, the solution, and the rubric) before grading each new LO. As they had all graded the same problems just a few weeks prior in their own sections, it is difficult to make any conclusions about their engagement in preparation activities. The participants were given a packet with the problems and solutions. All of the participants skimmed through the packet, some more quickly than others; however, it is entirely possible that those who reviewed the document more quickly already had the information more solidly internalized and did not need as much time. Thus, variabilities of these functions were not tracked and would likely not have been meaningful in the interview setting. More meaningfully, before each new learning objective, each grader did look over the new rubric; however, some skimmed and others carefully read each evidence item. This information was also not consistently recorded.

What was recorded was each time the graders re-read the rubric during grading. Fifteen of the 17 graders engaged in this activity anywhere from one to five times throughout the interviews. Most incidents occurred in LOs 1 (9 times), 4 (16 times), and 5 (7 times). Further, it almost always occurred while grading the first sample (32 out of 42 instances, or 76.1%), another eight instances (19.0%) occurred during grading of the second sample. Thus, it seems that graders feel fairly confident with their understanding of the problem, solution, and some aspects of the rubric after grading one sample. It should be noted, however, that the graders did frequently go back to the EIs to indicate which were achieved and which were not. Therefore, it is difficult to make claims about the tendency to read through the EIs more than once.

5.5.3 Evaluation of task performance

The context established by the outputs of all of the functions that occur in the background relative to the evaluation functions cause every LO-response pair to have a unique work-as-imagined instantiation. From a rigorous perspective, evaluation of each LO should require evaluation of each individual EI within the LO. As each EI in a rubric implies a different mode and object of evaluation, an ideal grading instance requires finding that object designated by the rubric and performing the evaluative task on that object as designated by the EI. For example, EI 2 of LO 1 suggests finding the invalid viscosity case and evaluating the description of the test case

and repeating the process for the laminar flow case. If the response, like sample 2, does not have one of the cases, the grader would be expected to not try to perform the evaluation of a description for a case that is not present—the grader would scan for the case, determine it is not there, and stop evaluating that EI. Alternatively, for EI 3 of LO 1, the grader just has to determine if the input arguments are listed in a valid form for those two cases. As such, they need to locate the appropriate case, and rather than evaluate the text, they are expected to determine whether the presentation format matches with their perception of a valid presentation format. In other words, they would engage in scanning then evaluating for EI 2 and scanning followed by matching for EI 3.

Figures 5.1 and 5.2 show examples of work-as-imagined instantiations expected to thoroughly evaluate fully correct responses for LOs 1 and 2, respectively. When applied to an imperfect sample response, there is a different work-as-imagined instantiation. For instance, when evaluating LO 1 for sample 2, the laminar flow case is missing (see Figure 5.3). The grader should still attempt to look for each test case so they will be able to give feedback to the student based on the missing cases. The missing case causes the first EI to be unmet rather than met (note the red circle indicating a missing input). The missing case prevents the evaluation of the description text, the matching of the input formats, and the evaluation of the input values for EIs 2, 3, and 4, respectively. Thus, all other EIs are unmet and the final score is ‘Insufficient Evidence.’

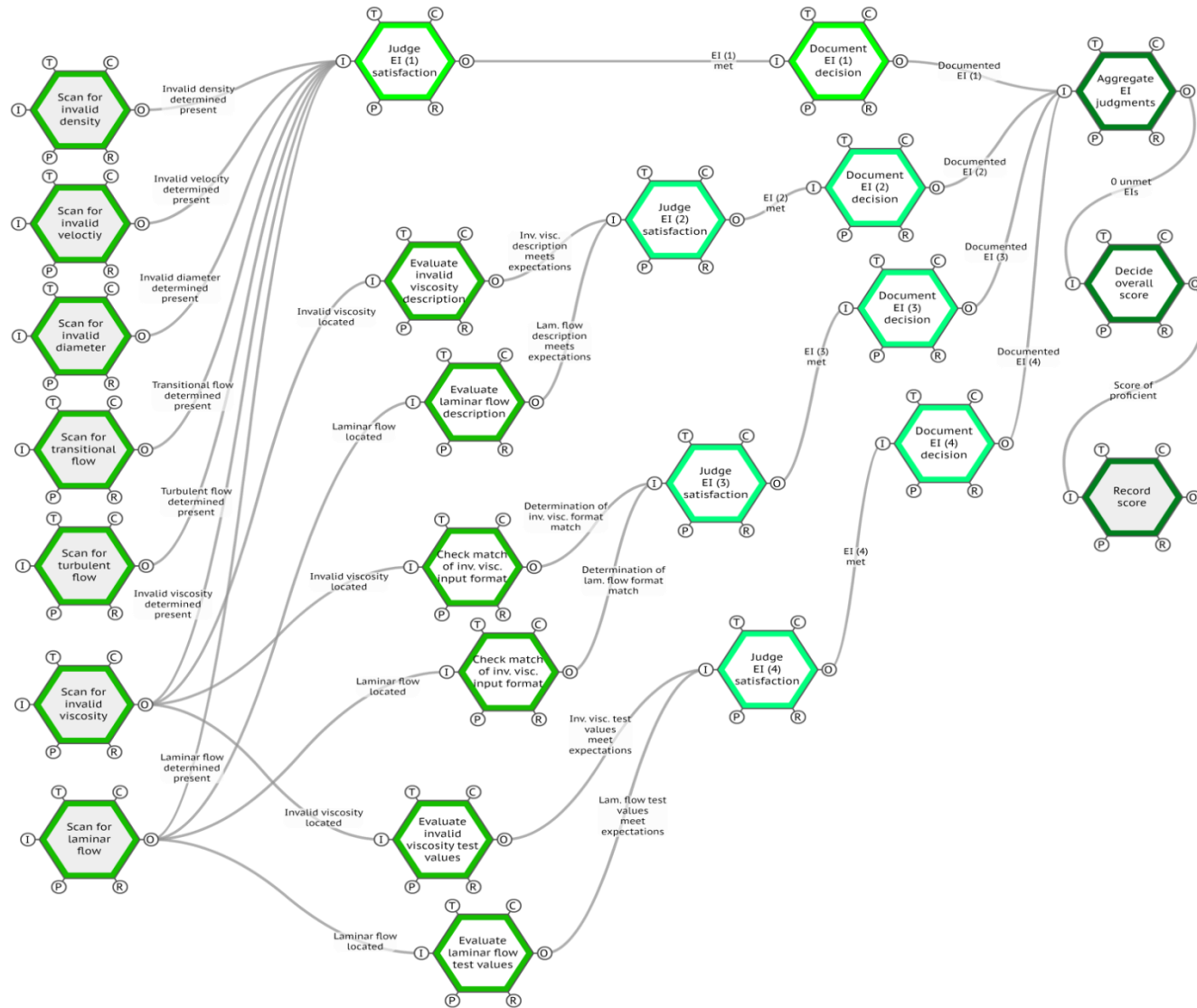


Figure 5.1. Work-as-imagined instantiation for “correct response” for LO 1.

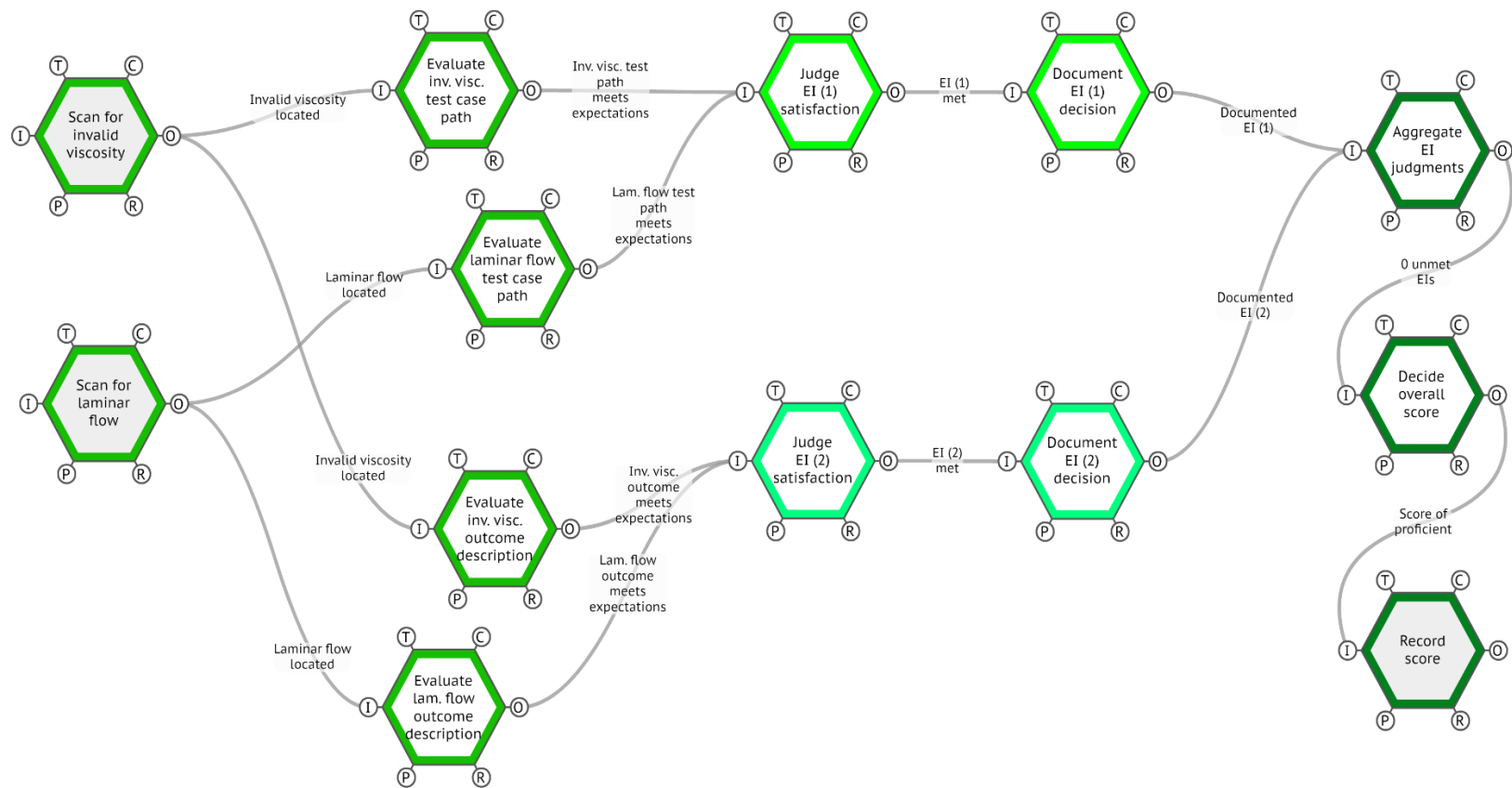


Figure 5.2. Work-as-imagined instantiation for “correct response” for LO 2.

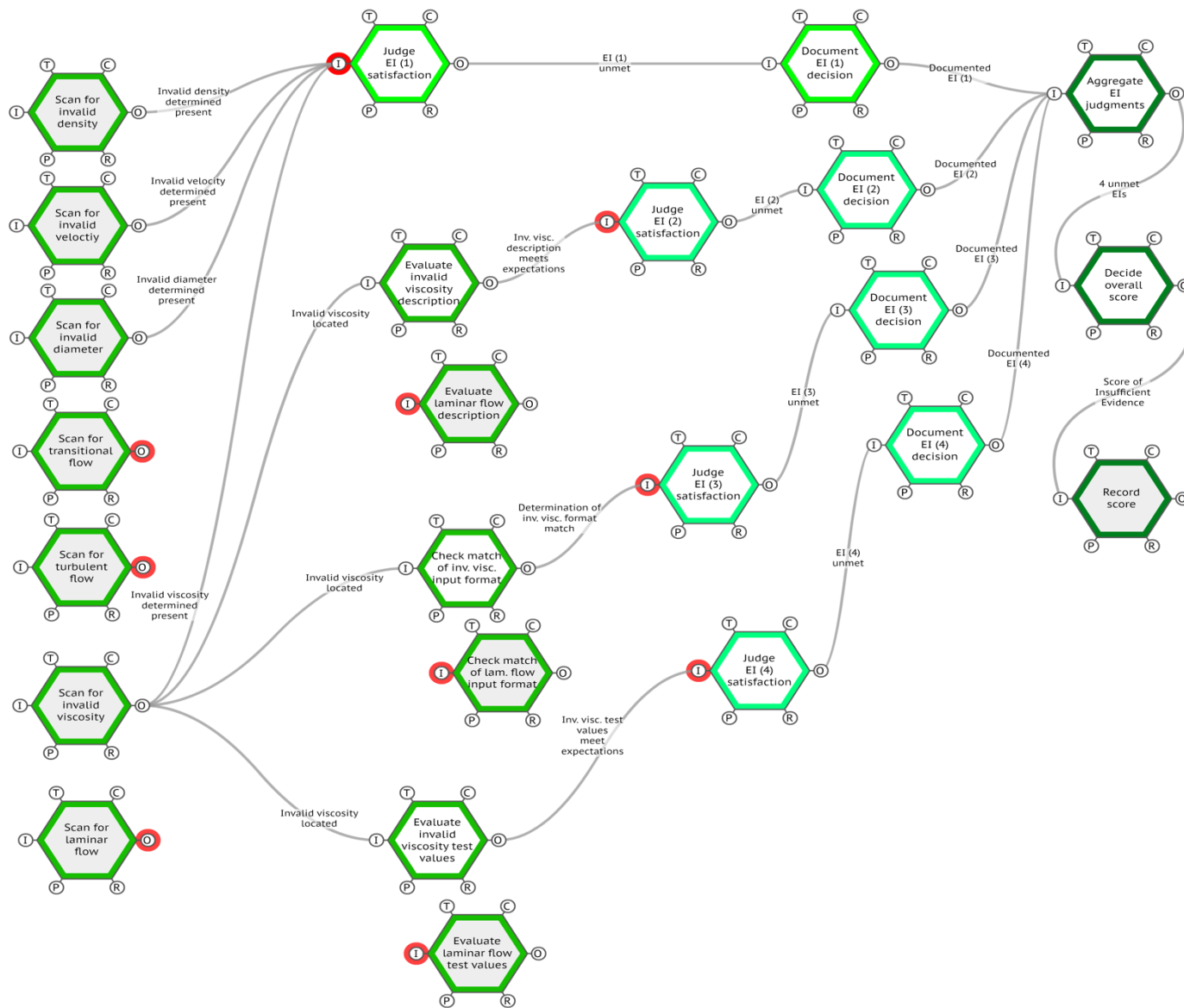


Figure 5.3. Work-as-imagined instantiation of grading LO 1 for sample 2

While Figure 5.3 shows the work-as-imagined instantiation for a thorough instance of grading LO 1, sample 2, it is more likely that a grader will identify that one of the necessary cases is missing and immediately decide all four EIs are not met, without verifying the presence of each of the other expected cases, evaluating the invalid viscosity test description, checking the format of the input values, or evaluating the input values of the invalid viscosity case. It is understandable that a grader would adjust their process in that way—it is significantly more time efficient. However, underlying all grading decisions should be grading expectations that establish the importance of giving the student the quality feedback they need to grow. If the grader stops after identifying the missing case, they will deprive the student of the opportunity to get feedback on all other aspects of their performance.

Despite these work-as-imagined instantiations, every grader had their own instantiation of grading each sample for each LO, totaling to 433 LO grading instantiations that encompassed 2,421 EI decisions (note: some LOs were graded without grading all EIs). Of the 2,421 EI evaluations, only 845 instances (34.9%) were evaluated using the same set of functions. Notably, 723 of those instances (85.6%) agreed with the definitive mark, despite 39 (5.4%) having unexpected outputs in at least one function. Of the 122 instances that used the same set of functions but came to a different scoring decision, only three (2.5%) had all the same function outputs as the work-as-imagined definitive marking and two of those occurred in the same LO-sample instantiation (i.e., LO 5, sample 2, grader 20). Both of those instantiations did result in a different overall score being assigned, as well. On the other hand, of the 158 instantiations that used the same set of functions but had more unexpected outputs, only 39 (24.7%) resulted in the definitive mark. Of the 1,725 instances (67.1%) that did not use the same set of functions, the definitive mark was selected only 1,207 times (70.0%). Thus, when using the same set of functions as the work-as-imagined instantiations, graders were 22% more likely to agree with the definitive mark ($p < 0.0001$).

In total, 1,930 EI decisions agreed with the definitive mark (79.7%) and 491 disagreed (20.3%). This is a rate of nearly four decisions that align with the definitive mark per disagreement. However, when considering the work-as-completed instantiations of each decision, the average number of modifications in comparison to the work-as-imagined instantiations (i.e., the total number of expected functions not used, unexpected functions used but not including the extraneous functions, and functions that led to unexpected outputs) was 2.97 when the definitive mark was

chosen versus 5.57 when the definitive mark was not chosen. This is an effect size of 0.73 fewer modifications from the work-as imagined instantiations when the grader agreed with the definitive mark than when they did not ($p < 0.0001$). In other words, when the final scoring decision agreed, the set of cognitive strategies used more closely resembled the work-as-imagined process.

Of the 2,421 EI evaluations observed, the 1,930 that agreed with the definitive mark were not distributed evenly across the LOs and EIs. Table 5.10 shows the breakdown of percentage of graders' scoring decisions that agreed with the definitive mark for each EI. These percentages were sometimes consistent for each sample within an EI but varied significantly in some cases. For instance, EI 1 in LO 2 has an overall agreement of 51.7%, but the agreement with the definitive mark by sample for this EI ranged from 86.7% for sample 3 down to 5.9% for sample 1. The overall breakdown highlights which EIs, on average, most typically disagreed with the definitive mark. This table also communicates, indirectly, the graders' agreement with one another. For example, the 23.5% agreement with the definitive mark for EI 4 of LO 1 corresponds to a 76.5% agreement between the graders. This means that percentages around 50% indicate the greatest disagreement between graders while low percentages in the table indicate that most graders made the same decision that happened to disagree with what was considered the "definitive" decision. The former scenario may be more difficult to address.

Table 5.10. Observed variability of criterion judgment outputs

Grader	Generalized function		Evaluate task performance			
	Cognitive function		Judge criterion satisfaction			
Output variability		Accuracy (% agreement with definitive mark)*	Accuracy (% agreement with definitive mark)*			
LO 1 (Prob. 1)	EI 1	Excellent (94.1%)	LO 6 (Prob. 2)	EI 1	Moderate (61.9%)	
	EI 2	Moderate (72.5%)		EI 2	Moderate (76.2%)	
	EI 3	Moderate (72.5%)		EI 3	Moderate (64.3%)	
	EI 4	Poor (23.5%)		EI 4	Moderate (61.9%)	
LO 2 (Prob. 1)	EI 1	Weak (51.7%)	LO 7 (Prob. 2)	EI 1	Moderate (77.8%)	
	EI 2	Moderate (75.1%)		EI 2	Strong (80.4%)	
LO 3 (Prob. 1)	EI 1	Excellent (96.1%)	LO 8 (Prob. 2)	EI 1	Excellent (100%)	
	EI 2	Strong (82.4%)		EI 2	Excellent (95.2%)	
	EI 3	Excellent (96.1%)		EI 3	Excellent (95.2%)	
	EI 4	Excellent (100%)		EI 4	Excellent (97.6%)	
	EI 5	Strong (88.2%)		EI 5	Excellent (95.2%)	
	EI 6	Excellent (96.1%)		EI 6	Poor (33.3%)	
LO 4 (Prob. 2)	EI 1	Excellent (100%)		EI 7	Strong (85.3%)	
	EI 2	Strong (86.2%)		EI 8	Excellent (97.6%)	
	EI 3	Excellent (97.9%)		EI 9	Excellent (90.5%)	
	EI 4	Moderate (64.4%)		EI 10	Weak (56.7%)	
	EI 5	Strong (80.1%)	LO 9 (Prob. 3)	EI 1	Excellent (93.9%)	
	EI 6	Moderate (73.5%)		EI 2	Poor (22.8%)	
	EI 7	Strong (81.9%)		EI 3	Weak (39.4%)	
	EI 8	Moderate (75%)		EI 4	Weak (39.4%)	
	EI 9	Moderate (75.8%)		EI 5	Excellent (93.9%)	
	LO 5 (Prob. 2)	EI 10	Strong (83.3%)	LO 10 (Prob. 3)	EI 1	Poor (23.9%)
		EI 11	Weak (46.9%)		EI 2	Moderate (72.2%)
EI 1		Excellent (97.8%)	EI 3		Strong (80.6%)	
EI 2		Moderate (63.9%)	EI 4		Moderate (66.7%)	
EI 3		Excellent (93.3%)	EI 5		Moderate (69.4%)	
EI 4		Excellent (91.1%)	EI 6		Moderate (66.7%)	
EI 5		Weak (41.0%)	EI 7		Excellent (93.9%)	
	EI 6	Poor (39.9%)	EI 8		Excellent (97.2%)	
			EI 9		Moderate (66.1%)	
			EI 10		Strong (88.9%)	

Note: For any percent agreement below 50%, the complementary percentage (i.e., 100% - reported percentage) indicates the percent agreement between graders. For instance, the 22.8% agreement with the definitive mark for LO 9, EI 2 indicates a moderate 77.2% agreement across the observed graders, which likely corresponds to the overly stringent, literal adherence to the rubric when creating the definitive marks.

The majority of the glaringly inconsistent EIs shown in Table 5.10 have related to the issues that have been presented and discussed in the previous sections (e.g., the skipping of EI 4 in LO 1 due to confusing meaning and no specifying text; the need for system 2 processing in EI 1 of LO 2, EI 11 of LO 4, EIs 5 and 6 of LO 5, EIs 6 and 10 of LO 8, EIs 2, 3, and 4 of LO 9, and EI 1 of LO 10). As such, one can reasonably argue that poor agreement between graders or with definitive marks is not random but due in large part to issues with rubric design or grader training. Also, when these percentages are broken down into the percentages of agreement for each individual sample, many of the lower levels of agreement occurred when the samples presented atypical responses (e.g., sample 1 for LO 2, where one of the two specified cases was absent), suggesting one of the most significant factors for inconsistency relates to insufficient rubric robustness.

Some of the differences in scoring outcomes could also be attributed to other issues. For instance, using the same set of functions as the definitive work-as-imagined instantiation does not guarantee the right outcome, as the internal variability of a function can lead different graders to produce different function outputs, which can then aggregate through later function inputs. Of the 12,047 observations where graders used functions they were expected to use based on the work-as-imagined instantiations, 1,600 (13.3%) resulted in unexpected outputs. Alternatively, across all interviews, there were 135 separate instances where graders either intentionally or accidentally skipped evaluating an entire EI. While this can increase efficiency when done purposefully (in cases where, enough evidence items have been identified as unmet that the LO score will be ‘Insufficient Evidence’ regardless of additional EIs, e.g., sample 2 in LO 9), it deteriorates the quality of feedback to the students. As such, applying the perspective assumed to generate the definitive marks, no amount of skipping error, whether intentional or due to human error, should be acceptable because students need feedback to correct misunderstandings or to reinforce proper understanding.

A similar and very common behavioral trend observed was the use of a more holistic grading approach, despite training that reinforces an analytical approach. Grading decisions were made holistically at the EI level in 912 separate EI instances (35.5%). Holistic grading came in multiple forms. At the EI level, the grader set out specifically to evaluate the individual EI, but rather than applying the functions in the work-as-imagined instantiation or some alternative set of function, only scans or evaluates the entire response or a large chunk and then makes a bigger picture judgment about the EI achievement. Alternatively, they make the judgment without

referring back to the sample, relying on their general impression of the work. This happened often when evaluating the EIs in LO 5 given that they had already looked over the same flowchart for LO 4. By relying on their general impression, they are not returning to the work to verify that the EI was consistently demonstrated. That said, holistic grading at the EI level still resulted in agreement with the definitive mark in 730 instances (80.0%) as opposed to agreement in 1,200 out of 1,523 (78.8%) non-holistic instances, though difference is not significant ($p = .480$).

Some graders performed the holistic approach slightly differently. For instance, some graders would, on occasion, look over the sample response, then go to the rubric and go down the list of EIs, saying yes or no to each EI they believed to be or not to be achieved based on their general memory and impression of the student's work. Alternatively, sometimes the grader would read through all of the EIs, then scan through the response looking to see if any unmet EIs caught their attention, then would go back to the list of EIs and note which ones they believed were met or unmet. Still another approach was a hybrid between holistic and analytic grading. In this approach, the grader would start trying to grade from a holistic approach, but then might specifically grade a few of the EIs more closely. While the data indicates that graders tend to have similar levels of accuracy using the more holistic approach, Joe et al. (2015) noted the tendency for graders to reduce their focus to a small number of features over time, in line with a limited working memory capacity. It is difficult to know if the same accuracy rate would hold up in a more natural context.

5.5.4 Scoring decisions

Once the graders have evaluated each of the EIs, they can aggregate the EIs and assign a total LO score. While it would seem unlikely, it is not terribly uncommon for this function to occur unexpectedly. In LO 1 alone, graders selected an overall LO score that did not align with their EI evaluations in seven out of 50 instances (14%). On the other hand, it is entirely possible that these surprising instances were the result of a grader intentionally overruling the rubric and not stating so explicitly. Table 5.11 shows the variability of the entire system's output in terms of the average severity (measured in terms of the Mean Actual Difference), the average magnitude of variability (measured by the Mean Absolute Difference), and the overall consistency of assigned grades (measured in terms of the standard deviation). For all three of these measures, the scores are adjusted to a 5-point scale. This allows for differentiation between 'No Attempt' and 'Insufficient

Evidence’ while equally spacing the difference so that the difference between ‘Insufficient Evidence’ and ‘Emerging’ is the same as between ‘Developing’ and ‘Proficient,’ which would not be true using the points assigned for each in the class. This paints a more accurate and consistent picture for interpreting variability of grading decisions.

Table 5.11. Observed variability of scoring output

Grader		Generalized function	Record score	
		Cognitive function	Decide overall score	
Output variability		Severity (MAcD)	Magnitude (MAbD)	Consistency (standard deviation)
LO 1 (Prob. 1)	Sample 1	Slightly lenient (0.389)	Moderate (0.722)	Moderate- (0.756)
	Sample 2	Slightly lenient (0.278)	Small (0.278)	Moderate (0.558)
	Sample 3	Slightly lenient (0.889)	Moderate (0.889)	Moderate (0.657)
LO 2 (Prob. 1)	Sample 1	Lenient (1.444)	Large (1.444)	Moderate- (0.896)
	Sample 2	Slightly lenient (0.444)	Moderate (0.556)	Moderate- (0.831)
	Sample 3	Slightly harsh (-0.667)	Moderate (0.667)	Low (1.106)
LO 3 (Prob. 1)	Sample 1	Slightly lenient (0.611)	Moderate (0.833)	Moderate- (0.951)
	Sample 2	Slightly lenient (0.278)	Small (0.278)	Moderate+ (0.448)
	Sample 3	Slightly harsh (-0.111)	Small (0.111)	Moderate+ (0.458)
LO 4 (Prob. 2)	Sample 1	Slightly lenient (0.333)	Moderate (0.667)	Moderate- (0.816)
	Sample 2	Lenient (1.471)	Large (1.471)	Moderate- (0.915)
	Sample 3	Slightly harsh (-0.313)	Small (0.313)	Moderate+ (0.464)
LO 5 (Prob. 2)	Sample 1	Lenient (1.111)	Large (1.111)	Moderate (0.737)
	Sample 2	Slightly lenient (0.625)	Moderate (0.875)	Moderate- (0.927)
	Sample 3	Lenient (1.500)	Very large (1.500)	Moderate (0.612)
LO 6 (Prob. 2)	Sample 1	Slightly lenient (0.533)	Moderate (0.533)	Moderate- (0.806)
	Sample 2	Slightly harsh (-0.133)	Small (0.133)	Moderate+ (0.340)
	Sample 3	Lenient (1.800)	Very large (1.800)	Low (1.046)
LO 7 (Prob. 2)	Sample 1	Slightly lenient (0.533)	Moderate (0.533)	Moderate- (0.806)
	Sample 2	Neutral (0.000)	Zero (0.000)	Very consistent (0.00)
	Sample 3	Slightly harsh (-0.714)	Moderate (0.714)	Low (1.161)
LO 8 (Prob. 2)	Sample 1	Slightly lenient (0.333)	Small (0.333)	Moderate (0.699)
	Sample 2	Slightly lenient (0.333)	Small (0.467)	Moderate (0.596)
	Sample 3	Slightly lenient (0.462)	Small (0.462)	Moderate+ (0.499)
LO 9 (Prob. 3)	Sample 1	Slightly lenient (0.538)	Moderate (0.538)	Moderate+ (0.499)
	Sample 2	Very lenient (2.462)	Very large (2.462)	Low (1.082)
	Sample 3	Very lenient (2.455)	Very large (2.455)	Low (1.157)
LO 10 (Prob. 3)	Sample 1	Slightly harsh (-0.308)	Small (0.462)	Moderate (0.722)
	Sample 2	Slightly harsh (-0.769)	Moderate (0.769)	Moderate (0.576)
	Sample 3	Slightly lenient (0.636)	Moderate (0.636)	Moderate+ (0.481)

* With respect to the definitive mark

Table 5.11 shows that there was unanimous agreement on only one of the 30 samples. It also shows that graders tended to err on the lenient side, with 22 of 30 samples (73.3%) being graded slightly to very leniently, on average, compared to the definitive mark. Given that the

definitive mark was based off of a rather strict and unforgiving interpretation of the rubric, this may be a reasonable outcome. Also, 23 of the samples (76.7%) were graded within one proficiency level of the definitive mark. Still, four of the samples (sample 3 of LO 5, sample 3 of LO 6, and samples 2 and 3 of LO 9) were graded, on average, more than 1.5 marks above the definitive mark and two others (LO 3 of sample 7 and sample 3 of LO 2) received a broad spread of grades. These instances are, perhaps, the strongest examples of aggregation of variability throughout the system, as they each relate to variable outcomes of background functions discussed in previous sections.

As a whole, while the graders only made one EI evaluation that differed from the definitive mark for every four that agreed, their overall LO score disagreed with the definitive mark slightly more often than not (47.8% in agreement). The majority of the LO differences were by no more than one proficiency level (31.4% of all scores). Still, given that across the eight unique LOs in the assignment had an average of 5.5 EIs, a 20.2% discrepancy rate means at least one discrepancy per LO, on average. Thus, it would make sense that the overall score differed by one almost a third of the time.

Lastly, one other holistic approach to grading was observed: grading the entirety of the LO holistically. For instance, in some cases, the student sample looked nearly identical to the model response. In these cases, it makes sense that the grader can quickly recognize that the sample holistically matches expectations and assign a score of ‘Proficient’ without investigating each EI individually. This is also more acceptable because it does not result in insufficient feedback to the student. However, there are also instances where a grader attempted to use this approach and came to a completely inaccurate conclusion. For example, grading LO 9 for samples 2 and 3, grader 4 used holistic evaluation for the entire LO and gave both samples scores of ‘Proficient’ when the definitive mark was ‘Insufficient Evidence.’

5.6 Function Variability and Aggregation

While all of the functions identified in the model were unable to be observed through the study, it is expected that all of the functions have some possibility of variability. Table 5.12 summarizes the variabilities that were observed for each of the IST functions, as well as the observed or potential (in italics) impact on the system as a whole. Tables 5.13 and 5.14 summarize the same information for the student and grader functions, respectively. Note that functions in the model that were not observable through this study were excluded from these tables, as speculation

of the potential variability of those functions are included in the model's presentation in chapter 4. Observed trends are stated based on multiple regression analyses regressing observed variables on behaviors and observed variables and behaviors on agreement with definitive scores for EIs and LOs.

Table 5.12. Observed variability and impact on the system for IST functions

Agent: IST			
Cognitive function	Dimension	Observed Variability	Observed Impact on Overall System Output
Articulate LOs	Clarity	Small	Moderate: Despite graders not seeming to pay attention to LOs, greater clarity corresponded to greater likelihood of holistic grading and agreement with EI definitive marks
	Breadth	Large	Small: Graders were slightly less likely to deviate from work-as-imagined or use holistic approach; however, there seemed to be no impact on agreement with EI or LO definitive marks
	Alignment	Small	Negligible: No trends observed, but variability may have been too small
Articulate EIs	Clarity	Small	Small: A few observations of confusion on unclear EIs Moderate: Agreement with the definitive mark decreases slightly with greater precision; some graders exhibited emotional responses to students being penalized repeatedly for minor errors
	Precision	Large	Moderate: More coverage corresponded with less holistic grading but also less agreement with either the EI or LO definitive marks; also corresponds with expression of confusion and overruling
	Coverage	Small	Large: Graders are more likely to follow the work-as-imagined process when evaluating system 2 EIs, but express some confusion and agree significantly less with the EI definitive mark
Select appropriate task	Complexity	Large	Negligible: Did not vary enough to observe trends
	Alignment	Negligible	Moderate: Openness associated with deviation from work-as-imagined and holistic approach; lower agreement with the definitive LO mark
Develop task context	Open-endedness	Moderate	Very large: More understandable greatly increased likelihood of deviation from work-as-imagined and holistic grading and corresponded to a large increase in agreement with the definitive EI and LO marks
	Understandability	Small	Moderate: Clearer instructions slightly increased use of the holistic approach, slightly increased disagreement with EI and moderately increased disagreement with LO
Write task instructions	Clarity	Small	Moderate: More scaffolding related strongly to a moderate increase in likelihood of disagreement with the definitive EI mark
	Scaffolding	Moderate	

Table 5.12. continued

Agent: IST			
Cognitive function	Dimension	Observed Variability	Observed Impact on Overall System Output
Develop model response	Comprehensiveness	Moderate	Moderate: Greater model comprehensiveness corresponded to moderate disagreement with the LO definitive mark
	Accuracy	Mostly negligible	Small: Greater model accuracy corresponded to slightly increased disagreement with the LO definitive mark
Differentiate performance levels	Discriminability	Large	Small: Greater discriminability corresponded to graders making fewer deviations from the work-as-imagined instantiations
	Specificity	Moderate	Small: Greater specificity corresponded to a greater likelihood of disagreement with the EI definitive score
Assemble rubric elements	Robustness	Small	Large: Greater robustness related to moderate deviation from work-as-imagined process but decreased use of the holistic approach; also related to greater likelihood of disagreement with definitive EI and LO marks
	Usability	Moderate	Moderate: Greater usability related strongly to fewer deviations from the work-as-imagined process and decreased likelihood of holistic grading but greater disagreement with both the definitive EI and LO marks
Create sample problem	Alignment	Small	Negligible: Too little variation in alignment across samples to identify trends; <i>ideally graders will be able to generalize, but graders did express concern about their ability to do this when the samples differed from the actual assignments</i>
Create examples	Representativeness	None	None: Variability in representativeness across samples too small to detect impact
Create quiz feedback	Accuracy	Small	None: Variability in accuracy too small to detect impact
	Specificity	None	None: Variability in specificity of feedback too small to detect impact

Table 5.13. Observed variability and impact of student task performance

Agent: Student			
Cognitive function	Dimension	Observed Variability	Observed Impact on Overall System Output
Perform task	Quality	Large	Very large: Graders are far more likely to grade holistically for very high-or low-quality work and are significantly less likely to agree with the definitive mark for either the EI or LO for mid-level work
	Typicality	Large	Negligible: Graders were slightly more likely to employ holistic approach but were in slightly stronger agreement with increasing typicality; negligible association agreement of EI or LO
	Clarity	Moderate	Small: More likely to cause the graders to grade carefully and engage in scrutinizing function; almost no effect on agreement with EIs; slightly better agreement with LO as clarity increases

Table 5.14. Observed variability and impact of grader cognitive functions

Agent: Grader			
Cognitive function	Dimension	Observed Variability	Observed Impact on Overall System Output
Calibrate grading decisions	Timing	Moderate	Negligible: Did not appear to have any meaningful effect on the system
	Experience	Small	Negligible: Did not appear to have any meaningful effect on the system
Interpret problem	Alignment	Moderate	Negligible: All graders initially looked over the problem, but some more intently than others; was not coded in a way to allow for meaningful differentiation
Develop model of acceptable solution	Alignment	Moderate	Hard to determine: There were instances where graders clearly gave credit for incorrect responses, but it was hard to tell if this was due to their model being off or something else
Review rubric	Alignment	Moderate	Moderate: Variable levels of care were taken in reading the rubrics, graders occasionally missed key details that led directly to grading errors
Scan for aspect of response	Accuracy (of detection)	Moderate	Large: Sometimes the object being scanned for varied and sometimes the conclusion of presence was inaccurate--hard to identify what caused either error, but the result was disagreement with definitive EI mark; occasionally used when higher-level evaluative function should be used
	Accuracy (of location)	Small	Large: Typically done correctly, but occasionally the wrong object was located, leading to disagreement with definitive EI mark
Check for exact match	Accuracy	Small	Large: Typically done correctly, but occasionally the wrong object was compared or the wrong conclusion was made, leading to disagreement with definitive EI mark
Check for effective match	Accuracy	Small	Large: Typically done correctly, but occasionally the wrong object was compared or the wrong conclusion was made, leading to disagreement with definitive EI mark
Evaluate meaning of response	Accuracy (interpretation and output)	Small	Large: Typically done correctly, but occasionally a wrong conclusion was made, leading to disagreement with definitive EI mark; occasionally not used in favor of a simpler strategy or used to holistically appraise a response
Scrutinize response to infer student knowledge	Alignment	Moderate	Large: Often not used in situations that likely warranted use; occasionally resulted questionable conclusion leading to disagreement with definitive EI mark
Judge criterion satisfaction	Alignment	Large	Large: Occasionally judgments are made based on gathering the wrong information or a judgment is made based on information that is inconsistent with the definitive mark; ultimately, the result is a disagreement with the definitive EI mark which could lead to disagreement with the LO mark

Table 5.14. continued

Agent: Grader			
Cognitive function	Dimension	Observed Variability	Observed Impact on Overall System Output
Document criterion satisfaction	Format	Large	Moderate: Sometimes documented physically with a check mark, other things it is documented mentally; in a few instances, mental "documentation" led to aggregated LO scores that disagreed with their EI decisions along the way
	Accuracy	Very small	Large: In a few instances, graders stated an EI was met or not met and wrote down the opposite, leading to a different LO score than would have been intended
Aggregate criteria	Accuracy	Small	Large: Occasionally graders miscalculated the number of EIs unmet (usually due to mental documentation), but not common; resulted in error in LO score
Decide overall score	Severity	Small	Large: Some graders made final overall score decisions based on a more holistic evaluation or adjusted their score despite the EIs to better align with their overall sense of the student performance (or voiced frustration that they could not)
Bring [item] to working memory	Likelihood	Hard to determine	Hard to determine: Graders had to revisit documents occasionally on all of the LOs, though more frequently while grading the first sample; it is impossible to know if they always reviewed documents when they forgot something and needed to do so
Question meaning/translate to support understanding	Likelihood/precision	Hard to determine	Hard to determine: Occurred relatively uniformly across all LOs; however, some graders were more likely to engage in this action than others; it is unclear whether all graders have sufficient self-awareness to do this consistently
Overrule interpretation	Likelihood	Large	Moderate: Some graders significantly more likely to overrule than others, typically leads to disagreement with a definitive mark or other graders
	Likelihood	Moderate	Small: Some graders were more likely than others to be self-reflective.
Reassure self about score	Accuracy	Moderate	Small: Graders reassured both when correct and incorrect, but there were a few more instances when it prompted an appropriate correction than vice versa; occurs more often with system 2 EIs
Modify score	Precision	Very small	Large: Improved score decision in most cases (9 out of 13 times) but was infrequent

Across these tables, there are a few key features to note. First, the functions that exist within the model but are excluded from these tables are still believed to make important contributions to the variability of the system. For instance, it is strongly suspected that the way activities and learning are supported during class has a strong influence over the quality of student learning and

possibly the distribution of responses the graders will ultimately have to evaluate. Second, it is important to reiterate that in many cases, the observed variability was based on the spread of samples within the study's context. By process, the clarities of the different materials were forced to vary across a set range of values. Thus, the size of the observed variabilities should be interpreted with that in mind. Finally, it's important to notice that the observed variabilities and observed impacts can contradict one another in magnitude. For instance, the final function shown, "Modify score" has a "very small" observed variability because it only occurred 13 times in the entire study, and in 69.2% of those times, it led to an improved score. The impact is considered large because it directly affects the output of the system. On the other hand, the variability in breadth of articulated LOs was large, ranging from two EIs to 11 EIs being associated; however, the analyses demonstrated that different breadths made very little difference to the use or outcomes of future functions or the system as a whole.

6. DISCUSSION

This study set out to explore the reliability of grading open-ended engineering tasks in large classes that rely upon many graders to improve the quality and meaningfulness of grading in those contexts. As the analysis and model presented in Chapters 4 and 5 should demonstrate, large courses, such as the one in this study, are complex, dynamic, socio-technical systems. When considering grading and the surrounding activities, there are several interacting agent groups, performing highly dynamic functions with multiple internal and external sources of variability. Attempts to minimize the variability of the system's output require a deep understanding of the system that extensively maps out all of the system's components and how they can contribute to the system's variability. As such, the first stage of this study employed Hollnagel's (2012) Functional Resonance Analysis Method designed for complex socio-technical systems. This process focused the analysis of each function on the various inputs and outputs and the interactions between them. As Human Reliability Analysis techniques like the FRAM are typically grounded in more industrial or physical applications, this application benefited from expanding beyond the traditional modes of output variability and defining "failure" modes that relate to the novel context, as encouraged by Hollnagel (2012, pp. 71).

Chapter 4 communicated, in detail, the comprehensive process model (RQ 1), including all of the functions involved (RQ 1a), who performs them, how they vary (RQ 1c), and how that variability can, in theory, aggregate and resonate within the system (RQ 1d). Chapter 5 provided a glimpse into the work-as-imagined instantiations and how they related to the contexts established by the course developer and student functions (RQ 2a). It broadly addressed how work-as-completed differed from work-as-imagined (RQ 2b) and the background functions most substantially impacted subsequent functions (RQ 2c). Chapter 5 also provided statistics that will be useful in understanding system resilience (RQ 2d). This chapter aims to round out the remaining questions and provide connections to previous research findings regarding how:

- 1) the model extends previous models of grader cognition (RQ 1b)
- 2) the variables contributing significantly to variability relate to the conclusions of previous research studies (RQ 2c)
- 3) the theory supports the observed differences between imagined and completed grading instances (RQ 2b)

- 4) resilient is the system to variability and its points of weakness (RQs 2d & 3), and
- 5) what mechanisms could reduce variability (RQ 3a)

The chapter concludes with a brief discussion of the generalizability of the study's findings, based on the ecological validity of the research and the degree to which aspects of the study are unique to the context studied.

6.1 Extensions Beyond Previous Models

This research extends previous research on grader cognition in four distinct ways: (1) it expands the context to the evaluation of open-ended engineering tasks; (2) it extends cognitive behaviors to actions performed by the grader and others before actual grading; (3) it elaborates upon and adds nuance to the previously identified cognitive behaviors; and (4) it relates graders' cognitive behaviors directly to variable contexts. The model also provides evidence to validate aspects of previous models, as many of the same behaviors occurred during this study.

6.1.1 Grader cognition in context

This study examines grader cognition in the previously unexplored context of open-ended engineering tasks. While there have been several studies exploring grader cognition conducted over the past few decades, most analyses have occurred within the context of grading verbal performances. Charney (1984) and Lumley (2002) focused on the evaluation and rating of written language. Joe et al. (2011) analyzed raters of undergraduates giving speeches. Meanwhile, Cumming (1990), Vaughan (1992), Milanovic et al. (1996), and Orr (2002) all looked at raters of writing and speaking for learners of English as a second language. Only a few studies have branched out to other subjects in higher educational contexts, including mathematics and physics (Laming, 1990; Suto, Greateorex, & Nádas, 2009), business studies (Greateorex & Suto, 2008), social sciences, and law (Webster et al., 2000).

The extension of context provided by this work is noteworthy, as grading behaviors are dependent upon context. In their work analyzing raters of General Certificates for Secondary Education, Greateorex and Suto (2008) and Suto et al. (2009) determined that differences in the types of problems presented to students and the frequency and use patterns of cognitive marking strategies vary across disciplines (see §2.3). While this study did not compare those frequencies or

patterns with other disciplines, the model can be applied to understand how cognitive strategies utilized while evaluating open-ended engineering tasks might differ.

6.1.2 Cognition outside of grading

This study situates grading within a broader ecosystem of agent groups and behaviors. Previous studies on grading cognition focused solely on the behaviors of the grader, alone. To apply the FRAM, a complete model required inclusion of agents involved in producing all of the materials and activities a grader must interact with or engage in while grading student work. Within the context of a large system, this model demonstrates an enormous number of actions taken by others that directly or indirectly affect the grading process. The model explicates what are likely often subconscious or implicit decisions in multiple stages of course development and implementation, bringing conscious awareness to those agents of the potential consequences of their decisions. Breaking the model into multiple levels of abstraction should communicate a bigger picture of each agent's goals to the other agents in the system without the overwhelming degree of detail needed only for that particular agent group.

The FRAM model demonstrates, based on the number of functions and interactions, that the agents involved in making curricular decisions and designing course materials are the most influential in the system. Those actions (i.e., developing content, schedules, assignments, grading guidelines, and training) directly influence teaching, learning, and grading. When the agents responsible for curricular development either do not also teach or grade or are not alone in those activities, it is essential that they carefully plan and communicate all materials clearly and transparently and set realistic schedules. The clearer and more effective the designed lecture content (i.e., slides and activities), the more likely instructors will implement classes as envisioned. Further, assessment tasks that are well aligned with the course content and within a realistic estimation of the students' capability development make students more likely to learn effectively and produce less variable, higher-quality task performance. Finally, the course developer's design of and communication about organizational expectations (Lumley, 2002), the overall grading scheme (Ahmed & Pollitt, 2011; Lengh, 2010; Thompson et al., 2013), the rubric (Goldberg, 2014; Joe et al., 2011; Menéndez-Varela & Gregori-Giralt, 2018; Moskal, 2003; Popham, 1997; Tierney & Simon, 2004), and training (Alshuler, 2016; Baird et al., 2017; Joe et al., 2011) all influence

grader behavior and consistency. The effects of rubric design, in particular, were well supported by observations within this study.

The teaching team and students also play an important role in grading consistency. While the model demonstrates that there are no direct interactions between the teaching team and the graders, the teaching team's actions have the most direct effects on the students as they deliver all of the content and course materials. The teaching team's responsibility is to ensure that the students are presented the material in the most effective fashion possible and are supported sufficiently through interactions and in-class activities. The teaching team's support should lead to a more consistent understanding and a firmer grasp of the content by the students. Student understanding leads to what this study demonstrated to be one of the strongest influences on grader consistency: the student's performance. As Russell et al. (2017) and Cooksey et al. (2007) noted, mid-quality or unanticipated work is the most difficult to grade consistently.

By developing all of the interactions between functions and agents, this model shows how each function interacts with the other actions taken by each other agent group. Careful examination of each function's output can indicate various ways that outputs can be focused upon to attempt to produce conditions for grading that will most facilitate grading consistency.

6.1.3 Additional cognitive functions and nuance

The FRAM model was developed primarily through iterative analyses of interview observations and transcripts and the course materials used within, interpreted through the lens of personal experience but attempting to disregard the theories put forth in the literature (i.e., avoid *a priori* coding) as recommended for the FRAM by Hollnagel (2012). Despite this, functions identified throughout the model development process coincide well with previous frameworks, effectively marrying many of the elements of multiple models of cognitive grading strategies while offering some refinement and positioning the process within a broader perspective of human factors engineering.

The first notable observation from the interviews was how the graders used the *scanning* function. It was observed consistently throughout the interviews that *scanning* was always the first action taken by graders, whether they needed to grade a small portion of a response (e.g., LO 8) or an entire response (e.g., LO 10). Additionally, graders performed the *scanning* function in a variety of ways. The simplest form is when the grader only needs to identify if the student has included

something within their response, which may be referred to as *scanning for presence*. Based on this interpretation, the *scanning* function subsumes within it the *no response* function from Greatorex and Suto's (2006) framework as one of the possible outputs of *scanning for presence* (i.e., that the response is not present). In other instances of *scanning*, the goal was not to detect presence but to locate a portion of a response to perform further cognitive functions. *Scanning* in this way produced a different output (i.e., a location of the object to be graded). Two other forms of *scanning* were observed but will be discussed subsequently in the context of holistic grading.

Once the graders *scanned* the student's response, the next function they used depended upon the output of the *scanning* function and the need dictated by the evidence items. Assuming the evidence item required more than detection of presence (and that the grader interpreted the EI as such), the grader would typically engage in some form of *matching*. This sequence is notable because it conflicts with Greatorex and Suto's (2006) argument that *matching* is the simplest, function followed by *scanning*. While it is arguable whether *matching* or *scanning* requires less cognitive effort, Greatorex and Suto's presentation gives the impression that *scanning* is likely to follow *matching*, based on being more effortful. One might argue that the different sequence observed in this study occurred because the rubrics often specified portions of a longer response to grade; however, it is notable that even when specific portions were not specified (e.g., LO 10), graders performed one of the more holistic forms of *scanning* (to be discussed subsequently). Thus, *scanning* always occurred first.

The *matching* behaviors observed by graders in this study indicate that *matching* can either take the form of looking for an exact match or an effective or approximate match. *Exact matching* may be a common approach for very closed-ended contexts, but in the assignment used for the think-aloud interviews, even the most closed-ended task allowed slight room for variation in the students' responses (e.g., differences in variable names). As such, it would be rarer for a student's response to be identical to the solution compared to the context of a mathematics problem with only one acceptable answer. With the slight possibility of an exact match, graders often *scanned* to find the designated portion of the student's response first and tried to perform *exact matching*. However, once they ruled out an exact match, the grader applied the *evaluation* function (discussed in more depth subsequently) to develop an interpretation of the student's answer and, in some cases, transitioned to *approximate matching* to determine if their interpretation of the response matched with a more generalized interpretation of acceptable solutions. This observation extends

and adds some nuance and context to the previous description of *matching* as an initial and immediate determination of awarding points or needing to engage in other strategies that occur at the start of the evaluation. Considering Suto et al.'s (2009) observed variability of function use across disciplines, the nuance and difference in sequence may stem from specific aspects of the open-ended engineering tasks compared to the problems studied by Greateorex and Suto (2006).

The discrepancy between *exact matching* and *approximate matching* leads to interesting features of the *evaluation* function. As shown in Chapter 4, the *evaluation* function had several possible outputs, depending on the function's inputs (i.e., the quality of the response and the evaluative task designated by the EI). In some cases, the *evaluation* function immediately resulted in a determination of whether or not the response meets expectations for criterion satisfaction. In this form, *evaluation* resembles an intuitive, system 1 process. In other cases, the function either output an interpretation of the response or, if *evaluating* was insufficient to derive meaning, an uncertain interpretation. When the grader obtained an interpretation from *evaluating*, *approximate matching* was applied to determine if the interpretation resembled a generalized model of acceptable responses. Alternatively, when an uncertain interpretation occurred, *scrutiny* was needed to identify the extent to which the response met expectations. In either case, the grader engaged in a slower, more effortful system 2 process. To an extent, these observations aligned with Greateorex and Suto's (2006) acknowledgment that *evaluation* occurs as either a system 1 (i.e., quick, effortless) or a system 2 (i.e., slow, effortful, reflective) cognitive process, depending on whether the grader would need to utilize external knowledge or information to make an accurate judgment. However, the findings of this study base the distinction on the function's output.

The variable outputs of *evaluation* highlight aspects of the *scrutinizing* function. The use of *scrutinizing* observed in the think-aloud interviews often occurred when the student response was unexpected, unaligned, or partially incorrect; however, it also occurred in other situations, such as when a grader misinterpreted some aspect of the assignment, the rubric, or the student's response. That is, *scrutiny* does not only occur because of the student's response but can also be due to partially inattentive engagement by the grader. Additionally, creating work-as-imagined instantiations as part of the FRAM led to an important observation: *scrutinizing* is never present in an ideal case—it only occurred when something had gone wrong on the student or the grader's behalf. These results provide nuance to Greateorex and Suto's (2006) descriptions of *scrutiny* as occurring primarily due to unexpected, partially incorrect, or not adequately aligned responses.

The function identification stage of the FRAM also helped to illuminate a previously unidentified function. Regardless of which function or functions were used along the way, the grader always concluded the evaluation process by *making a judgment* about the appropriateness of awarding points. This distinction of *judging* as its own unique function is necessary. For example, even if a grader determined a match using the *matching* function, they could, and occasionally did, make the inappropriate judgment that the student's work did not merit points. Thus, the action of *judging criterion satisfaction* warrants its own function that can have a variable output. Greateorex and Suto's (2006) framework, on the other hand, clumped judgment in with the *scrutinizing* strategy, arguing that judgment only occurred within the act of *scrutinizing*.

In addition to functions included in the direct evaluation of student work, graders in this study also demonstrated the cognitive patterns that fall into the model's pre-evaluation functions. Before grading, graders typically skimmed the grading documents to develop an understanding of the problems, what they needed to evaluate, and what features to look for in the student's response. These actions often occurred before a new LO but were occasionally revisited as needed, particularly early in grading a new LO. These functions coincide with what Lumley (2002) referred to as management behaviors.

This study also revealed grader's use of post-evaluation functions. Graders often *revised scores* and *reassured* themselves about their scoring decisions. The reassurance behavior occurred more often than not when graders made final scoring decisions. While reassuring themselves, graders often recognized errors they made and adjusted their scoring decisions. Occasionally graders revised scores in the wrong direction, but this was generally not the case. Either way, it helped the grader feel at ease with their decision and move on. These observations also align with Lumley's (2002) claim that graders justify scoring decisions concerning criterion and scale descriptors and reconsider, revise, or confirm their scoring decisions. Lumley argued that graders rely upon these strategies to feel confident that their decisions reconcile any contradictions or misalignments between the student work, the grading guidelines, and their perception of the general organizational expectations.

Throughout the interviews, the graders also frequently used holistic strategies, which revealed two other nuanced forms of *scanning*. There were three typical ways that graders graded holistically: (1) they initially scanned the whole student response for general and specific features and then *scanned* the list of EIs to see if any flaws in the student's work warranted point deduction;

(2) they *scanned* the rubric to get a sense of LO performance expectations, then *scanned* the student's work, trying to spot flaws related to the rubric; or (3) in line with Charney's (1984) description of grading as highly idiosyncratic, graders read the LOs, without paying attention to the EIs, then *evaluated* the work as a whole and assigned a score. The first two strategies illustrate that, in addition to checking for presence or locating a feature of a response, the *scanning* function can also have the purpose of gathering a general sense of performance or trying to spot specific flaws. Interestingly, these holistic approaches strongly align with the three general grading behaviors identified by Lumley (2002).

Further contributions of the FRAM model include functions identified that occur before grading and unexpectedly during grading. There are several pre-grading cognitive operations (e.g., *engaging in training, familiarizing with the task, familiarizing with the rubric, familiarizing with the solution*) that have significant consequences when neglected. The model illustrates how not engaging in those functions can affect evaluation and scoring functions, leading to potentially faulty scores. The model also includes several cognitive operations that can occur unexpectedly throughout the grading process (*forgetting, reviewing, translating, overruling, modifying, or reassuring*). While irregular and unpredictable, these functions exert a substantial influence over the effectiveness of all the functions used during grading (e.g., having something in working memory).

Collectively, the idiosyncratic holistic approach and the additional cognitive operations included in the model, which occur before and during grading, demonstrate how general erroneous human actions cause differences between work-as-imagined and work-as-completed instantiations. As Sharit (2006) notes, erroneous actions, or those that result in undesirable outcomes, are typically the result of challenges posed by contextual factors or basic human fallibility. The limitations associated with policies, culture, competing demands on time, need for knowledge, procedural clarity, training, or communication can all interfere with ideal practice. Within this context, communication directed toward graders can be unclear with respect to expectations of student performance, background knowledge needed to grade a problem, the philosophy behind grading, and expectations of grading procedures. Additionally, humans have limited sensory, attentional, and working memory capacities, which can be exacerbated by personal dispositions or fatigue and competing obligations. Due to human limitations, when procedural protocols become overwhelming, graders often resort to heuristics, which save time and effort but increase the

likelihood of error (Sharit, 2006). This tendency perfectly explains why many graders employed holistic approaches more frequently in the longer LOs and later within the interviews.

6.1.4 Connections between functions

By discussing different work-as-imagined instantiations for grading based on features of the rubric and student work, this study demonstrates how contextual variables established by background functions contribute directly to the cognitive processes that should be utilized by the grader. For example, when the evidence item requires a system 2 level of processing to evaluate, such as interpretation of a textual response or appraisal of a set of test values aligned with the designated flowchart path, the grader should engage in the *evaluation* function. Meanwhile, if the student's work is extremely misaligned with reasonable expectations, the grader is expected to use the *scrutinizing* function to examine the extent to which the evidence items are demonstrated.

By articulating all of the inputs, time constraints, controls, execution conditions, preconditions, and outputs, the model also shows how each of the cognitive functions used within grading interacts with one another. For instance, it shows that when a grader attempts to determine if a response matches the model response, if the match is not exact, the output should lead the grader to engage in the *evaluation* function. Thus, it helps to outline an expected sequence of cognitive functions that the graders should utilize to evaluate a student's performance. It also provides useful language to describe when a particular instantiation has not occurred as expected by noting an output for each function. Following the same example, if the grader recognizes that there is not an exact match and immediately moves to judge that the criterion is not achieved, then it can be easily asserted that the grader has skipped over a necessary cognitive process to determine achievement. It may be the case that their ultimate conclusion was correct, but it was a matter of chance rather than thoroughness.

The six function aspects also highlight how functions performed by the grader before grading (e.g., training or grading preparation) or by other agents (e.g., setting grading schedules) can influence the actions performed by the grader. In an ideal case, one should know the specific actions a grader will take by knowing the context of the LO/EI and the response they are grading, either assuming the grader is adopting thorough or efficient strategies (e.g., not grading all EIs once they determine that the score must be 'Insufficient Evidence'). These connections illustrate

how an output varying in an unfavorable direction can lead to the misuse of subsequent functions, or the use of appropriate functions with inappropriate inputs, to produce an unfavorable result.

6.2 Impactful Variables

All of the functions identified in this study have some possibility of variability. Hollnagel (2012) states that, in general, human functions are subject to many physiological and psychological sources that could contribute internally to performance variability and are likely to be high in frequency and large in amplitude. On the other hand, organizations may have many function-specific or culturally based internal sources of variability that are likely to be low in frequency but large in amplitude. Human functions are also subject to many social and organizational external sources of variability that are likely to be high in frequency and large in amplitude. Meanwhile, organizations face many instrumental or culture external sources of variability that are likely to be low in frequency but large in amplitude. Hollnagel also noted that human functions are most likely to have acceptable levels of precision but are occasionally imprecise and rarely have high levels of precision. On the other hand, organizational functions are likely to be imprecise but are possibly acceptable and unlikely to be precise.

Section 5.5 presented tables that summarized the observed variability of each of the functions analyzed in this study. Many of the functions can impact the system substantially, some related to the variability of each system instantiation and others relating to the system's overall outcome (i.e., the LO score assigned by the grader). In some cases, a function could lead to more significant instantiation variability. For instance, creating a highly usable rubrics appeared to help graders to employ a grading process that more closely resembled the work-as-imagined instantiation. In contrast, graders were more likely to use a holistic approach to grading with highly typical student responses (i.e., widespread approaches that are likely to be seen repeatedly and anticipated). Notably, it could be the case that the outputs of some of these functions were heavily confounded due to the selection of samples and the context of the assignment. Thus, any trends that were or were not observed may have been partly due to limitations in the data, such as the fact that the operational breakdowns of each LO (i.e., the set of EIs) were relatively strong for all the LOs in the assignment. Had all of the course's LOs been involved, the observed variability may have been broader, altering the scale of the variable and the observed relationship with others.

There were several outputs of background functions that established contexts impacting not only the grader's final LO score but also the actions they took to assign that score. There were many features associated with the content, assignments, rubrics, and student work that all contributed to notable variability in the system, either observationally or statistically. These features were described in the previous two chapters but should be connected to the relevant literature or discussed more thoroughly. Also note that due to limited variability in samples studied, the claims that do not directly support previously argued positions warrant more targeted future research. Conversely, due to the limited sample variability, a lack of observation of significant impact does not necessarily imply that a given function is not impactful—just that this study did not reveal the magnitude of the function's impact.

6.2.1 Learning objectives

The learning objectives' clarity and the complexity, coverage, and precision of the evidence items all influenced grader behavior and system outcomes in meaningful ways. The overall LO clarity, while varying minimally across the eight LOs in the study, corresponded to an increased likelihood of agreement with the definitive mark. Interestingly, LO clarity also related strongly with increased deviation from work-as-imagined grading and the adoption of more holistic approaches. These two trends contradict trends across the broader dataset: (1) holistic and analytic grading approaches agreed with the definitive mark at similar rates; and (2) the more closely the grading resembled the work-as-imagined instantiation, the more likely it agreed with the definitive mark. This conflict suggests that greater LO clarity allows a grader to grade holistically more accurately than they would otherwise. It may be that a high degree of LO clarity supports intuitive grading by establishing a strong understanding of the construct under evaluation, which could explain Charney's (1984) findings that gut instinct grading was more accurate than slow and deliberate grading. This explanation may also extend Joe et al.'s (2011) finding that experienced graders possessed more confidence relying heavily on holistic intuition to include less experienced graders for very clearly articulated LOs.

The extent to which each LO was operationalized into a set of discernable and observable behaviors (i.e., the LO's EI coverage) had multiple implications. Broader coverage corresponded to less deviation from the work-as-imagined process and less frequent holistic grading. At the same time, it also related to greater disagreement with both the EI and LO definitive marks. These trends

were statistically significant, the graders never commented, directly or indirectly, on the adequacy of the LOs' operational coverage. While the stronger resemblance to the work-as-imagined instantiations supports Goldberg (2014) and Sadler's (2010) logical and reasonable recommendations to operationalize every construct as fully as possible to prevent the exclusion of any essential features of performance, the reduced accuracy seems to conflict. The observation may be a result of the specific grading structure of the course under study. Graders occasionally impose additional idiosyncratic requirements for criterion satisfaction that are not stated in grading guidelines (Charney, 1984; Joe et al., 2011). As each operational dimension corresponds to an additional grading decision in the studied course, fuller coverage generally corresponds to more opportunities to impose unique perspectives. Given the size and number of graders involved in the course, there is a wide variability of background knowledge and experience that leads to an increased likelihood of observing idiosyncratic variability. As a result, there are, inevitably, more observations of differing decisions.

While graders did not explicitly express concern about EI coverage, EI precision caused considerable, observable consternation for graders. To a minimal effect, overlapping or indistinct EIs corresponded to stronger resemblance to the work-as-imagined grading process and agreed less frequently with the definitive EI mark. Graders repeatedly expressed frustration by EIs that required achievement of other EIs in order to be achieved. For instance, LO 4's inclusion of two EIs that start with "arrows must..." after a previous EI based on the presence of arrows upset several graders—they felt the rubric excessively penalized the student response that lacked arrows. LO 8's implicit dependencies between EIs 3–7 and 9 produced similar frustration. Forsythe et al. (2015) and Lerner et al.'s (2015) reports of neurological and emotional effects on decision making in response to perceived unfairness may explain why some graders chose to defy the rubric in these situations. The prominence of this theme within the graders' comments suggests a more significant effect than was observed, statistically, which may have been ironically due to the imprecise definition of EI precision. Including both overlap and inadequate precision of wording (i.e., use of terms like "minimal," "concise," or "valid" without clarifying descriptions) into the same dimension may have weakened the measured effect.

The last notable aspect about the learning objectives was whether or not the EIs required system 1 (*scanning, matching, or low-level evaluating*) or system 2 (*high-level evaluating or scrutinizing*) processing to evaluate. The EIs that clearly required higher levels of processing to

evaluate (e.g., evaluating the passing of information between functions, that a selection structure dealt with all the necessary paths, or that test values were consistent with test descriptions) tended to be graded with extra attention, as graders' actions closely resembled the work-as-imagined instantiations. However, the graders expressed more frequent confusion and agreed with the definitive mark less frequently when grading the system 2 EIs than the system 1 EIs. These trends reinforce Suto and Grestorex's (2008) findings that grading consistency suffers when grading requires system 2 processing. It is reasonable to assume that EI complexity imposes greater cognitive demand on the graders. To parallel the cognitive demand frameworks of Smith and Stein (1998) and Tekkumru-Kisa et al. (2015), the more complex EIs require a stronger connection to graders' own understanding of the content. As such, accurate grading of complex EIs requires graders to command high levels of content proficiency and familiarity. In large-scale courses, where selectivity may be limited in order to obtain the necessary number of graders, it is unreasonable to expect all graders to have the requisite content proficiency and unfeasible to develop it in a single semester. On the other hand, with so many graders available, the strongest could be leveraged to focus on grading the more complex EIs.

6.2.2 Assigned tasks

Assigned tasks can vary across many factors, including open-endedness. Some of this possible variability existed in the problems used in this study. Tasks such as creating a flowchart from scratch or writing complete code without a provided flowchart were significantly more open-ended than directly converting a provided flowchart to code or identifying the output for test cases given a flowchart with only limited possible outcomes. As the data showed, grading of open-ended tasks deviated further from the work-as-imagined instantiations and employed holistic strategies more frequently than the more closed-ended tasks. Further, overall LO scores assigned for the more open tasks disagreed more frequently with the definitive marks. These findings support Suto and Nádas (2010), Black et al. (2011), and Menéndez-Varela and Gregori-Giralt's (2018) assertions that questions, themselves, affect grading consistency and that less constrained, more objective questions with a wider range of complex acceptable answers are less reliably graded.

Related to problem open-endedness is the comprehensiveness of the "correct" solution. As a problem becomes increasingly open-ended, it becomes increasingly challenging to create a reference solution that encompasses the full range of acceptable responses. As a result, there is a

very strong negative correlation between the two variables ($r = -.883, p < .001$). Despite this connection, the model comprehensiveness only related strongly to the agreement with the overall LO score, in the opposite direction than one might expect. The more comprehensive the solution, the more likely graders were to disagree. This is difficult to reconcile but may be due to the fact that graders rarely relied much on the provided solutions given that the rubric included a “correct response” in the “What to grade” portion of each rubric, and these two solutions did not always perfectly align. For instance, for LO 1, the portion “solution” shown in the “what to grade” portion of the rubric was more comprehensive than the official solution—the official solution gave an example of a correct answer, while the rubric showed the entire range of acceptable values for different inputs. This latter point illustrates that, though the particular problem was only mildly open-ended, there are ways to write solutions that more-or-less address a range of acceptable responses.

The difficulty of the problem also contributed to consistency. The three problem contexts were estimated to be of variable difficulty for the students to understand based on the clarity of their descriptions, their expected familiarity to the students, and their required pre-requisite knowledge. The third problem was the hardest to understand as it relied on understanding content the students had just recently learned, had a moderate level of clarity in its description, and was in a context the students were not likely to be familiar with (i.e., specific features of contact lenses). On the other hand, problem 2 may have been the least clearly written, but likely familiar (i.e., atmospheric layers they were likely exposed to in earth science in grade school) and required the least background knowledge or skills. While results may have been skewed and confounded by only having three problems, the data suggest that greater understandability of the problem leads graders to stray further from the work-as-imagined grading process and more regularly apply a holistic approach. However, like LO clarity, it also associated with a stronger agreement with the definitive marks for both EI and LO, possibly for the same reasoning. Greater context understandability for the student also means greater understandability for the graders, who are not far removed from being students themselves, which may allow for more confidence to make intuitive judgments of performance. On the other side, Suto and Nádas (2010) speculated that more difficult problems require the graders themselves to apply deeper understanding and knowledge to grade, which can hurt consistency.

6.2.3 Grading guidelines

There are several features of grading guidelines that can contribute to consistency, including discriminability between performance levels. The rubrics in this study discriminated between performance levels relatively well across all items, and the discriminability did not impact the overall outcome of the system significantly. However, weaker discriminability did correspond to deviation from the work-as-imagined instantiations, and graders occasionally seemed to apply their own schemes for discriminating levels of performance. This was demonstrated most clearly in LOs 2 and 7 where one of the performance levels was disallowed despite the potential for student work to fall between the two allowable levels. Graders expressed frustration when the student's response demonstrated half of the EIs but received the same score on the rubric as the student who demonstrated none of the EIs. This supports Goldberg's (2014) argument for evenly spaced performance levels and Menéndez-Varela and Gregori-Giralt's (2018) assertion that too few or too many performance levels can make scoring decisions difficult.

The robustness of the grading guidelines indicates how well different student solutions are handled. In theory, it should relate not only to the open-endedness of the problem but the comprehensiveness of the model solution. Across the materials in this study, that did not appear to be the case, as the correlations between the measures were all weak. The robustness had little effect on grading behavior but did seem to have a notably negative effect on the accuracy of grading with respect to the definitive mark for both EIs and LOs. This discrepancy is not easy to reconcile. One possible explanation is that none of the LOs were glaringly weak, but the recognition of their robustness may not have been as recognizable to the graders—that is, had the graders been the ones to rate all of the documents in this study in terms of each of these variable dimensions, they may not have rated the LOs as robust as they ended up being rated. The rating of robustness of the rubrics was under the assumption of applying the admittedly strict perspective used in setting the definitive marks. Under that perspective of such a literal interpretation of the rubric, many of the EIs were either met or not (e.g., if a flowchart did not have arrows, then EIs that implicitly assumed the presence of arrows were not met). However, the graders who did not hold such an aggressive stance may have perceived the implicit assumptions as less robust, which would have affected their scoring decisions relative to the definitive mark.

The last important feature of the grading guidelines is usability. Human factors play an important role in rubric design. Throughout the interviews, the LOs with the strongest usability

corresponded to the greatest similarities to the work-as-imagined processes. When rubrics had usability issues, there were specific instances where crucial pieces of information were skipped (e.g., LO 9, where the vast majority of graders overlooked the specific lens ID they were supposed to evaluate; LO 10, where the majority overlooked which code they were supposed to evaluate). Some graders even admitted that when they see too much information on a rubric, such as a large number of evidence items, it feels like a “wall of text” and they tend to ignore it, which likely occurs more frequently in natural settings where the graders are not being directly observed. This makes sense when considering that humans have limited sensory abilities, attention, focus, expertise, and working memory, possessing a cognitive load capacity of five plus or minus two items (Sharit, 2006; Sweller, 1994). It also supports Joe et al. (2011) and Sadler’s (2010) recommendations to limit the amount of information on a rubric in terms of the number of criteria or performance levels.

Despite the evidence that rubric usability improves adherence to an appropriate grading procedure, greater usability also corresponded to increased disagreement with the definitive EI and LO marks, to a rather considerable degree. Like LO clarity and understandability, the more usable rubrics that featured less information and utilized color to highlight key information may have instilled a false sense of confidence in the graders and led them to make careless decisions; however, these results may be conflated. Part of the usability of a rubric was the amount of information on the rubric, which corresponded to the number of evidence items. If there is some probability of error on any given EI, fewer EIs means that a mistake in scoring one EI is more likely to affect the overall scoring decision. Thus, more usable rubrics provide less room for error.

6.2.4 Student work

The typicality, clarity, and overall quality of the student responses contributed considerably to the system. These three factors partially relate to one another. Even though they represent distinct constructs, a less typical or less clear response tends to also be lower in quality. Still, graders deviated from the work-as-imagined process for more typical, less clear, or extremely high- or low-quality work, often adopting a holistic approach when a response was immediately discernable as very good or very bad. Clarity did not correspond to adopting a holistic approach, as a less clear response required additional scrutiny to understand. Highlighting this trend, the majority of instances of confusion observed occurred when grading unclear work. Also, overall,

agreement with the definitive LO mark was significantly higher for clear and, even more so, for high- or low-quality work, supporting the findings of Cooksey et al. (2007), Black et al. (2011), and Russel et al.'s (2017) related to work quality and consistency.

6.3 Differences Between Imagined and Completed Grading

The work-as-imagined instantiations were created by applying a strict, literal interpretation of the rubrics with the intention of thoroughly and rigorously attending to every aspect of each sample to fully evaluate it with respect to every aspect indicated within the rubric. This decision was two-fold. First, the graders vary significantly in their knowledge of the content and their grading experience and are, in terms of a hierarchical organization, serving under the GTA, instructor, and the course coordinators. As such, any non-literal interpretations of the rubric (i.e., taking any liberties with the wording of EIs, such as ignoring the use of the word “arrow” to look at the more relevant, operative feature of the EI that is the connectedness of flowchart elements), makes an assumption about autonomy that cannot be assumed consistently across graders. Any assumption about the level of autonomy would be arbitrary, and the most deferent of graders will assume no autonomy. As such, it makes the most sense to assume the latter mentality. Second, it is assumed that the underlying purpose of grading is not to assign a numeric grade but to provide the student with the formative feedback necessary to enhance learning. As such, while it will result in the same grade, and is more efficient, for a grader to skip all remaining EIs the moment the students has failed to demonstrate enough EIs to put them in the ‘Insufficient Evidence’ performance level, doing so would deprive the student of all possible feedback on their work. Thus, the 30 work-as-imagined instantiations for all LO-sample pairings are likely to be more thorough than anyone could reasonably expect from a grader in any experimental or natural setting.

Given the excessive or unreasonable thoroughness of the work-as-imagined instantiations, it is unsurprising that graders’ processes only matched the work-as-imagined instantiations in approximately one-quarter of the work-as-completed instantiations. This section is devoted to exploring the literature as a means to explain this observed deviation from work-as-imagined and to speculate the potential implications. Thus, this section primarily presents what the literature says about erroneous actions, generally, and about grading, more specifically to support the observed results.

There were many erroneous actions performed across the entirety of the think-aloud interviews. These erroneous outcomes could be attributed to any number of sources, depending on contextual variables or the nature of the action itself. The literature states that in the context of human performance of cognitive tasks, humans are highly prone to erroneous actions as a result of a large number of possible internal and external factors (Sharit, 2006). At a macrocognitive level, erroneous actions can relate to detection or noticing, understanding or sensemaking, decision making, action implementation, or team coordination (Liu et al., 2017). Mistakes in detection or noticing are either internally due to inherent limitations in cognitive capacity (Sharit, 2006; Sweller, 1994) or externally due to weaknesses in the design of the objects in use (Sharit, 2006). Issues with understanding and sensemaking may be internally due to weaknesses with background knowledge, excessive cognitive demand, or secondary to issues with detection or noticing (Liu et al., 2017; Sharit, 2006; Smith & Stein, 1998) or externally due to an organizational failure to provide adequate information or training (Hollnagel, 1998). Erroneous decision making could result from internal issues with any of the previous macrocognitive processes or from emotional complications (Lerner et al., 2015), which can be externally influenced by emotion-inducing circumstances or unclear communication of organizational expectations (Hollnagel, 1998). Action implementation is subject to all the same threats as the previous macrocognitive processes but can suffer externally due to poor object design or unsatisfactory work conditions (Hollnagel, 1998). Finally, team organization is purely a consequence of organizational actions, such as unclear communication or support (Hollnagel, 1998).

The functions identified through the FRAM help to draw connections between the disparate literature associated with grading and human cognition to highlight how erroneous actions occur. The alignment between Liu et al.'s (2017) macrocognitive functions and Suto and Greatorex's (2008) cognitive strategies of graders that were elaborated upon by the model in this study is no coincidence. The first two macrocognitive level actions subsume *scanning* and *matching*, and *evaluation* and *scrutinization*, respectively. The third level corresponds to the added function of *judging criterion satisfaction*. Further, the fourth level can be attributed to the external forces that act on the grader as they grade (e.g., other pressing obligations, fatigue, discomfort), but also relates to the design of course materials provided by the course coordinators and the support provided by the graders' GTAs. Meanwhile, team coordination issues relate entirely to the organization and support provided by both GTAs and the course coordinators.

Given these parallels, it is no wonder that so many of the possible erroneous actions were observed in this study or can easily be recognized as likely outcomes of authentic implementations of the model (i.e., real-life grading instances). The variability of the background functions illustrates many of the ways that the course coordinators' or GTA's actions affect the likelihood of graders making various erroneous actions. For instance, if the course coordinators design a rubric to contain amounts of information that overload the graders' cognitive load capacities, it is more likely that they will commit an erroneous action involving detection or noticing. If the GTA and course coordinators do not support the graders in their training, they could easily commit erroneous actions with understanding or decision making.

Throughout this study, there were numerous times when graders overlooked information within a rubric or within a sample response or decided something matched that did not actually match (i.e., erroneous outcomes for *scanning* or *matching*/noticing or detecting). On the other hand, there were also numerous occasions when graders misinterpreted an EI or some aspect of a sample response (i.e., erroneous outcomes for *evaluating* or *scrutinizing*/understanding or sensemaking). Further, there were instances when graders made incorrect decisions based on what appeared to be appropriate interpretations of the right portion of work (i.e., erroneous outcomes for *judgment* or decision making). These poor decisions were most prevalent with LO 10 ("adhere to programming standards"), where several graders explicitly stated that they grade programming standards "leniently" and based on a "good-faith effort." Such perspectives are undoubtedly the result of expectations communicated by their instructor or GTA. Overt and intentional subversions of expectations expressed by the rubric, which could be acts of emotional defiance (Lerner et al., 2015), are not considered by Sharit (2006) to be erroneous actions; however, adopting the broader view of the system as a whole, it is an erroneous action within the context of designing and supporting elements to produce the intended outputs.

Another major trend observed across the think-aloud interviews was the use of holistic approaches to grading. The human factors literature acknowledges that in response to excessive cognitive demand or cognitive load, such as lengthy or overwhelming protocols, humans often resort to heuristics in an effort to save effort and time (Sharit, 2006). In the context of grading, this explains why graders in this study and as described throughout the literature often rely on pre-existing cognitive frameworks (Joe et al., 2011), holistic strategies (Bloxham et al., 2011; Hay &

Macdonald, 2008), overall impressions of a student's work (Humphry & Heldsinger, 2014), or, worst of all, general impressions of the specific students themselves (Meier et al., 2006).

When comparing all of the grading data from the interviews, holistic approaches were nearly indistinguishable from analytic approaches in terms of agreement with the definitive mark. This middle-of-the-road result reflects the disagreement about holistic grading within the literature. Sharit (2006) claims that heuristics tend to be more error-prone, supporting Baird et al.'s (2017) observations that analytic grading leads to lower mean absolute score differences than holistic grading. However, Barkaoui (2011) found stronger interrater agreement with holistic rubrics. That said, when modeled with other variables to account for variability, holistic grading was, in fact, correlated with greater disagreement for both EI and LO scoring decisions with respect to the definitive mark.

While there are inconsistencies in reports of consistency of grading with holistic approaches, again adopting the perspective chosen in establishing the work-as-imagined instantiations and definitive marks, there are reasons why the use of holistic grading is concerning. In this study, some graders who used holistic strategies would return to the EI list and check off items, sometimes skipping over various EIs, which varied from instance to instance and grader to grader. This observation, once again, echoes the literature. Joe et al. (2011) observed that graders occasionally attend to aspects of a student's work that are not stated within the rubric. This can save time but may lead to inconsistency across graders. Further, even within the set of explicated features to evaluate (i.e., the EIs), graders tend to focus on their own unique subsets (Orr, 2002). In Joe et al.'s (2011) estimation, graders typically focus on no more than three rubric features when employing holistic strategies. Thus, holistic scores between graders may represent entirely different aspects of proficiency.

The use of the holistic approach over the analytic approach appeared to relate to multiple factors. Holistic grading occurred most frequently when a response was very high or low in quality and when the problem was more understandable. This may explain the observed agreement rates—graders employed the more error-prone holistic approach on the easier to grade samples, moderating one another. The use of holistic grading also increased over time during each interview for every participant. This trend to increasingly grade holistically over time coincides with Joe et al.'s (2011) observation that graders' attention to specific details of the rubric varies over time and with experience. While the experience of graders in this study was not tracked, it was certainly

true that some graders were far more likely to employ holistic approaches than others. That said, Interestingly, there was not a clear association between holistic grading and LOs or EIs with higher cognitive loads or demands. This differs from Joe et al.'s (2011) finding that graders tend to abandon the rubrics for holistic approaches when rubrics become overly complex or demand the graders to consider multiple pieces of information simultaneously.

There is one last thread to address with respect to variability across instantiations that could not be reasonably observed in this study but is likely very relevant: differences in personalities and situational circumstances. Grooten and Suto (2006) note the importance of personal factors that influence grading, such as teaching and grading experience, experience grading a specific assignment, personal preferences, deference to authority, or attention to detail. On the other hand, Sharit (2006) notes all the circumstantial personal factors (e.g., fatigue, distraction) and external environmental factors (e.g., discomfort) that can influence the likelihood of committing erroneous actions. These factors may be ubiquitous to grading and are significantly harder to control, so it should be expected that some amount of erroneous action is inevitable and unavoidable, especially as the size of the system increases and becomes inherently more complex.

6.4 System Resilience

The overall resilience of the system is less than desirable. The final output of the system agreed with the definitive mark only 47.8% of the time, despite the fact that, overall, agreement at the EI level was 79.7%. Thus, even though the LO score is effectively the sum of EIs, which have a fairly acceptable agreement rate, the final output of the system has an unacceptable agreement rate. Even if one were to dismiss the definitive mark as being too harsh and embrace the most agreed-upon system output scores across the interviews as the “correct” definitive scores, the overall agreement of system output would only increase to 62%. Meanwhile, adopting the same standard for the EIs, agreement for EIs would increase to 86.8%. Thus, either way, there is a significant drop in consistency of the system output from the near last to final output.

This limited resilience should not be unexpected. If the rate of erroneous decisions for EIs is about 1 in 5, one would expect any LO with five EIs to have at least one incorrectly identified EI. For LOs with only five EIs, performance levels are differentiated by a single EI. As such, unless the grader makes two mistakes in opposing directions (i.e., grades one too harshly and one too leniently), it is expected that graders will, on average, incorrectly grade LOs with five EIs.

When fewer EIs are present, the expected number of incorrect EIs decreases, but the score difference may be greater (e.g., LO 2 had a 0.5-point swing for each EI). On the other hand, LOs with many EIs had some built-in buffer for error. LO 10's 10 EIs allowed graders to, in some circumstances, but not all, make one EI mistake and obtain the same overall LO score, as some performance levels spanned demonstration of multiple EIs (i.e., achieving seven or eight EIs resulted in the same performance level).

As this system is so large and interconnected, it would be unwise to make overall conclusions based on the observed data. But the model does demonstrate how variability in the output of one of the early functions, such as articulation of an LO, can affect many variables and very clearly aggregate throughout the system. If an LO were to be unclear, it could align with assigned tasks. It could affect interpretation by the teaching team, which, in turn, could affect the way it is communicated by the teaching team to the students. If the message delivered by the teaching team conflicts with the student's own interpretation of the LO, which also conflicts with the assigned tasks, this can cause considerable confusion for the student, who could fail to learn the content properly and perform the assigned task in an unpredictable way. This could then be graded by one grader in a completely different way than it might be graded by another grader who interpreted the LO differently from one another. Though this particular type of instance was not and could not have been observed in this study, the model allows one to see how it could happen.

Another important note is that of the grades that disagreed with the definitive mark erred disproportionately on the lenient side—of EI disagreements, 499 out of 643 (77.6%) were too lenient, and of LO score disagreements, 185 out of 232 (79.7%) were more lenient. Diefes-Dux et al. (2010) observed similar issues with leniency. It is reasonable to assume that grades that are too harsh will lead to grade change requests by the students, while grades that are too lenient are extremely unlikely to be reported or result in a regrade request. This outcome sends a misleading signal to the instructor that students understand the content better than they do when an instructor may need to re-address content that was not adequately covered. It also drives grade inflation, as errors are not uncommon and tend toward the positive direction. Finally, it is crucial to note that this trend results in unfairness to students who may already feel disenfranchised in the “chilly climate” of engineering (Malicky, 2003). The assumption that students will request regades when graded too harshly is likely a weak assumption—students with more timid personalities have been shown to react differently to critical feedback based on feelings of anxiety or views toward

authority (Garza & Lipton, 1978). As a result, instructors need to be well aware of this potential bias.

6.5 Dampening mechanisms

Hollnagel (2012) presents four general approaches to adjust a system upon variability analysis: (1) elimination, or removal of a part of the system that is producing unfavorable variability, making the variability impossible; (2) prevention, or adding a barrier to keep the unfavorable outcome from occurring; (3) facilitation, or redesigning the system to make correct use easier or incorrect use more difficult, such as decreasing task complexity or providing operational support; and (4) protection, or improving outcomes after an unfavorable outcome occurs. Some useful mechanisms are monitoring performance indicators, implementing interventions to reduce internal variability of functions, or decoupling functions that frequently see down-stream upstream amplification of variability. Based on these principles, analysis of materials, observations from interviews, and statistical analyses, a table is presented in each of the following sub-sections outlining relevant recommendations to dampen variability in the grading system.

One important note is that many of these challenges and recommendations are compounded by the size of the system. As many graders become involved, it becomes increasingly difficult to control and reduce potential error. Some error will always be an inevitable consequence of human limitations and fallibility. As a result, the purposes of these mechanisms are to impose as much control over the system as possible, to prevent erroneous outcomes generating as a result of the system, rather than despite the system, and to include barriers to prevent erroneous outcomes from making it through the system. Still, it can be expected that some erroneous outcomes will occur, so students should be empowered to advocate for the accuracy of their grades when they feel they have been mis-graded, without fear of negative consequence or retribution.

6.5.1 Course content

Figure 6.1 shows recommendations for designing course content based on this study's observations and a synthesis of the literature. Some aspects may be applicable or adaptable to other contexts. Regardless of the context, it is important to identify course learning objectives and

articulate them in a way that is observable and measurable. That does not mean the rubrics need to fully resemble those included in this study, where an overall proficiency level relates to an enumeration of demonstrated evidence items. An alternative format may be to allow for indication of achievement of each evidence item separately, which would increase differentiation between score levels and remove a level of decision making for the graders while giving more specific feedback to the students. Further, there is no reason these need to be limited to dichotomous achievements. If it is impossible to break an observable behavior into two distinct performance levels, break the performance into however many levels can be consistently and uniquely differentiated. Either way, it will be helpful to develop a streamlined set of learning objectives that are conceptually distinct and can be clearly interpreted and measured by all members of the system who will need to interpret them.

6.5.2 Assignments

Figure 6.2 presents a set of observations and literature-based recommendations for the design of assignments to reduce variability. As the data showed, grading of more open-ended tasks deviated further from the work-as-imagined instantiations and were associated with an increase in the likelihood of holistic grading. Further, overall LO scores assigned for the more open tasks showed greater disagreement with the definitive marks. This disagreement for open-ended tasks poses a considerable challenge given that engineers typically encounter “wicked” or unstructured, open-ended problems (Rittel & Webber, 1984). As such, high-quality education of engineers demands the development and assessment of open-ended tasks (Darling-Hammond et al., 2013; Hansen, 2011). As a result, this may be a necessary trade-off in engineering coursework. It will be up to the teaching team or the IST to determine what amount of inevitable variability is acceptable for the benefit of providing authentic experiences to the students.

Function	Recommendation
Articulate learning objectives	<ul style="list-style-type: none"> • Ensure the learning objectives are easy to understand for all agents within the system and fully capture the construct they represent. <ul style="list-style-type: none"> ○ Write LOs as a teaching team and seek feedback ○ Ask graders and potential students to read the LO aloud and to explain their interpretation and what skills they believe it encompasses—modify if interpretation is incorrect ○ Avoid LOs that encompass too many sub-behaviors—if it spans too many sub-behaviors, break it into smaller LOs ○ Ensure that LOs are conceptually distinct from one another ○ Avoid an excessive number of LOs—identify those which are most important for the course and avoid redundant or unnecessary LOs ○ Remove grading system that only allows selection of overall LO score in favor of system that can evaluate at the EI level
Articulate evidence items	<ul style="list-style-type: none"> • Identify a small, yet complete, set of observable, measurable behaviors that represent achievement of the LO <ul style="list-style-type: none"> ○ Write EIs as a teaching team and seek feedback ○ Ask graders and potential students to read each EI aloud and to explain their interpretation—modify if interpretation is incorrect ○ Ask graders if they associate achievement of the LO with any other behaviors that are not included in the EI list—add to list if relevant or clarify description of LO to remove possible interpretation ○ Avoid an excessive number of EIs—if LO encompasses too many, break the LO into smaller outcomes ○ Ensure that the EIs are conceptually distinct from one another ○ Avoid multi-dimensional EIs that could complicate determination of achievement or allow for partial achievement in the grading system ○ Avoid vague terms or give them clear, measurable definitions (e.g., “minimal” hardcoding means <10% of assigned variables are hardcoded)

Figure 6.1. Recommendations for design of course learning objectives and ancillary content to reduce grading variability

6.5.3 Grading guidelines

A list of grading guideline recommendations is included in Figure 6.3. It is important to note, however, that the specific grading scheme used does not need to resemble the one demonstrated and analyzed in this study. In anything, simplification of the grading scheme is preferred as the complexity of the grading scheme and rubric is one of the primary threats to validity, as it contributes to graders choosing to abandon the scheme in favor of their general impressions of the work (Joe et al., 2011). It is encouraged to explore new alternatives to guide grading that dampen sources of variability. For instance, large courses have employed mixtures of

self- and or peer-assessment to some degree of success (Jonsson & Svingby, 2007). The reliability of grades for peer-grading has been shown to correlate with instructor scores when four or more peers grade; meanwhile, the process reaps the benefit of helping students to internalize the content (Jonsson & Svingby, 2007; Schunn et al., 2016).

Function	Recommendation
Select appropriate task	<ul style="list-style-type: none"> • Ensure the task and the LOs intended to measure the task are fully aligned <ul style="list-style-type: none"> ○ Ensure that the LO has been/will be clearly communicated to the students ○ Ensure that any or all LOs associated with a task have been taught in alignment with expectations ○ Ensure that all LOs are assessed through a task ○ Ensure that LOs are conceptually distinct from one another ○ Decide, as a teaching team, how open-ended tasks should be with recognition that open-ended tasks may be inherently more difficult to grade consistently • Consider including some closed-ended tasks and grading them automatically, if it aligns with course learning objectives
Develop task	<ul style="list-style-type: none"> • Develop task context <ul style="list-style-type: none"> ○ Poll students at the beginning of course to have a sense of contexts that are or are not familiar and content that is or is not known—alternatively, base this information only on required prerequisites ○ If a context is chosen that is unfamiliar or requires background knowledge, take the time to introduce the context in class and teach any background knowledge that will be needed to ensure all students are on relatively equal footing. ○ Ensure that the task description can be clearly understood by students or devote time in class to checking in with students on their interpretation to ensure consistency • Develop task instructions <ul style="list-style-type: none"> ○ Decide, as a teaching team, on an acceptable amount of scaffolding to provide to the students in a problem, noting the trade-off that increased scaffolding reduced authenticity and decreased scaffolding with increase the variability of student responses ○ Ensure that task instructions are easily understandable to all members of the system

Figure 6.2. Recommendations to support task development to reduce grading variability

6.5.4 Grader training

While grader training was not specifically observed in this study, graders did express some perspectives about training throughout their interviews. Their sentiments, along with needs

identified through observation of the system and ideas presented throughout the literature, were synthesized to produce the recommendations in Figure 6.4.

Function	Recommendation
Develop model response	<ul style="list-style-type: none"> • Create a solution model that encompasses the widest range of acceptable responses (that is, don't just create a sample of a solution with one possible acceptable response) <ul style="list-style-type: none"> ○ Collect past student work on similar problems, particularly in the mid-level range ○ Ask graders to attempt problem ○ Collect responses to develop model solution that articulates the parameters or descriptors of acceptability ○ Create “model” solutions for each level of performance for each LO <ul style="list-style-type: none"> ▪ This could be done using snippets of responses to highlight the differences between achievement and non-achievement ▪ Can include descriptions to help guide differentiation between performance levels • Clearly articulate an appropriate number of performance levels <ul style="list-style-type: none"> ○ Use the same resources to develop an encompassing solution model to identify the range of possible responses from students ○ Based on responses expected from students, make sure that distinct levels of performance for any given evidence item is represented so graders do not feel a performance falls in an unavailable level between others (can vary across evidence items) <ul style="list-style-type: none"> ▪ Couple performance levels with examples that illustrate each level
Assemble grading guidelines	<ul style="list-style-type: none"> • Assemble grading guidelines to optimize usability <ul style="list-style-type: none"> ○ Reduce the cognitive load on graders—keep the amount of information communicated to the bare minimum of essential information ○ Highlight the most important information to draw attention and help users find the information quickly when needed ○ Build a more holistic view into the grading guidelines—rather than designating a single, specific instance of a learning objective performance to evaluate, allow for the evaluation of the skill generally over the entirety of the response ○ Ask a few graders to read through grading guidelines and to apply to a mid-level sample to identify potential weaknesses <ul style="list-style-type: none"> ▪ Observe use to ensure graders are attending to all necessary features • Streamline the process as much as possible

Figure 6.3. Recommendations for the design of grading documents and guidelines to reduce grading variability

Function	Recommendation
Identify training problems and samples	<ul style="list-style-type: none"> • Select representative examples to use for training <ul style="list-style-type: none"> ○ Avoid overburdening graders with excessive training that will burn them out or lead to unauthentic engagement ○ Rather than providing entire problem length responses, select small portions of work to illustrate individual evidence items ○ Provide examples of each performance level of a given evidence item—indicate the intended performance level and provide explanation ○ Use a variety of problem contexts to help graders understand how to generalize evaluation of the constructs across different problems • Deliver training to graders in a way that is reliable, consistent, and effective <ul style="list-style-type: none"> ○ Train in group settings—include a GTA (or very experienced grader who may be more knowledgeable of the system than an inexperienced GTA) as a group leader <ul style="list-style-type: none"> ▪ Discuss grading guidelines to achieve a consistent interpretation ▪ Discuss samples of work along with how and why they should be assigned the recommended performance level
Deliver training	<ul style="list-style-type: none"> ○ Train on the same LOs a few times throughout the semester to maintain and refresh interpretation—note that this would become overburdening with an excessive number of LOs ○ Approach grading as a process and attempt to have graders follow a consistent process ○ Calibrate decisions and make recommendations for providing feedback ○ Give graders specific, individualized feedback through training sessions that help them learn based on their needs

Figure 6.4. Recommendations for design of grader training to reduce grading variability

6.5.5 General organizational policies

Figures 6.5 and 6.6 provide a list of recommendations for grading procedures and organizational policies that influence the grading process, respectively. The list is based on sources of variability in the system, either due to internal function variabilities or external influences on functions, with respect to a synthesis of recommendations in the literature.

Function	Recommendation
Assign grading	<ul style="list-style-type: none"> • Select an appropriate grader for each problem or LO <ul style="list-style-type: none"> ○ Despite training all graders for all LOs, reserve more expert or experienced graders for the higher order LOs or EIs ○ When possible, grade “horizontally” rather than “vertically”—that is, one grader grades all students for a subset of problems or LOs rather than all problems for a subset of students ○ Oversee graders, tracking metrics such as average scores assigned ○ Spot check assigned grades that occur at variable ranges of performance, but focused on mid-level scores ○ Based on metrics, give graders feedback if they are being consistently lenient, harsh, or erratic
Communicate expectations	<ul style="list-style-type: none"> • Clearly communicate expectations for and purposes of grading <ul style="list-style-type: none"> ○ Describe how the grading guidelines are expected to be used ○ Indicate the degree to which graders have the authority to exercise autonomy and apply judgment versus follow the guidelines precisely ○ Emphasize the purpose of the grading—ideally, even in summative contexts, it is still important to provide feedback to students of their weaknesses in order to learn and improve (even on final exams—the goal of student learning should never end) ○ Grading is for the students, and feedback on performance is most effective when it is prompt, but should not correspond to sacrificing quality or accuracy • Teach graders to be aware of the effects of emotions on grading decision making <ul style="list-style-type: none"> ○ Encourage self-awareness and self-monitoring of emotions, noting their effects on grading ○ Avoid carry-over of incidental emotions from previous experiences by not grading when experiencing significant emotions (positive or negative) ○ If identifying a student response that triggers an emotional response, notify GTA to address issue and/or move to next student and return later

Figure 6.5. Recommendations for grading procedures to reduce grading variability

Function	Recommendation
Provide supports	<ul style="list-style-type: none"> • Provide a comfortable environment for graders to grade <ul style="list-style-type: none"> ○ Offer a location with multiple monitors to help graders reduce cognitive load of switching between screens ○ Keep environment comfortable (i.e., no overpowering odors, comfortable temperature, work-conducive atmosphere) • Encourage graders to communicate their circumstances and offer support when possible <ul style="list-style-type: none"> ○ Ask graders to indicate when overwhelmed with other obligations so reassignment of grading can occur • Maintain a supportive atmosphere with safe and open lines of communication <ul style="list-style-type: none"> ○ Encourage asking questions and admitting when uncertain ○ Respond promptly to questions to reinforce the behavior
Collect data and revise	<ul style="list-style-type: none"> • Collect feedback and data throughout the semester on all course documents—do not wait until the end of the semester <ul style="list-style-type: none"> ○ Encourage instructors, GTAs, and graders to log reflections on different course materials (e.g., was time sufficient? were course materials unclear? what worked effectively? what did not work effectively?) ○ Collect samples of student work to supplement examples of performance levels for training ○ Identify EIs or LOs that were demonstrated or not demonstrated by the vast majority of students to flag for revision or removal • Collect survey data from students about specific aspects of the course that worked well or did not work well, possibly including questions at the end of individual units specific to a subset of assignments • Make revisions based on data and feedback <ul style="list-style-type: none"> ○ Based on collected data, identify course materials (i.e., LOs, EIs, assignments, grading guidelines) that could benefit from improvement and supplementation and revise to reduce issues ○ Keep a database of archived versions of documents with annotations regarding feedback or data ○ Be careful to apply consistent updates across all documents affected by modifications (i.e., a change in the assignment may require updates to the grading guidelines)

Figure 6.6. Recommendations for organizational policies to reduce grading variability

6.6 Generalizability of Findings and Recommendations

There are two primary considerations with respect to the generalizability of this study: (1) the ecological validity of the collected data, and (2) the applicability of the model, observations, and recommendations for contexts that differ from the context studied. There were definite limitations in the data collected for this study that require some of the findings reported in Chapter

5 to be interpreted with caution—the strongest takeaways are based on the qualitative observations rather than any statistically-based arguments. Meanwhile, the model that is presented in Chapter 4 was a large-scale synthesis of multiple bases of literature, observations, and experience with the system as a whole, and specific qualitative observations of graders grading. As such, there should be few concerns about the ecological validity of the model itself, with the data collected for Chapter 5 providing more of an illustration of possible system instantiation variability rather than an encompassing description of how the system definitively varies. With that in mind, while the model was built within a specific context that is, most likely, more complex than most course contexts, the model over-specifies concerns of variability for most contexts, but the concerns of smaller contexts should be subsumed by and visible within the model.

The model itself was synthesized using an extremely large set of information, grounded in personal experience, academic literature on human factors, cognitive capabilities, decision making, grading systems, grader cognition, and factors that affect graders, specifically. The model was developed in coordination with an undergraduate assistant who had been a student in the system and served in teaching assistant and grading capacities in the course studied for three years. Thus, even though the observations of grading through think-aloud interviews were arguably artificial, the majority of the data that contributed to the development of the model was based on empirical research and experience with the natural system. This satisfies Hoc's (2001) general description of ecological validity for cognitive engineering studies.

The data collected through think-aloud interviews that were used to identify model instantiations and make conclusions about significant sources of variability in the model did suffer a few limitations. First, because the model was built after the interviews occurred—by necessity, as observations helped to identify the system's functions—the problems, rubrics, and samples of student work used in the interviews did not optimize the potential for inference making. Given that nearly three dozen variables were identified in the system, the samples that were purposefully selected to illustrate variable solution approaches did not effectively vary the variables ultimately identified to allow for adequate isolation of the effects of individual variables. Observed variability was a product of the elements of the interview rather than being a true representation of the entire span of variability for each function. Further, any observed trends for some variables were potentially inextricable from other variables. For example, the accuracy of model solutions varied by an extremely small degree. Most were perfectly accurate, two had minor typo-level errors that

were likely not even noticed by graders, and one completely lacked a solution. This small distinction between the most extreme cases created a variable that overemphasized differences between variables and made it difficult to discriminate between the effect of the solution accuracy rather than effects from many other variables. These challenges are all in addition to the inherent limitations imposed by the think-aloud process's potential alteration of actual cognition.

One important thing to note with respect to the value of the findings in Chapter 5 is that because the data was collected from only three responses to items on a single assignment in a class that was developed by many highly knowledgeable and experienced faculty, they cannot be expected to represent the most significant sources of variability in every system. A source of variability being identified as impactful in this system may have been an artifact of the assignment and samples selected. On the other hand, a function not being identified as an impactful source of variability does not, in any way, suggest that the function is not meaningful to the system. This study had no way to observe variability in the schedule setting functions; however, setting a far too restrictive schedule without the flexibility to modify the schedule could have significant effects.

Despite these limitations to ecological validity, the purpose and benefit of using the interview observations to create model instantiations served more to show how the system *might* play out in practice. The observed instantiations were not expected to reflect all possible variabilities of the system definitively, but to highlight how the system varies based on contextual factors of background functions, how those variabilities aggregate within the system, and to offer insights about what functions might be more or less impactful. The specific observations of how grading is conducted, particularly the dependence on holistic approaches, provide more critical information than any statistical conclusions. Thus, the statistical analyses were primarily used to guide the discussion rather than to make any definitive claims about system variability. In this respect, they can be viewed as illustrative case studies of the system's utility.

Viewing the observed instantiations of the system as a particular set of case studies helps to highlight the generalizable utility of the system across multiple contexts. The instantiations shown in Chapter 5 illustrate how all of the pieces of the system can be analyzed and characterized. In a broader application, it might be that the LOs or EIs described as “clear” or “unclear” all fell somewhere in the middle relative to other instances. This does not ultimately matter—the value of the system is in its ability to draw attention to crucial aspects and design considerations for each of the components in the system and to recognize how erroneous decisions with regards to one

component can affect others. Those relationships will be maintained regardless of the specifics of the system. Further, because this study focused on such a large system with so many different agents, a smaller or less complex system should be entirely subsumed within the model presented. In other words, in a small class in which the course developer, course coordinator, teaching team, and graders are all one person, some of the functions may become irrelevant (e.g., training functions), but that instructor still needs to consider many of the same features of content, assignment, and rubric design. In that scenario, they would not need to be concerned with consistency across graders but would need to be concerned with self-consistency and still create a rubric to communicate expectations to students transparently. As the model was built in the context of an extremely complex system, the system can be easily reduced for simpler systems.

7. CONCLUSION

7.1 Overview

Evaluation of student learning is fundamental to student learning. It provides feedback to learners to adjust their understanding or pursue further study, to instructors to adjust their teaching, and to many other stakeholders who make important and expansive decisions based on evaluation data. Given these stakes, the importance of evaluation data being valid and meaningful cannot be overstated. This validity depends heavily on whether what is intended to be assessed, on whether the measures can be conducted reliably, and on whether the measures are fair to different populations. Reliability, at the core of validity, is challenging to achieve when evaluating performance on the open-ended tasks that are central to authentic, high-quality engineering education. This challenge is exacerbated when classes become large and a single individual cannot feasibly be relied upon to perform all of the evaluations. Tools like rubrics are used to help attain stronger reliability when multiple graders are needed, but research indicates that even with thorough rubrics, there may be large levels of unreliable application for open-ended engineering tasks. Thus, an exploration of evidence for validity, focused on sources that threaten validity, is warranted.

To explore the validity, a focus on reliable application of rubrics across many graders of open-ended engineering tasks relies on a recognition that this context involves a large system that coordinates many humans who are all prone to erroneous actions. This study frames grading as a socio-technical system consisting of an instructional support team that designs most of the course materials, a teaching team that delivers the content to students, students who learn the content and perform assessment tasks, and graders who evaluate the students' performances. Interpretation of this system requires an understanding of the different elements within it. That is, it is necessary to understand human erroneous actions and the internal and external factors that can cause them. It is also important to understand how student work is evaluated using various grading schemes and tools, like rubrics. Further, it is necessary to understand what factors have been shown to affect consistency of evaluation, including the assignments eliciting task performance, the quality of student work, specific features of the grading tools, and personal features of the graders,

themselves. Further, it helps to understand the way graders engage in evaluation of student work at a cognitive level.

Given the view of grading as a system, this study embraces the human factors engineering approach of Human Reliability Analysis to identify where unreliability occurs. Specifically, it utilizes an approach called the Functional Resonance Analysis Method that is used to model all of the actions taken within a system, how those actions interact with one another, and how the outputs of those actions might vary individually and affect subsequent actions. To make the model useful to different groups within the system and to highlight the general goals of each group in addition to each individual action taken, the model is organized using an abstraction hierarchy approach. Functions (i.e., actions) are organized by agent, or group, at different levels of abstraction (from the agent's overall general purpose, to the broad sets of actions needed to achieve those general purposes, to all the specific individual actions that must be completed). Each function is defined with respect to what it does, as well as what information it needs to take in to be performed, what controls, time constraints, preconditions, and resources or execution conditions are used during, guide, or constrain the process, and the output of the function. Every function output must be used in some capacity by another function and every aspect that affects any function must be the output of another function. Through this process, the system is modeled extensively in terms of all the different interactions between actions and how an issue with one action's output can support or hinder another.

This study built a FRAM model for the grading system involved in a course consisting of over a dozen large sections at a large midwestern university. The model was constructed based on a synthesis of literature about the topics referenced above, personal experiences in several roles within the system, analyses of many documents, and observations of graders grading samples of student work on a real assignment from the course. At the deepest layer of abstraction, 60 functions were identified across the four agent groups, most of which related to the course developers/coordinators and graders. While the frequency of each of the functions varies (some occur once per semester, others occur repeatedly throughout), the majority of the functions identified can be reliably expected to occur, but the outputs of each function can vary in ways that can easily aggregate within the system. Some of the evaluative functions performed by graders will vary depending on the context established by the outputs of the other functions. For instance, the course developer's design of evidence items for a given learning objective may require the

grader to engage in different cognitive strategies to appropriately evaluate the student's demonstration of the evidence.

Taking into consideration the full model of all functions and how every function can theoretically vary, the study also takes direct observations from think-aloud interviews with graders to build specific instantiations of the model. The interviews utilized a purposefully selected assignment from the course and purposefully selected samples of student responses to each assignment task. The assignments, rubrics, and student responses were observed to be the outputs of the background course developer and student functions in the system. Taken collectively, the outputs of these functions created a particular instantiation of the system that dictated an expected sequence of actions for the grader. More specifically, the rubric indicated parts of a student's response that needed to be evaluated and the way in which it needed to be evaluated while the student sample should have affected the interpretations and decisions the grader should have made (e.g., if the specified portion of student work was missing, the grader should have looked for the portion and identified its absence). Through this process, work-as-imagined instantiations were created for all 30 LO-sample pairings. Next, every observed grading instance across 17 interviews with undergraduate graders were coded with respect to the cognitive functions they did utilize to evaluate each evidence item and learning objective for each student sample, creating several hundred work-as-completed instantiations. To get a stronger sense of variability, the instantiations were compared with the work-as-imagined instantiations to identify the number of functions that led to unexpected outcomes, the use of unexpected functions, or the lack of use of expected functions. These were added to identify the overall deviation from the work-as-imagined. The graders' determinations of EI achievement and overall LO scores were also compared to the "definitive scores" determined by applying the work-as-imagined process. Analysis of these instantiations allowed a general sense of how variations in the background functions influenced variations in function use and output by the graders. It further demonstrated large levels of variability in the cognitive actions utilized by graders and relatively low levels of agreement of final scores for students, despite moderate levels of agreement on individual evidence items.

Collective analysis of the model and the observed instantiations allowed for a focus on which functions seemed to affect variability in the system most significantly, both in terms of how the system was enacted and the final output of the system. This process identified some of the most important functions to be the student's performance of the task, the understandability of the task

context, the open-endedness of the task, the usability of the rubric, and the clarity of the learning objectives, among others. Based on these impactful variables, in conjunction with understandings of human behavior and grading systems presented throughout the literature, a set of potential mechanisms to dampen variability associated with content design, assignment design, grading guidelines, grader training, and organizational practices. While the specific observed variabilities and some of the recommendations are strongly connected to the context of the course under study and the assignment and samples observed in the interviews, the model itself provides all of the possible mechanisms of variability that can be useful for analysis of any assignment or student work in the system. Further, less complex course contexts are subsumed within the model and can be teased out by moving layers of abstraction or removing functions that are not relevant for an alternative context.

Ultimately, this study illustrates the complexity of grading at large scales and all the different actions that must be taken and how they interact. It demonstrates that the grades of open-ended engineering tasks at large scales cannot be reliably interpreted, given an overall agreement of less than 50% across final LO scores observed in the interviews. As such, the validity and meaningfulness of grades may be less than desirable. However, the extensiveness of the model highlights the potential sources that contribute to variability and provides an effective tool for evaluating the system and generating mechanisms to improve the reliability, which can be extended to alternative systems.

7.2 Major Contributions

One of the major contributions of this study is to the understanding of grader cognition. As the majority of grader cognition studies have focused in language arts, this study is the first to explore grader cognition in the context of evaluating open-ended engineering tasks. Given that previous research has suggested some aspects that affect grading consistency, this study extends the past research to connect the actions taken by graders during grading to actions taken outside of grading. The study further extends previous models of grader cognition by developing a model that encompasses multiple previous models, contributes additional nuance to the performance of each of the previously identified cognitive strategies, and adds a new cognitive strategy as well as several cognitive actions that regularly occur throughout the grading process. Finally, by employing the FRAM, it establishes a way cognitive strategy usage can be reasonably predicted

for specific contexts and demonstrates specifically how features of the actions taken outside of grading can affect the actions performed during grading and can vary in ways that affect the reliability of the grading process.

This study ties together multiple disparate sources of literature. It integrates literature about course design, course materials, grading systems, and rubrics with grader reliability and grader cognition and with systems engineering, human factors engineering, and cognitive engineering. In this way, it demonstrates not only the cognitive strategies graders utilize, but also the ways those strategies may go wrong and why. This provides more of an empirical and theoretical basis for designing course materials such as assignments and rubrics.

The study also shows that grades of open-ended engineering tasks at large scales may not achieve the levels of reliability that should be expected. However, the model serves as a useful tool that can be used specifically for large systems like the one analyzed in this study but can be easily adapted for any course context. The tool is useful in reminding a course designer of all the features that must be attended to during the design of course materials and how inattention can affect the system. It can also serve as a retrospective tool to backtrack from an undesirable outcome to try to identify the source of variability.

Finally, this study demonstrates a novel connection between human factors engineering and engineering education. Applying the FRAM to the grading process shows the utility of applying methods to study human actions within an educational context to improve the design of the components of the education system. While FRAM and other human reliability analysis methods are often used in more industrial contexts with higher stakes events, it can be applied to study how different actors within the education system function and find ways to optimize performance.

7.3 Practical Implications and Recommendations

This study offers a set of recommendations that can be used by course designers and instructors to improve reliability of grading in large-scale engineering courses. However, the recommendations can be reasonably generalized for instructors in smaller scale contexts as they design their own course materials. Attempts to improve reliability of grades should be viewed as a continuously iterative process that should be receive data-driven revisions each semester. Through improvement of reliability, grades can be viewed as more meaningful for students,

instructors, administrators, and policy makers. Indirectly, it may help to reduce perceptions of unfairness or the chilly climate of engineering that tends to perpetuate underrepresentation of some groups in engineering.

Ultimately, the recommendations can be simplified as measured to reduce cognitive demand, cognitive load, and emotional reactions in graders as a result of issues with interactions between student work and course materials. Learning objectives need to be clearly articulated and operationalized, conceptually distinct, and streamlined. Complexity in the form of difficult-to-understand text, excessive numbers of learning objectives or evidence items can threaten reliability by overloading the grader cognitively, leading the grader to employ more holistic approaches rather than focus on the details they are expected to evaluate. Any time a grader is prompted to adjust their behavior, variability can occur, as all graders are different. Thus, designing rubrics to be user friendly with minimal information, providing encompassing solutions to minimize judgments needed to be made, and ensuring that learning objectives will not repeatedly punish students for individual mistakes help graders to not feel overburdened or react emotionally. The more comprehensive list of recommendations is included in Section 6.5.

Despite measures to improve reliability, it is also important to emphasize that to err is to be human. As long as humans are involved in the grading system, erroneous outcomes are inevitable (and, perhaps, erroneous outcomes should be anticipated even when humans are not directly involved). Purely due to the statistics of expected outcomes, as the size of a system increases and more graders are involved, the variability will increase, and the number of erroneous outcomes will scale up along with it. As undesirable outcomes are unavoidable, it is important to establish a culture that accepts that grading errors can and will occur and encourages students to feel safe disputing such outcomes. A reduction in undesirable outcomes may be achievable by following the recommendations in this study—a goal that should be strived for to improve overall grade validity—but they can never be expected to fully disappear.

7.4 Future Work

There are several natural follow-ups to this research. The instantiations observed through think-aloud interviews were an artificial context that may have affected the behaviors that were demonstrated and the grading decisions that were made. For instance, it is very likely that many graders were a bit more detail-oriented than they might have otherwise been without an observer.

Data regarding grader training on the remainder of the course learning objectives has already been collected and can be analyzed to expand the spectrum of variability for some of the background variables (e.g., the LO and EI related variables) and to determine if expectations for increased variability correspond to greater disagreement of selected scores. This will lend credence to the generalizability of the model.

Additionally, revisions to the course grading structure can be made and analyzed in a similar fashion to determine how well a revised structure fits the previously developed model. Further, it can demonstrate whether or not the recommendations actually led to improvement. This will, in turn, revise the set of recommendations for improvement based on what was observed to be effective or ineffective and, potentially, why.

The need to improve reliability and validity of grading is an iterative process that should be continued indefinitely as perfect reliability is sought. That said, it will likely be a never-ending process in the context of engineering education. As this study demonstrates, there are several important trade-offs between authentic education and assessment and the ease of obtaining reliable scores. Real-life engineers face wicked, ill-defined, open-ended problems all the time. Preparing effective engineers means preparing engineering students to face these types of problems. However, increasing open-endedness seems to be inversely related to grading reliability. Similarly, while additional details in a rubric give graders more information to grade consistently, they also decrease the likelihood that graders will pay attention to the details or can even feasibly manage all the details within their working memory if they tried. Ultimately, designing course and grading materials, like any other engineering design process, may simplify to recognizing these conflicting variables, weighing potential outcomes, and making trade-offs.

APPENDIX A. RECRUITMENT EMAIL

Hi peer teachers and graders,

My name is Nathan Hicks and I am an engineering education Ph.D. student. As part of my research, I'm trying to improve the grading and training process to make the grades fairer for students and make the process of grading easier for you. In the next couple weeks, you'll be emailed a survey about the training process you've done this semester where you can give entirely honest feedback (it will be anonymous) about what aspects of the training have or have not worked and how you think it can or should be improved.

Before that, however, I am conducting interviews to get a better understanding of how you use the rubrics and grade student work. I am *the only* person who will know you participated and will remove any and all identifying information that might link you to what the interviews. There will be no personal judgment or evaluation of you or the job you do. These interviews are intended purely for the purpose of improving the process for you and future PTs/graders so you/they may do the job more easily and effectively.

The interview should take approximately **1 hour** and you will be given **\$20 cash immediately following the interview!** Please respond to my email if you would like to participate so we can arrange a time.

Thanks,

Nathan

APPENDIX B. OPEN CODING SAMPLE

Table B.1 shows a sample of the open coding process following the methods described by Saldaña (2016). The first four columns indicate which interview is being coded, the item number from the rubric, the evidence item the grader is evaluating, and the student sample the grader is evaluating, respectively. A blank in a row indicates that the value has not changed from the row above. The fifth column is the open-coding column, where a gerund-based description of what the grader was doing at each time is stated. The sixth column provides quotes that demonstrate the action being captured by the open code. Some actions did not have corresponding quotes, but were demonstrated by the annotations the participant made or were obvious intermediate steps; for example, in the first set of actions shown in Table B.1, the grader went from looking at the evidence items in the rubric to counting the number of test cases, meaning that the grader clearly shifted focus from the rubric to the sample. The final column represents general notes that were taken during the coding process to capture additional information, such as the obvious occurrence of an error performed by the grader.

Table B.1. Sample of initial open coding

Interview	Item #	Evidence Item(s)	Sample	Tasks Performed	Details	Note
1	1	1	1	Scanning EI	"Initially seven cases."	May have just looked at blue text
				Looking at student sample		
				Counting # of test cases	"1, 2, 3, ... 5, 6, 7. Does have seven so that's met."	Did not look at outcomes being tested
				Determining EI is met		
				Making mental note of EI being met		Did not immediately make a mark

Table B.1 continued

1	1	2	1	Scanning EI	"Use English to describe tests for invalid viscosity and laminar flow"	Focused on first clause of EI
				Looking at student sample		
				Locating one of specified cases		
				Determining that the specified case was tested	"They do test invalid viscosity"	Not actually what was asked for in the rubric
				Checking for accuracy of answer	"I'd check to see what the numbers they tested would be valid."	Not actually what was asked for in the rubric
				Reviewing problem set	"Let's see... Gotta check the problem set"	Not actually what was asked for in the rubric
				Identifying relevant information in problem set	"Viscosity between 0.001 and 25 for this test"	Not actually what was asked for in the rubric
				Locating test values		Not actually what was asked for in the rubric
				Comparing test values in specific case in sample to expected solution	"They did not test invalid viscosity" ($\mu = 1$ tested)	Not actually what was asked for in the rubric
				Determining EI not achieved (partially)		Based achievement of EI on details not asked for
				Looking at other test case specified by blue		
				Locating other specified test case		This suggests a compound EI (multiple distinct things to check)

Table B.1 continued

				Comparing test values in specific case to expected solution	"Uses proper fluid density, proper diameter, velocity, viscosity"	
				Reflecting on personal knowledge of subject	"This is specific to my knowledge from civil engineering, but that would be laminar flow there."	
				Making mental note of EI achievement		There was no statement for this, it is inferred, he quickly moved onto next EI
1	1	3	1	Reading EI	"Lists input arguments in a valid format, invalid value in viscosity"	
				Re-reading EI	"... okay, let's see, laminar flow test see note about testing inputs"	This is inferred from pause
				Checking accuracy of student work	"They don't have the specific ranges but I guess that's alright"	Did not seem to check both specified cases
				Making mental note of EI achievement		Again, no statement, just moved on.

APPENDIX C. INITIAL FOCUS CODING

Table C.1 shows a sample of the focus coding process following the methods of Saldaña (2016). This step followed the open coding process shown in Appendix B, where the task performed was converted into a broad task category (column 5), a specification of that category (column 6), and a contextual clarification for the task (column 7). The first four columns represent the interview number, the item within the assignment being evaluated, the evidence item being evaluated, and the sample being evaluated by the grader, respectively. Blank spaces indicate that the value did not change from the row above. The final column provides quotes or descriptions of actions taken as noted evidence for the task designation. Not every task required a contextual clarification in the “tasks performed” column or details to support in the “details” column.

Table C.1. Sample of focus coding

Interview	Item #	EI	Sample	Task	Task Specification	Tasks Performed	Details (quote(s) or action(s))
7	1	1	1	Orienting Translating	EI orienting Translating		"So we're looking at the number of test cases"
				Shifting Matching	Shifting Number matching	To sample	"Looks like there are 7."
				Shifting Orienting	Shifting Specified task orienting	To rubric	"There should be 7, so..."
				Scoring Annotating	EI scoring Annotating	Achieved	

Table C.1. continued

Interview	Item #	EI	Sample	Task	Task Specification	Tasks Performed	Details (quote(s) or action(s))
7	1	2	1	Orienting	EI orienting		"Then we're looking for the invalid viscosity test and the laminar flow test."
				Orienting	Specified task orienting		
				Shifting	Shifting	To sample	
				Scanning	Scanning to locate	Finding invalid viscosity	"This I would normally check with the code... let's see, [mutter], 0 to 10, 0.05 to 0.2, okay, so that looks okay."
				Evaluating	Evaluating quality		
				Scoring	EI scoring	Partially achieved (invalid visc.)	
				Shifting	Shifting	To rubric	"Normally I'd create some code to check these values to see if they actually produce this flow. For this I'd say it's correct."
				Annotating	Annotating		
				Shifting	Shifting	To sample	
				Scanning	Scanning to locate	Finding laminar	
				Evaluating	Evaluating quality	Checking values (should be checking English description)	
				Scoring	EI scoring	Partially achieved (laminar)	
				Shifting	Shifting	To rubric	
				Annotating	Annotating		

APPENDIX D. FINAL FOCUSED CODES

D.1 High-Level Focused Codes

Table D.1 shows the initial set of high-level focused codes that were developed using the initial open codes. The cognitive processes listed in the first column are the codes used in the “task” column in Table C.1 in Appendix C. The second column provides a detailed operational description of the cognitive process.

Table D.1. High-level focused codes

Cognitive Process (Foreground functions)	Description
Orienting	Tasks to orient the grader regarding the task, expected performances, or specified portions of a response
Matching	Checking to see if or how well a response or portion of response compares to the expected/correct response
Scanning	Looking through a response to find specific details or chunks of the response
Evaluating	Determining if an entire response or portion of a response meets a general or broad standard for performance or acceptably demonstrates proficiency
Scrutinizing	Analyzing to understand a response and infer respondents' understanding, knowledge, or intention
Scoring	An appraisal of a response
Reassuring	Convincing self of the appropriateness of a scoring decision
Second-guessing	Questioning a grading decision or returning to/revisiting a previous item or response after revised understanding of criteria or expectations
Rescoring	Changing a previous scoring decision in light of revised interpretation
Overruling	Consciously overriding specifications of a rubric based on autonomous judgment of appropriateness of score with respect to quality of student response or fairness of the specifications
Annotating	Making an actual physical annotation of EI achievement
Noting	Taking mental note of EI achievement
Shifting	Switching attention from one document (i.e., problem set, solution, rubric, sample response) to another
Questioning	Expressing confusion regarding one of the documents or part of one of the documents
Error spotting	Finding an unexpected part of a response
Translating	Stating EI/LO/etc. in simpler language based on understanding
Other	Tasks that do not fit into the other categories

D.2 Detailed Focused Codes

Table D.2 provides nuanced distinctions between some of the high-level tasks listed and described in Table D.1. These tasks were used in the focused coding samples demonstrated in the “task specification” column in Table C.1 shown in Appendix C. The Description column in Table D.2 provides a clearer description of the specified detailed-level task code.

Table D.2. Detailed focused codes

Cognitive Process	Task Specification	Description
Orienting tasks	EI orienting	Reading/re-reading/skimming the evidence item
	General solution orienting	Considering what constitutes an acceptable solution (without clear indication of how)
	LO orienting	Reading/re-reading/skimming the learning objective
	Mental model solution orienting	Thinking about grader's own mental model of what constitutes an acceptable solution
	Problem orienting	Reading/re-reading/skimming the problem, including the details provided to the student and the specific task asked of them
	Performance orienting	Considering student's previous performance to evaluate current performance
	Provided solution orienting	Reading/re-reading/skimming the exact provided solution
	Specified task orienting	Reading/re-reading/skimming the "what to grade" or blue specifying text
Matching tasks	Block matching	Comparing whole blocks or code or chunks of response
	Exact matching	Comparing an entire response to the expected solution or to another response
	Individual word/line matching	Comparing just a single word or line of code, or part of a flowchart
	Memory matching	Comparing memory of response with expected response
	Number matching	Comparing number(s) of items
	Scanning error correspondence	Looking through evidence items (after scanning response for errors) to determine if errors identified correspond to evidence for achievement
Scanning tasks	Scanning for errors	Looking through a large chunk or entire response for unexpected or incorrect aspects or features
	Scanning for gist	Looking over a whole response to get a sense of student's overall answer
	Scanning in memory	Using memory of response to determine if feature/aspect is present
	Scanning to locate	Looking through a response for a specific portion or task or piece of evidence to demonstrate achievement (failure to locate corresponds to a "no response" task)

Table D.2. continued

Cognitive Process	Task Specification	Description
Evaluating tasks	Evaluating comparability	Determining if an unexpected or unconventional response is equivalent to the expected response (e.g., lines of code are similar but in a different order for an EI that does not refer to order of commands)
	Evaluating quality	Determining if a qualitative property of the response is present or represents the entire response (e.g., sufficient commenting)
	Interpretive scrutinizing	Examining the response to infer the intentions or underlying understanding or knowledge of the respondent
Scrutinizing tasks	Qualitative scrutinizing	Reflecting upon whether or to what extent an unexpected or unconventional response meets the EIs, expectations, or learning objective
	Sensory scrutinizing	Examining the content closely to ensure proper decoding (i.e., when a response is handwritten and hard to read)
	Aggregating criteria	Adding up the number of correct or incorrect EIs or errors
Scoring tasks	EI scoring	Deciding whether an EI is, is not, or is partially achieved
	Holistically assessing	Assessing the overall, holistic quality of a response not based on individual EIs
	Memory scoring	Scoring by memory of response rather than direct concurrent inspection
	Overall scoring	Making decision on overall LO score for a sample

APPENDIX E. WORK-AS-IMAGINED CODING EXAMPLE

Table E.1 shows an example of how each rubric item and sample pairing was coded in Excel for the work-as-imagined instantiations. That is, this represents the set of functions that would need to be utilized and the outputs that should occur if a grader were to look for each aspect designated by the rubric and appropriately evaluate the performance demonstrated in the sample. The first column indicates the item on the assignment, the second indicates the specific evidence item that is being coded, or tasks that occur prior to any evaluation of evidence items or after evaluating all evidence items. The third column represents which of the three samples for the particular item is being coded. The final set of columns indicates the function that should be used, abbreviated using the coding numbers shown next to the cognitive functions for graders (see §4.3.2). Note that in multiple cases, the functions were utilized multiple times toward different aspects of the students response (for example, the grader should have scanned for the presence of (code 5) and found (indicated by the “+”) seven different test cases when evaluating the first evidence item.

Table E.1. Example of work-as-imagined instantiation coding in Excel

Item	EI	Sample	Function								
1	Start		1	2	3						
1	1	1	5(+) (dens.)	5(+) (vel.)	5(+) (diam.)	5(+) (visc.)	5(+) (turb.)	5(+) (lam.)	5(+) (trans.)	9(+)	10(+)
1	2	1	5(0) (visc.)	6(+) (visc.)	5(0) (lam.)	6(-) (lam.)	7(+) (lam.)	6b(+) (lam.)	9(+)	10(+)	
1	3	1	5(0) (visc.)	6(+) (visc.)	5(0) (lam.)	6(+) (lam.)	9(+)	10(+)			
1	4	1	5(0) (visc.)	7(-) (visc.)	6b(-) (visc.)	5(0) (lam.)	7(+) (lam.)	6b(+) (lam.)	9(-)	10(-)	
1	End	1	12	11(Dev.)	13						

APPENDIX F. WORK-AS-COMPLETED CODING EXAMPLE

Table F.1 shows an example of the coding of the observed grading performed by graders in the think-aloud interviews using the functions created for the FRAM model. The first column indicates the interview being coded, the second indicates the evidence item (or actions prior to looking at evidence items or after evaluating evidence items), and the third indicates the sample being graded. The final set of columns show the cognitive level grader functions that were observed, following the short-hand described in chapter 4 (see §4.3.2). To streamline the process, work-as-imagined codes were used in Excel for all interviews (cells in white). When an expected function was observed with the appropriate output, the cell remained white. When the function was not observed, the cell was highlighted red. When the function was observed but had an unexpected output, the cell was modified, sometimes including clarifying text, and highlighted yellow. When additional, unexpected functions were performed, a new cell was added at the end of the row (note that chronology of functions performed is not indicated here) and highlighted green. Occasionally, functions were not observed directly due to insufficient verbalizations of the participants but were inferred as occurring based on other contextual indicators. These instances were highlighted in blue (not shown in the sample in Table F.1).

Table F.1. Example of work-as-completed coding in Excel

Interview	EI	Sample	Function									
1	Start		1	2	3							
1	1	1	5(+) (dens.)	5(+) (vel.)	5(+) (diam.)	5(+) (visc.)	5(+) (turb.)	5(+) (lam.)	5(+) (trans.)	9(+)	10(+) (M)	5(+) (7 cases)
1	2	1	5(0) (visc.)	6(-) (visc.) *matched test values	5(0) (lam.)	6(+) (lam.) *matched test values	7(+) (lam.)	6b(+) (lam.)	9(-)	10(-) (P)		
1	3	1	5(0) (all)	6(+) (all)	5(0) (lam.)	6(+) (lam.)	9(+)	10(+)	3a *Refers to note in what to grade section of rubric			
1	4	1	5(0) (visc.)	7(-) (visc.)	6b(-) (visc.)	5(0) (lam.)	7(+) (lam.)	6b(+) (lam.)	9(-)	10(-)	6(-) *mentally	16 *States this was already accounted for in EI #2. Grades holistically to developing
1	End	1	12	11(Prof.)	13							

APPENDIX G. RUBRICS

LO 1: Problem 1 - Reynolds Number

Learning Objective	15.09 Create test cases to evaluate a flowchart												
What to Grade:	PS09_answer_sheet.docx > Problem 1												
	Prob 1, Step 2a-b:												
	Grade the test-case table for Problem 1. In particular, look at												
	<ul style="list-style-type: none">The number of total test casesThe details in the invalid viscosity test row (first two columns)												
	<table><tr><th>Test Case</th><th>Input Arguments</th><th>Flowchart Output</th></tr><tr><td>Test the validity of the viscosity input by using an invalid viscosity value</td><td>() () ()</td><td rowspan="2">GRADE BELOW in LO 15.02</td></tr><tr><td>All other inputs are valid</td><td>()</td></tr></table>		Test Case	Input Arguments	Flowchart Output	Test the validity of the viscosity input by using an invalid viscosity value	() () ()	GRADE BELOW in LO 15.02	All other inputs are valid	()			
	Test Case	Input Arguments	Flowchart Output										
Test the validity of the viscosity input by using an invalid viscosity value	() () ()	GRADE BELOW in LO 15.02											
All other inputs are valid	()												
<ul style="list-style-type: none">The details in the valid test case for laminar flow (first two columns)													
<table><tr><th>Test Case</th><th>Input Arguments</th><th>Flowchart Output</th></tr><tr><td>Test a laminar flow (Re< 2300)</td><td>() ()</td><td rowspan="2">GRADE BELOW in LO 15.02</td></tr><tr><td>Uses valid inputs that produce a laminar Re</td><td>()</td></tr></table>		Test Case	Input Arguments	Flowchart Output	Test a laminar flow (Re< 2300)	() ()	GRADE BELOW in LO 15.02	Uses valid inputs that produce a laminar Re	()				
Test Case	Input Arguments	Flowchart Output											
Test a laminar flow (Re< 2300)	() ()	GRADE BELOW in LO 15.02											
Uses valid inputs that produce a laminar Re	()												
NOTE: you will need to check that the input arguments for the laminar flow. You may want to use Excel to set up a formula to check the value of Re from the inputs provided by the student.													
NOTE: you're grading the selection of the test cases. The output will be graded in the next LO.													
Proficient		Developing	Emerging	Insufficient Evidence	No Attempt								
1 pt		0.8 pt	0.5 pt	0 pt	0 pt								
Evidence items for proficiency: 1. Creates thorough set of test cases to test all possible outcomes in the flowchart There are seven cases 2. Use English to describe each test and how the information moves through the flowchart for that test Column 1: invalid viscosity test Column 1: laminar flow test 3. Lists input arguments in a valid format Column 2: an invalid value in the viscosity input Column 2: laminar flow test, see note above about testing inputs 4. Test values are consistent with the test description		1 (of 4) missing or incorrect item from the proficient list	2 (of 4) missing or incorrect items from the proficient list	3 or more (of 4) missing or incorrect items from the proficient list	Did not attempt the graded item								

LO 2: Problem 1 - Reynolds Number

Learning Objective	15.02 Track a flowchart with a selection structure									
What to Grade:	PS09_answer_sheet.docx > Problem 1									
	Prob 1, Step 2a-b:									
	Grade the test-case table for Problem 1. In particular, look at									
	<ul style="list-style-type: none">The details in the invalid viscosity test row (last column ONLY)									
	<table><tr><th>Test Case</th><th>Input Arguments</th><th>Flowchart Output</th></tr><tr><td>graded above</td><td>graded above</td><td>Error: invalid viscosity</td></tr></table>		Test Case	Input Arguments	Flowchart Output	graded above	graded above	Error: invalid viscosity		
	Test Case	Input Arguments	Flowchart Output							
graded above	graded above	Error: invalid viscosity								
<ul style="list-style-type: none">The details in the valid test case for laminar flow (all three columns)										
<table><tr><th>Test Case</th><th>Input Arguments</th><th>Flowchart Output</th></tr><tr><td>graded above</td><td>graded above</td><td>Laminar flow</td></tr></table>		Test Case	Input Arguments	Flowchart Output	graded above	graded above	Laminar flow			
Test Case	Input Arguments	Flowchart Output								
graded above	graded above	Laminar flow								
NOTE: The output should not have any code outputs. It should be a description of the flowchart output, not actual MATLAB results.										
Evidence items for proficiency:		Not Used for Assessing Student Work	1 (of 2) missing or incorrect item from the proficient list	2 (of 2) missing or incorrect items from the proficient list OR MATLAB code results	Did not attempt the graded item					
1. Identify correct path given the test value(s) Check for both the invalid viscosity and the laminar flow cases										
2. Describe the outcomes(s) in English with resulting values when appropriate (not code results) Check for both the invalid viscosity and the laminar flow cases										

LO 3: Problem 1 - Reynolds Number

Learning Objective	16.01 Convert between these selection structure representations: English, a flowchart, and code				
What to Grade:	PS09_reynolds_report.pdf				
	Prob 1, Step 3:				
	Grade how the code compares to selection structure for the Reynolds number flow type to the flowchart in the problem set.				
	<pre>% Test Re values and print flow type if Re < 2300 fprintf('Flow type: Laminar\n\n') elseif Re > 4800 fprintf('Flow type: Turbulent\n\n') else fprintf('Flow type: Transitional\n\n') end % end test of Re values</pre>	<pre>graph TD Start(()) --> D1{Is flow laminar?} D1 -- yes --> P1[/Print flow type: Laminar/] D1 -- no --> D2{Is flow turbulent?} D2 -- yes --> P2[/Print flow type: Turbulent/] D2 -- no --> P3[/Print flow type: Transitional/] P1 --> Stop([Stop]) P2 --> Stop P3 --> Stop</pre>			
Proficient	Developing	Emerging	Insufficient Evidence	No Attempt	
1 pt	0.8 pt	0.5 pt	0 pt	0 pt	
Evidence items for proficiency: 1. Recognize that a diamond structure with one input arrow and two output arrows (labeled Yes/No or True/False) translates to an if or elseif statement 2. The number of diamonds in the flowchart translates exactly to the number if and elseif statements 3. Recognize that the first 1-in/2-out diamond in a flowchart (or first following other non-decision instructions or the first on a Yes path following a decision) is an if statement 4. Recognize that all immediately following 1-in/2-out diamonds on the No or False path are elseif statements 5. Recognize an else statement is implied if there are operations between the only or last diamond and the convergence of the flowchart connecting lines. 6. Recognize that a convergence of the entire No or False path with the entire Yes or True path translates to an end statement	1 (of 6) missing or incorrect item from the proficient list	2 (of 6) missing or incorrect items from the proficient list	3 or more (of 6) missing or incorrect items from the proficient list	Did not attempt the graded item	

LO 4: Problem 2 - Atmosphere (Flowchart)

Learning Objective	15.10 Construct a flowchart using standard symbols and pseudocode				
What to Grade:	PS09_answer_sheet.docx > Problem 2 > Flowchart				
	Prob 2, Step 1:				
	Grade the Problem 2 flowchart for items listed in the proficient list below.				
	In addition to requiring the correct use of symbols and formatting, a student’s flowchart must generally attempt to meet the requirements of the problem. A student’s flowchart is considered totally incomplete if it is missing <ul style="list-style-type: none">2 or more paths OR3 or more <i>unique</i> (as in, not repeated across paths) computation or output instruction steps. Such a flowchart will be considered to be at an Insufficient Evidence level. If a student’s flowchart is missing only 1 path or 1-2 unique steps, the flowchart will be considered partially incomplete and just count as Evidence Item 11 as incorrect.				
Proficient	Developing	Emerging	Insufficient Evidence	No Attempt	
1 pt	0.8 pt	0.5 pt	0 pt	0 pt	
Evidence items for proficiency: <div>1. Flowchart symbols: Start and stop for the overall flowchart are represented by ovals</div> <div>2. Flowchart symbols: Inputs and outputs are represented by parallelograms</div> <div>3. Flowchart symbols: Decisions are represented by diamonds</div> <div>4. Flowchart symbols: Processes, such as calculations, are represented by rectangles</div> <div>5. Flowchart symbols: Operations are connected with arrows with points at one end to indicate flow</div> <div>6. Arrows must connect all flowchart elements and indicate a continuous flow from start to stop.</div> <div>7. Arrows must converge prior to stop so that there is only one arrow into the stop</div> <div>8. Flowchart process ends in one place (cannot end in multiple places)</div> <div>9. Text within the symbols is in concise English (not code or only math) that conveys the purpose of the step</div> <div>10. Decisions are accompanied by Yes/No or True/False text on the appropriate arrows</div> <div>11. Flowchart represents all possible outcomes required by the problem</div>	1-2 (of 11) missing or incorrect item from the proficient list	3-4 (of 11) missing or incorrect items from the proficient list	5 or more (of 11) missing or incorrect items from the proficient list OR Flowchart is incomplete (missing 2+ paths or 3+ instructions)	Did not attempt the graded item	

LO 5: Problem 2 - Atmosphere (Flowchart)

Learning Objective	15.01 Construct a flowchart for a selection structure using standard symbols and pseudocode			
What to Grade:	PS09_answer_sheet.docx > Problem 2 > Flowchart Grade the specifics of the selection structure within their flowchart <i>Prob 2, Step 1:</i>			
Proficient	Developing	Emerging	Insufficient Evidence	No Attempt
1 pt	0.8 pt	0.5 pt	0 pt	0 pt
Evidence items for proficiency: 1. Decisions that are part of a selection structure are represented with a diamond filled with a condition 2. Decision have one input arrow and two output arrows (one for Yes/True and one for No/False) 3. There are operations on the Yes/True path 4. For multiple related selections (i.e., if-elseif-else), there are no operations between the decisions along the No/False path 5. For multiple related selections (i.e., if-elseif-else), the Yes/True and No/False path arrows converge after all related decisions and (optionally) the operations for the else path 6. Operations are included in the selection structure as required by the problem	1 (of 6) missing or incorrect item from the proficient list	2 (of 6) missing or incorrect items from the proficient list	3 or more (of 6) missing or incorrect items from the proficient list	Did not attempt the graded item

LO 6: Problem 2 - Atmospheric Temperature

Learning Objective		15.09 Create test cases to evaluate a flowchart (2)				
What to Grade:	PS09_answer_sheet.docx > Problem 2 > Test Cases					
	Prob 2, Step 2:					
	Grade the test-case table for Problem 2. In particular, look at					
	<ul style="list-style-type: none">The number of total test cases (one case for each layer (5) and at least one invalid)The details in the invalid case					
	Test Case		Input Arguments		Flowchart Output	
	Test an altitude where $h < 0$ or $h \geq 51$		any value $h < 0$ or $h \geq 51$		GRADE BELOW in LO 15.02	
What to Grade:	<ul style="list-style-type: none">The details in the valid test case for higher stratosphere layer					
	Test Case		Input Arguments		Flowchart Output	
	Test an altitude where $32 \leq h < 47$, which is inside the stratosphere		any value $32 \leq h < 47$		GRADE BELOW in LO 15.02	
Proficient		Developing	Emerging	Insufficient Evidence	No Attempt	
1 pt		0.8 pt	0.5 pt	0 pt	0 pt	
Evidence items for proficiency: 1. Creates thorough set of test cases to test all possible outcomes in the flowchart There are six cases 2. Use English to describe each test and how the information moves through the flowchart for that test Column 1: invalid altitude Column 1: valid altitude in the higher layer of the stratosphere 3. Lists input arguments in a valid format Column 2: invalid altitude test, any value $h < 0$ or $h \geq 51$ Column 2: stratosphere test, any value $32 \leq h < 47$ 4. Test values are consistent with the test description		1 (of 4) missing or incorrect item from the proficient list	2 (of 4) missing or incorrect items from the proficient list	3 or more (of 4) missing or incorrect items from the proficient list	Did not attempt the graded item	

LO 7: Problem 2 - Atmospheric Temperature

Learning Objective	15.02 Track a flowchart with a selection structure (2)				
What to Grade:	PS09_answer_sheet.docx > Problem 2 > Test Cases				
	Grade the test-case table for Problem 2. In particular, look at				
	<ul style="list-style-type: none">The details in the invalid case, column 3				
	Test Case			Flowchart Output	
	graded above			Error: invalid altitude	
	<ul style="list-style-type: none">The details in the valid test case for higher stratosphere layer, column 3				
Test Case			Flowchart Output		
graded above			Stratosphere		
Proficient		Developing	Emerging	Insufficient Evidence	No Attempt
1 pt		xx pt	0.5 pt	0 pt	0 pt
Evidence items for proficiency: 1. Identify correct path given the test value(s) Check for both cases 2. Describe the outcomes(s) in English with resulting values when appropriate (not code results) Check for both cases		Not Used for Assessing Student Work	1 (of 2) missing or incorrect item from the proficient list	2 (of 2) missing or incorrect items from the proficient list OR MATLAB code results	Did not attempt the graded item

LO 8: Problem 2 - Atmospheric Temperature

Learning Objective	16.02 Code a selection structure				
What to Grade:	PS09_atm_temp.pdf > CALCULATIONS				Prob 2, Step 3:
	Grade the student's selection structure code for calculating atmospheric temperature.				
	Do not compare their code to their flowchart (not assessed in this LO)				
Proficient		Developing	Emerging	Insufficient Evidence	No Attempt
1 pt		0.8 pt	0.5 pt	0 pt	0 pt
Evidence items for proficiency: 1. Begin a selection structure with an <code>if</code> 2. The <code>if</code> is accompanied by a condition for which a true result corresponds to code that immediately follows 3. <code>elseif</code> is used for a series of related conditions 4. Each <code>elseif</code> is accompanied by a condition which a true result corresponds to code that immediately follows 5. <code>elseif</code> is a single word – there is no space between <code>else</code> and <code>if</code> 6. An <code>else</code> is used to handle any condition(s) not addressed in the earlier parts of the selection structure and not used if no code is needed before the <code>end</code> 7. An <code>else</code> is not accompanied by a condition 8. <code>end</code> is used to terminate the selection structure 9. Statements between the <code>if</code> , <code>elseif</code> , <code>else</code> , and <code>end</code> are indented 10. A selection structure addresses all necessary paths for a given problem		1-2 (of 10) missing or incorrect items from the proficient list	3-4 (of 10) missing or incorrect items from the proficient list	5 or more (of 10) missing or incorrect items from the proficient list	Did not attempt the graded item

LO 9: Problem 3 - Contact Lens Decision

Learning Objective	11.11 Coordinate the passing of information between functions					
What to Grade:	<div>PS09_contactlens.pdf > FUNCTION CALLS</div> <div>Grade the function calls within the exec function (only grade for Lens ID LM17): lens_data = csvread('Data_newlensdesigns.csv',2,0); threshold = 0.02; % contact lens threshold % Create string variables for the design batch ID names % to use in the pcode and in the plot legend lens1 = 'LM17'; % lens design batch ID % Call stats UDF to get mean and std dev on each lens parameter [rbc1_mean,rbc1_std] = Solution_stats_io(lens_data(:,1)); [dial_mean,dial_std] = Solution_stats_io(lens_data(:,2)); % Call new contactlens decision code to determine acceptability decLM17 = Solution_contactlens_decision(lens1,rbc1_mean,rbc1_std,dial_mean,dial_std,threshold);</div>				Prob 3, Step 2:	
Proficient	Developing	Emerging	Insufficient Evidence	No Attempt		
1 pt	0.8 pt	0.5 pt	0 pt	0 pt		
<div>Evidence items for proficiency:</div> <div><div>1. Call to a user-defined function occurs in the proper function or script</div><div>Both the stats function and the PS09_contactlens_decision function are called in the exec function</div><div>2. Variables passed into a user-defined function are defined prior to calling the user-defined function</div><div>data, threshold, and lens batch ID variables are defined prior to use in a UDF</div><div>3. Variables passed into a user-defined function are defined prior to calling the user-defined function</div><div>STATS function is called before DECISION so that the statistics values are defined prior to running DECISION</div><div>4. User-defined functions are called in the order necessary to complete the coding task</div><div>STATS function is called before DECISION so that the statistics values are defined prior to running DECISION</div><div>5. No use of global variables (to circumvent proper passing of information through function calls)</div></div>	<div>1 (of 5) missing or incorrect item from the proficient list</div>	<div>2 (of 5) missing or incorrect items from the proficient list</div>	<div>3 or more (of 5) missing or incorrect items from the proficient list</div>	<div>Did not attempt the graded item</div>		

LO 10: Problem 3 - Contact Lens Decision

Learning Objective	11.03 Create a user-defined function that adheres to programming standards				
What to Grade:	PS09_contactlens_decision.m Grade the programming standards of the contact lens decision code (There is no published PDF of this code, so you will need to grade the m-file)				Problem 3
Proficient	Developing	Emerging	Insufficient Evidence	No Attempt	
1 pt	0.8 pt	0.5 pt	0 pt	0 pt	
Evidence items for proficiency: 1. Help lines contain input and output argument definitions, with units as appropriate 2. Help lines contain concise description of the program 3. Help lines show the call to the function 4. Complete programmer and contributor information in the header (names and emails) 5. Complete problem details including assignment number, problem number 6. Code items are in the correct section (e.g. Initialization, Calculations, ...) 7. Computed values are assigned to variables 8. Code blocks have explanatory comments 9. Variables have commented definitions and units 10. Minimal use of hardcoding	2 (of 10) missing or incorrect item from the proficient list	3-4 (of 10) missing or incorrect item from the proficient list	5 or more (of 10) missing or incorrect item from the proficient list	Did not attempt the graded item	

APPENDIX H. TASK PERFORMANCES

The figures in Appendix H demonstrate the task performances evaluated within the think-aloud interviews. Important takeaways for these task performances are discussed in section 5.3.1.

Test Case Description in English	Test Values (Flowchart Output
Test the validity of the density input by using an invalid density value All other inputs are valid	= 0.1 = 1 = 0.1 = 1	Error: invalid density
Test the validity of the velocity input by using an invalid value All other inputs are valid	= 1 = 15 = 0.1 = 1	Error: invalid velocity
Test the validity of the diameter input by using an invalid value All other inputs are valid	= 1 = 1 = 0.5 = 1	Error: invalid diameter
Test the validity of the viscosity input by using an invalid value All other inputs are valid	= 1 = 0 = 0.1 = 1	Error: invalid viscosity
Test when all inputs are valid and flow is turbulent	= 1500 = 10 = 0.05 = .001	Print “ = 1500, = 10, = 0.05, = .001, Re = 750000” Print “flow is turbulent”
Test when all inputs are valid and flow is laminar	= 1 = 1 = 0.1 = 1	Print “ = 1, = 1, = 0.1, = 1, Re = 0.1” Print “flow is laminar”
Test when all inputs are valid and flow is transitional	= 1200 = 2 = 0.1 = .1	Print “ = 1200, = 2, = 0.1, = 0.1, Re = 2400” Print “flow is transitional”

Figure H.1. Student sample 1 for LO 1 and LO 2.

Test Case Description in English	Test Values (Flowchart Output
Test the validity of the density input by using an invalid density value All other inputs are valid	= 0.1 = 1 = 0.1 = 1	Error: invalid density
Test the validity of the using an invalid velocity value < 0 or > 10 All other inputs are valid	= 1400 = 11 = 0.1 = 24	Error: invalid velocity
Test the validity of the using an invalid diameter value < 0.05 or > 0.2 All other inputs are valid	= 1400 = 9 = 0.04 = 24	Error: invalid diameter
Test the validity of the using an invalid diameter value < 0.001 or > 25 All other inputs are valid	= 1400 = 9 = 0.04 = 26	Error: invalid viscosity

Figure H.2. Student sample 2 for LO 1 and LO 2.

Test Case Description in English	Test Values (Flowchart Output
Test the validity of the density input by using an invalid density value All other inputs are valid	= 100 = 1 = 0.1 = 1	Error: invalid density
Test with all valid inputs	= 0.5 = 0 = 0.05 = 0.001	The Flow is Laminar
Test with invalid density input and all other valid input	= 1600 = 12 = 0.1 = 15	Error: invalid density
Test with invalid velocity input and all other valid input	= 1500 = 12 = 0.1 = 15	Error: invalid velocity
Test with invalid diameter input and all other valid input	= 1500 = 5 = 0.01 = 15	Error: invalid diameter
Test with invalid viscosity input and all other valid input	= 1500 = 5 = 0.1 = 30	Error: invalid viscosity

Figure H.3. Student sample 3 for LO 1 and LO 2.

```

else
    Re = (density*velocity*diameter)/viscosity;
    fprintf('The fluid density is %.2f (kg/m^3).\n',density)
    fprintf('The fluid velocity is %.2f (m/s).\n',velocity)
    fprintf('The pipe diameter is %.2f (m).\n',diameter)
    fprintf('The fluid viscosity is %.2f (Pa*s).\n',viscosity)
    fprintf('The Reynolds number is %.4f.\n',Re)
    if Re < 2300
        fprintf('Flow Type: Laminar')
    elseif Re > 4800
        fprintf('Flow Type: Turbulent')
    elseif (Re >= 2300) & (Re <= 4800)
        fprintf('Flow Type: Transitional')
    end
end

```

Figure H.4. Portion of student sample 1's code for LO 3.

```

if Re<2300 %Determines the type of flow
    fprintf('Laminar Flow\n')
elseif Re>4800
    fprintf('Turbulant Flow\n')
else
    fprintf('Transitional Flow\n')
end

```

Figure H.5. Portion of student sample 3's code for LO 3.

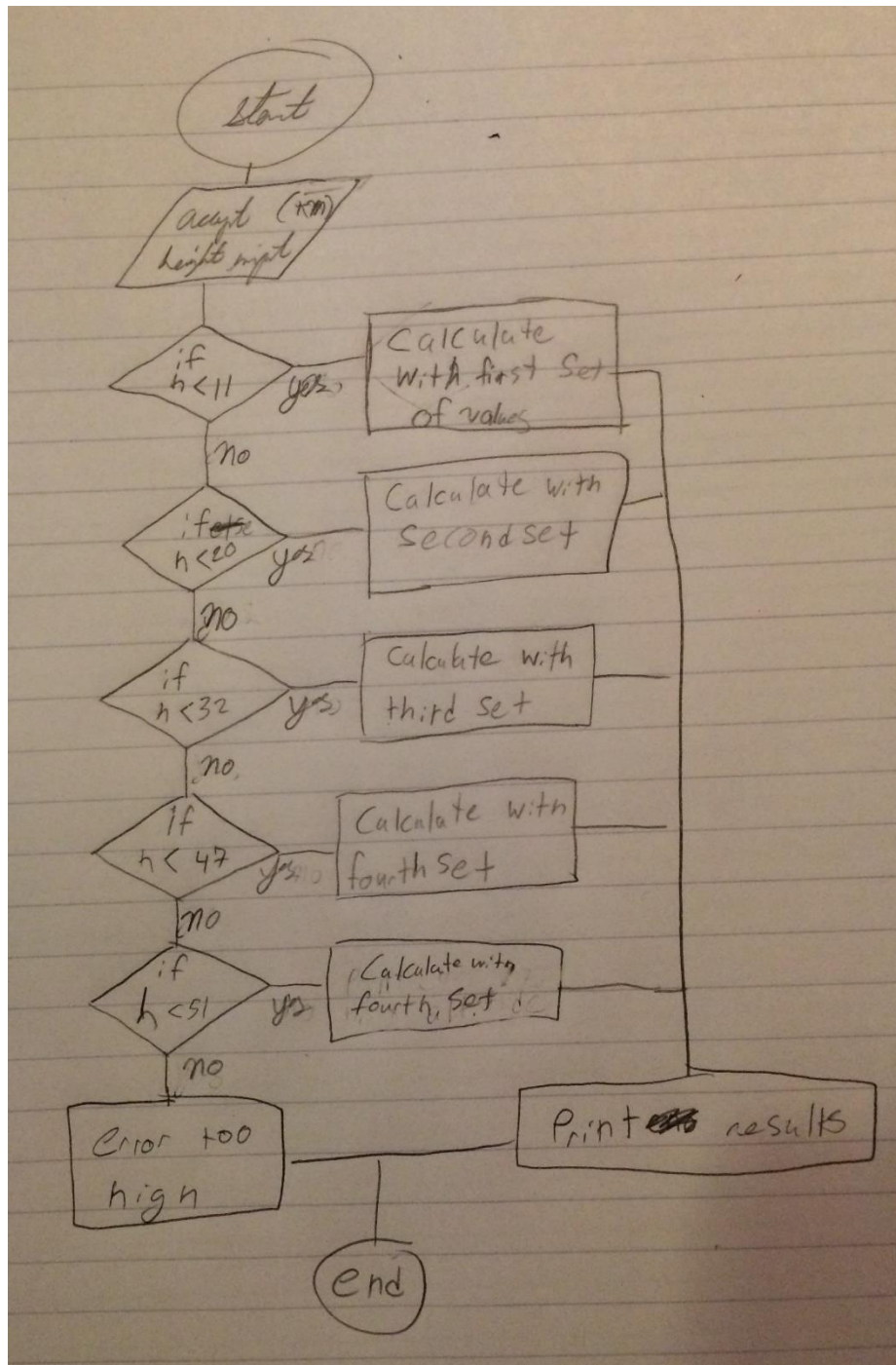


Figure H.6. Student sample 1 for LO 4 and LO 5.

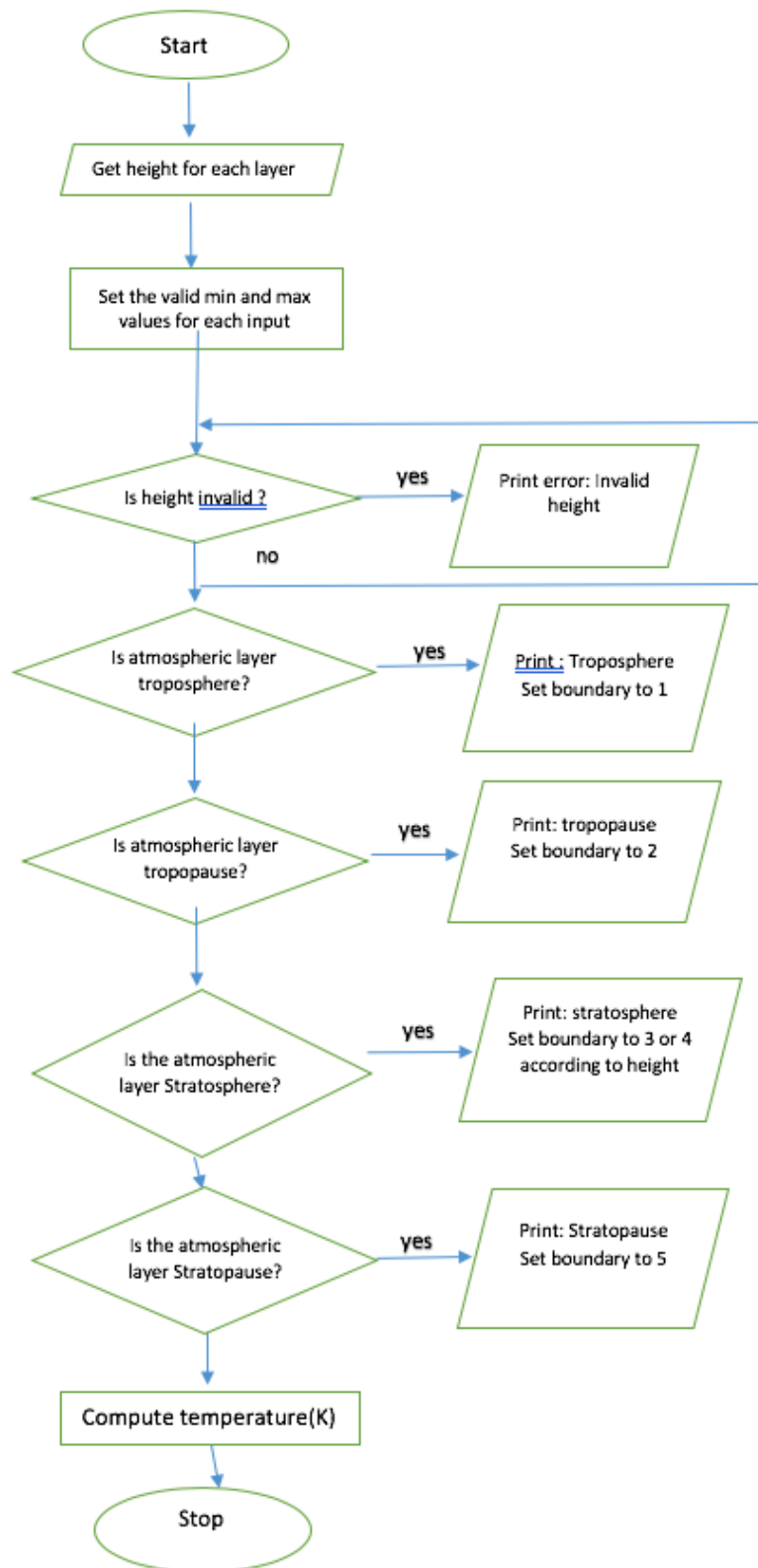


Figure H.7. Student sample 2 for LO 4 and LO 5.

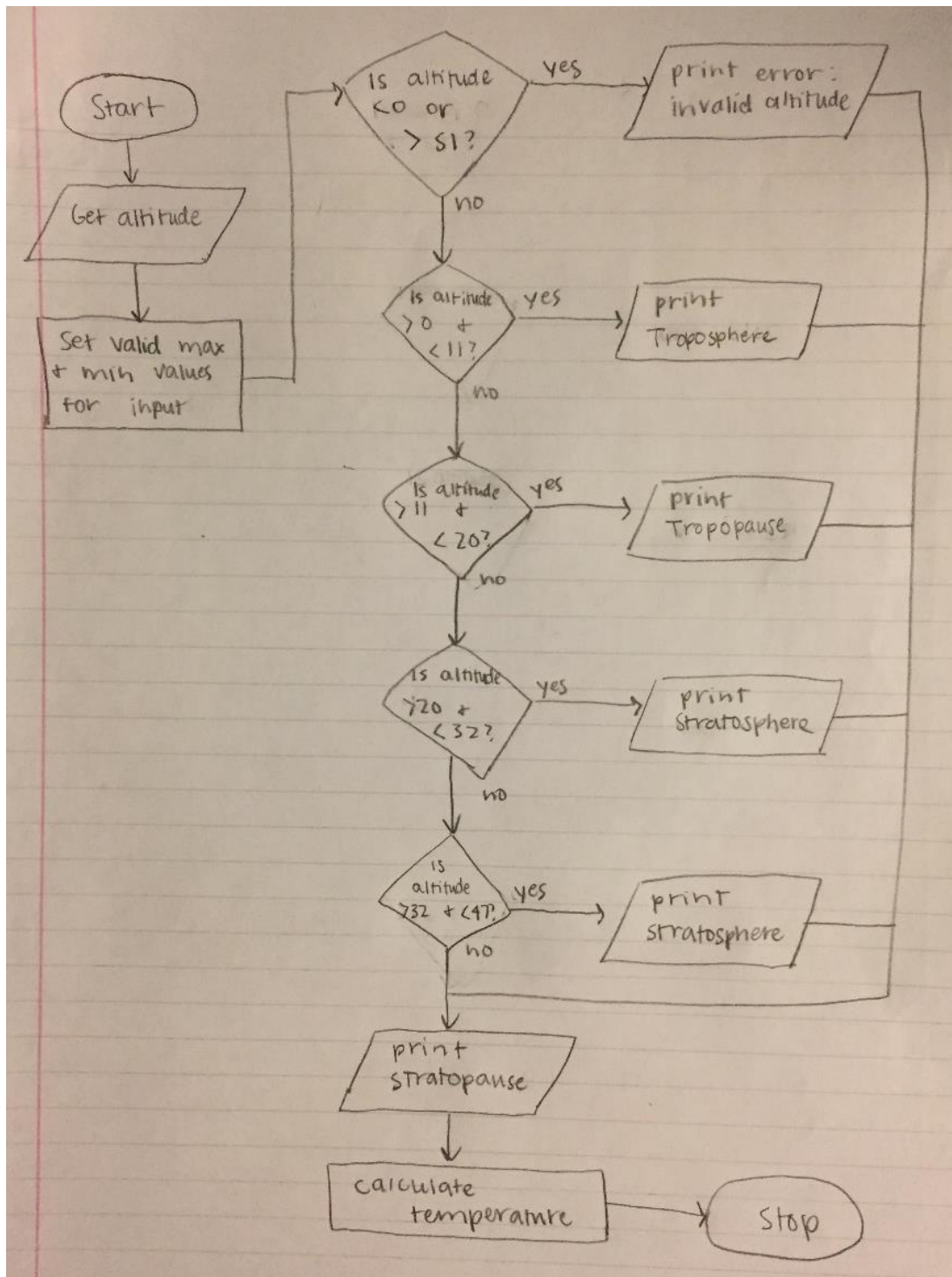


Figure H.8. Student sample 3 for LO 4 and LO 5.

Test Case Description in English	Input Argument (altitude (km))	Flowchart Output in English
Test when altitude is valid and in the troposphere $0 \leq h < 11$	$h = 10$	Atmospheric layer: troposphere
Test when altitude is valid and in the troposphere $11 \leq h < 20$	$h = 15$	Atmospheric layer: tropopause
Test when altitude is valid and in the troposphere $20 \leq h < 32$	$h = 25$	Atmospheric layer: Stratosphere with boundary 3
Test when altitude is valid and in the troposphere $32 \leq h < 47$	$h = 35$	Atmospheric layer: Stratosphere with boundary 4
Test when altitude is valid and in the troposphere $47 \leq h < 51$	$h = 48$	Atmospheric layer: Stratopause

Figure H.9. Student sample 1 for LO 6 and LO 7.

Test Case Description in English	Input Argument (altitude (km))	Flowchart Output in English
Test when altitude is invalid $h < 0$	$h = -14$	Error: Invalid Altitude
Test when altitude is valid and in the troposphere $0 \leq h < 11$	$h = 10$	Atmospheric layer: troposphere
Test when altitude is valid and in the tropopause $11 \leq h < 20$	$h = 14$	Atmospheric layer: tropopause
Test when altitude is valid and in the stratosphere boundary number 3 $20 \leq h < 32$	$h = 24$	Atmospheric layer: stratosphere
Test when altitude is valid and in the stratosphere boundary number 4 $32 \leq h < 47$	$h = 39$	Atmospheric layer: stratosphere
Test when altitude is valid and in the stratopause $47 \leq h < 51$	$h = 50$	Atmospheric layer: stratopause
Test when altitude is invalid $h \geq 51$	$h = 52$	Error: Invalid Altitude

Figure H.10. Student sample 2 for LO 6 and LO 7.

Test Case Description in English	Input Argument (altitude (km))	Flowchart Output in English
Test when altitude is valid and in the troposphere $0 \leq h < 11$	$h = 10$	Atmospheric layer: troposphere
Test when altitude is valid and in the troposphere $0 \leq h < 11$	$H = 9$	Atmospheric layer: troposphere
Test when altitude is valid and in the troposphere $11 \leq h < 20$	$H = 15$	Atmospheric layer: tropopause
Test when altitude is valid and in the troposphere $20 \leq h < 32$	$H = 30$	Atmospheric layer: stratosphere
Test when altitude is valid and in the troposphere $32 \leq h < 47$	$H = 36$	Atmospheric layer: stratosphere
Test when altitude is valid and in the troposphere $51 \leq h$	$H = 60$	Atmospheric layer: ERROR!

Figure H.11. Student sample 3 for LO 6 and LO 7.

```

%Determines whether the height is a valid input
if height<troposphere_baseH | height>max_height
    error('Invalid Height')
end
%Calculates the temperature of the height in an atmospheric layer
if height>=troposphere_baseH & height<tropopause_baseH
    temp = troposphere_baseL*(height-troposphere_
baseH)+troposphere_baseT;
    atm_layer = 'Troposphere Layer';
end
if height>=tropopause_baseH & height<stratosphere_baseH_3
    temp = tropopause_baseL*(height-tropopause_
baseH)+tropopause_baseT;
    atm_layer = 'Tropopause Layer';
end
if height>=stratosphere_baseH_3 & height<stratosphere_baseH_4
    temp = stratosphere_baseL_3*(height-stratosphere_
baseH_3)+stratosphere_baseT_3;
    atm_layer = 'Stratosphere Layer';
end
if height>=stratosphere_baseH_4 & height<stratopause_baseH
    temp = stratosphere_baseL_4*(height-stratosphere_
baseH_4)+stratosphere_baseT_4;
    atm_layer = 'Stratosphere Layer';
end
if height>=stratopause_baseH & height<max_height
    temp = stratopause_baseL*(height-stratopause_
baseH)+stratopause_baseT;
    atm_layer = 'Stratopause Layer';
end
end

```

Figure H.12. Portion of student sample 1's code for LO 8.

```

if altitude < 51 %first if else construct
    if altitude < 11 %second if else construct
        layer = 'Troposphere'; %names the layer as a string
        Tb = boundOne(1); %Base Temperature
        Hb = boundOne(2); %Base Height
        Lb = boundOne(3); %Temperature Lapse
    elseif altitude >= 11 && altitude < 20 %elseif
        layer = 'Tropopause'; %names the layer as a string
        Tb = boundTwo(1); %Base Temperature
        Hb = boundTwo(2); %Base Height
        Lb = boundTwo(3); %Temperature Lapse
    elseif altitude >= 20 && altitude < 32 %elseif
        layer = 'Stratosphere'; %names the layer as a string
        Tb = boundThree(1); %Base Temperature
        Hb = boundThree(2); %Base Height
        Lb = boundThree(3); %Temperature Lapse
    elseif altitude >= 32 && altitude < 47 %elseif
        layer = 'Stratosphere'; %names the layer as a string
        Tb = boundFour(1); %Base Temperature
        Hb = boundFour(2); %Base Height
        Lb = boundFour(3); %Temperature Lapse
    else %else, there are no more conditions
        layer = 'Stratopause'; %names the layer as a string
        Tb = boundFive(1); %Base Temperature
        Hb = boundFive(2); %Base Height
        Lb = boundFive(3); %Temperature Lapse
    end %ends the second if construct
    Temp = Tb + Lb * (altitude - Hb); %calculates temp
    fprintf('At the height %ikm, we are at the %s and the temperature
is %.2fK.\n', altitude, layer, Temp) %the output
else %else, there are no more conditions
    fprintf('ERROR! Your input is out of range.\n') %this is the
    output if the input is out of range
end %ends the first if else construct

```

Figure H.13. Portion of student sample 2's code for LO 8.

```

if h<11 %seperates each height into the seperate groups
    level = 'Troposphere';
    temp = 288.15 - 6.5*(h);
elseif h<20
    level = 'Tropopause';
    temp = 216.65;
elseif h<32
    level = 'Lower Stratosphere';
    temp = 216.65 + (h-20);
elseif h<47
    level = 'Upper Stratosphere';
    temp = 216 + 2.8*(h-32);
elseif h<51
    level = 'Stratopause';
    temp = 270.65;
else
    level = 'Above the atmosphere';
    temp = 'N/A';

```

Figure H.14. Portion of student sample 3's code for LO 8.

```
function [] = PS09_contactlens_█()
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% *REMOVED FOR SPACE*
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

INITIALIZATION

```
lensData = csvread('Data_newlensdesigns.csv',2);

LM17_bcr = lensData(:,1);
LM17_d = lensData(:,2);
LR283_bcr = lensData(:,3);
LR283_d = lensData(:,4);
LP06_bcr = lensData(:,5);
LP06_d = lensData(:,6);
LH44_bcr = lensData(:,7);
LH44_d = lensData(:,8);
```

LENS DESIGN PLOT

REMOVED FOR SPACE

FUNCTION CALLS

```
[Mean1,STD1] = PS04_stats_io_█(LM17_bcr);
[Mean2,STD2] = PS04_stats_io_█(LM17_d);
[Mean3,STD3] = PS04_stats_io_█(LR283_bcr);
[Mean4,STD4] = PS04_stats_io_█(LR283_d);
[Mean5,STD5] = PS04_stats_io_█(LP06_bcr);
[Mean6,STD6] = PS04_stats_io_█(LP06_d);
[Mean7,STD7] = PS04_stats_io_█(LH44_bcr);
[Mean8,STD8] = PS04_stats_io_█(LH44_d);

[dec] =
PS09_contactlens_decision_█('LM17',Mean1,STD1,Mean2,STD2,.02);
[dec] =
PS09_contactlens_decision_█('LR283',Mean3,STD3,Mean4,STD4,.02);
[dec] =
PS09_contactlens_decision_█('LP06',Mean5,STD5,Mean6,STD6,.02);
[dec] =
PS09_contactlens_decision_█('LH44',Mean7,STD7,Mean8,STD8,.02);
```

Figure H.15. Portion of student sample 1's code for LO 9.


```

function [dec] = PS09_contactlens_decision_----(lens_ID,bc_mean,bc_std,d_mean,d_std,threshold)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% ENGR 132 Program Description
% Program Description
% Determine if contact lens is acceptable or not.
%
% Function Call
% [dec] = PS09_contactlens_decision_----(lens_ID,bc_mean,bc_std,d_mean,d_std,threshold)
%
% Input Arguments
% 1. lens_ID
% 2. base curve mean
% 3. base curve standard deviation
% 4. diameter mean
% 5. diameter standard deviation
% 6. threshold
%
% Output Arguments
% 1. decision
%
% Assignment Information
% Assignment:      PS 09, Problem 3
% Author:         -----@purdue.edu
% Team ID:        -----
% Contributor:     N/A
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%
%% _____
%% CALCULATIONS & FORMATTED TEXT
bcrLR = bc_std/bc_mean; %Lens ratio for the base curve
dLR = d_std/d_mean; %Lens ratio for the diameter

if bcrLR < threshold && dLR < threshold
    dec = 1;
else
    dec = 0;
end

if dec == 1
    rdec = 'ACCEPTABLE';
else
    rdec = 'UNACCEPTABLE';
end

fprintf('\nLens Design %s is %s at threshold ratio %.2f.\n',lens_ID,rdec,threshold)
%%

```

Figure H.16. Portion of student sample 1's code for LO 10.

```
function[] = Solution_contactlens()
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
*REMOVED FOR SPACE*
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

INITIALIZATION

```
% Load the lens design data and set threshold value
lens_data = csvread('Data_newlensdesigns.csv',2,0);
threshold = 0.02; % contact lens threshold

% Create string variables for the design batch ID names
% to use in the pcode and in the plot legend
lens1 = 'LX18'; % lens design batch ID
lens2 = 'LF54'; % lens design batch ID
lens3 = 'LL107'; % lens design batch ID
lens4 = 'LA66'; % lens design batch ID
```

LENS DESIGN PLOT

REMOVED FOR SPACE

FUNCTION CALLS

```
% Call stats UDF to get mean and std dev on each lens parameter
% lens 1 (LX18) base curve radius & diameter stats (mm)
[rbc1_mean,rbc1_std] = Solution_stats_io(lens_data(:,1));
[dia1_mean,dia1_std] = Solution_stats_io(lens_data(:,2));
% lens 2 (LF54) base curve radius & diameter stats (mm)
[rbc2_mean,rbc2_std] = Solution_stats_io(lens_data(:,3));
[dia2_mean,dia2_std] = Solution_stats_io(lens_data(:,4));
% lens 3 (LL107) base curve radius & diameter stats (mm)
[rbc3_mean,rbc3_std] = Solution_stats_io(lens_data(:,5));
[dia3_mean,dia3_std] = Solution_stats_io(lens_data(:,6));
% lens 4 (LA66) base curve radius & diameter stats (mm)
[rbc4_mean,rbc4_std] = Solution_stats_io(lens_data(:,7));
[dia4_mean,dia4_std] = Solution_stats_io(lens_data(:,8));

% Call p-code to determine acceptability --this represents the
% quality control on the geometry in the manufacturing process
decLX18 =
    PS09_contactlens_decision_ (lens1,rbc1_mean,rbc1_std,dia1_mean,dia
    1_std,threshold);
decLF54 =
    PS09_contactlens_decision_ (lens2,rbc2_mean,rbc2_std,dia2_mean,dia
    2_std,threshold);
decLL107 =
    PS09_contactlens_decision_ (lens3,rbc3_mean,rbc3_std,dia3_mean,dia
    3_std,threshold);
decLA66 =
    PS09_contactlens_decision_ (lens4,rbc4_mean,rbc4_std,dia4_mean,dia
    4_std,threshold);
```

Figure H.17. Portion of student sample 2's code for LO 9.

```

function[dec] =
PS09_contactlens_decision_ (lens_ID,bc_mean,bc_std,d_mean,d_std,thres
hold)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% ENGR 132 Program Description
% Program Description
%     This function decided if the lens design is ideal or not. It
%     returns 1 if the design is ideal and 0 if not.
%
% Function Call
%
%     PS09_contactlens_decision_ (lens_ID,bc_mean,bc_std,d_mean,d_std,th
%     reshold)
%
% Input Arguments
%     1. lens_ID
%     2. bc_mean
%     3. bc_std
%     4. d_mean
%     5. d_std
%     6.threshold
%
% Output Arguments
%     1. dec
%
% Assignment Information
%     Assignment:      PS 09, Problem 3
%     Author:          @purdue.edu
%     Team ID:
%     Contributor:     Name, login@purdue [repeat for each]
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

```

CALCULATIONS & FORMATTED TEXT

```

lens_ratio_bc = bc_std / bc_mean; %Ratio between the mean base curve
ratio and standard base curve ratio
lens_ratio_d = d_std / d_mean; %Ratio between the mean diameter ratio
and standard diameter ratio

if lens_ratio_bc > threshold %comparing to the threshold
    dec = 1 ;
else
    dec = 0 ;
end

if lens_ratio_d > threshold %comparing to threshold
    dec = 1 ;
else
    dec = 0 ;
end

if dec == 1 %deciding if output if acceptable or unacceptable
    answer = 'ACCEPTABLE' ;
else
    answer = 'UNACCEPTABLE';
end

fprintf('\nLens Design %s is %s at threshold ratio
%f ' lens_ID answer threshold)

```

Figure H.18. Portion of student sample 2's code for LO 10.

```
function Solution_contactlens
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
*REMOVED FOR SPACE*
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

INITIALIZATION

```
*REMOVED FOR SPACE*

% Create string variables for the design batch ID names
% to use in the pcode and in the plot legend
lens1 = 'LX18'; % lens design batch ID
lens2 = 'LF54'; % lens design batch ID
lens3 = 'LL107'; % lens design batch ID
lens4 = 'LA66'; % lens design batch ID
```

LENS DESIGN PLOT

```
% Create the plot of the lens design data
*REMOVED FOR SPACE*
```

FUNCTION CALLS

```
% Call stats UDF to get mean and std dev on each lens parameter
% lens 1 (LX18) base curve radius & diameter stats (mm)
[rbc1_mean,rbc1_std] = Solution_stats_io(lens_data(:,1));
[dia1_mean,dia1_std] = Solution_stats_io(lens_data(:,2));
% lens 2 (LF54) base curve radius & diameter stats (mm)
[rbc2_mean,rbc2_std] = Solution_stats_io(lens_data(:,3));
[dia2_mean,dia2_std] = Solution_stats_io(lens_data(:,4));
% lens 3 (LL107) base curve radius & diameter stats (mm)
[rbc3_mean,rbc3_std] = Solution_stats_io(lens_data(:,5));
[dia3_mean,dia3_std] = Solution_stats_io(lens_data(:,6));
% lens 4 (LA66) base curve radius & diameter stats (mm)
[rbc4_mean,rbc4_std] = Solution_stats_io(lens_data(:,7));
[dia4_mean,dia4_std] = Solution_stats_io(lens_data(:,8));

% Call p-code to determine acceptability --this represents the
% quality control on the geometry in the manufacturing process
decLX18 =
contactlens_decision(lens1,rbc1_mean,rbc1_std,dia1_mean,dia1_std,threshold);
decLF54 =
contactlens_decision(lens2,rbc2_mean,rbc2_std,dia2_mean,dia2_std,threshold);
decLL107 =
contactlens_decision(lens3,rbc3_mean,rbc3_std,dia3_mean,dia3_std,threshold);
decLA66 =
```

Figure H.19. Portion of student sample 3's code for LO 9.

```

function PS09_contactlens_decision_
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% ENGR 132 Program Description
% Program Description
%
% Function Call
% ...
%
% Input Arguments
%     1. lens_id
%     2. mean_rad
%     3. std_rad
%     4. mean_diam
%     5. std_diam
%     6. threshold
%
% Output Arguments
%     1. accept
%
% Assignment Information
%     Assignment:      PS 09, Problem 3
%     Author:          @purdue.edu
%     Team ID:
%     Contributor:     Name, login@purdue [repeat for each]
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

```

CALCULATIONS & FORMATTED TEXT

REMOVED FOR SPACE

```

% Create string variables for the design batch ID names
% to use in the pcode and in the plot legend
lens1 = 'LM17'; % lens design batch ID
lens2 = 'LR283'; % lens design batch ID
lens3 = 'LP06'; % lens design batch ID
lens4 = 'LH44'; % lens design batch ID

```

FUNCTION CALLS

```

% Call stats UDF to get mean and std dev on each lens parameter
% lens 1 (LM17) base curve radius & diameter stats (mm)
[rbc1_mean,rbc1_std] = Solution_stats_io(lens_data(:,1));
[dial_mean,dial_std] = Solution_stats_io(lens_data(:,2));

*REMOVED FOR SPACE*

% Call p-code to determine acceptability --this represents the
% quality control on the geometry in the manufacturing process
decLM17 =
contactlens_decision(lens1,rbc1_mean,rbc1_std,dial_mean,dial_std,threshold);

```

Figure H.20. Portion of student sample 3's code for LO 10.

REFERENCES

- ABET. (2016). *Criteria for accrediting engineering programs, 2017–2018*.
<http://www.abet.org/wp-content/uploads/2016/12/E001-17-18-EAC-Criteria-10-29-16-1.pdf>
- Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice*, 18(3), 259–278.
<https://doi.org/10.1080/0969594X.2010.546775>
- Ai, H. and Lu, X. (2010). A web-based system for automatic measurement of lexical complexity. Paper presented at the 27th *Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10)*. Amherst, MA. June 8–12.
<https://doi.org/10.13140/RG.2.2.16499.07208>
- Ai, H. & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*, pp. 249–264.
<https://doi.org/10.1075/scl.59.15ai>
- Allen, J. (2005). Grades as valid measures of academic achievement of classroom learning. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 78(5), 218–223.
<https://doi.org/10.3200/TCHS.78.5.218-223>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, 57(5), 13–18.
http://www-tc.pbs.org/teacherline/courses/rdla230/docs/session_2_andrade.pdf
- Arffman, I. (2015). Threats to validity when using open-ended items in international achievement studies: Coding responses to the PISA 2012 problem solving test in Finland. *Scandinavian Journal of Educational Research*, 60(6), 609–625.
<https://doi.org/10.1080/00313831.2015.1066429>
- Arnold, J. (2002). Tensions between assessment, grading and development in development centres: A case study. *International Journal of Human Resource Management*, 13(6), 975–991.
<https://doi.org/10.1080/09585190210134318>

- Ashton, S., & Davies, R. S. (2015). Using scaffolded rubrics to improve peer assessment in a MOOC writing course. *Distance Education*, 36(3), 312–334. <https://doi.org/10.1080/01587919.2015.1081733>
- Baird, J.-A., Meadows, M., Leckie, G., & Caro, D. (2017). Rater accuracy and training group effects in Expert- and Supervisor-based monitoring systems. *Assessment in Education: Principles, Policy & Practice*, 24(1), 44–59. <https://doi.org/10.1080/0969594X.2015.1108283>
- Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, 18(3), 279–293. <https://doi.org/10.1080/0969594X.2010.526585>
- Baziuk, P. A., Nunez McLeod, J. E., & Rivera, S. S. (2018). Human reliability analysis based on human abilities theory model. *IEEE Transactions on Fuzzy Systems*, 26(2), 443–453. <https://doi.org/10.1109/tfuzz.2017.2685361>
- Berg, C. A. & Smith, P. (1994). Assessing students' abilities to construct and interpret line graphs: Disparities between multiple-choice and free-response instruments. *Science Education*, 78(6), 527–554. <https://doi.org/10.1002/sce.3730780602>
- Betts, J. R., & Costrell, R. M. (2001). Incentives and equity under standards-based reform. *Brookings Papers on Education Policy*, 2001(1), 9–74. <https://doi.org/10.1353/pep.2001.0001>
- Black, B., Suto, I., & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy & Practice*, 18(3), 295–318. <https://doi.org/10.1080/0969594X.2011.555328>
- Bloom, B. S. (1956). *Taxonomy of educational objectives. Vol. 1: Cognitive domain*. McKay.
- Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education*, 36(6), 655–670. <https://doi.org/10.1080/03075071003777716>
- Boren, M. T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, 43(3), 261–278. <https://doi.org/10.1109/47.867942>
- Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, 13(1), 10–18. <https://doi.org/10.2307/1164948>
- Bresciani, M. J., Oakleaf, M., Kolkhorst, F., Nebeker, C., Barlow, J., Duncan, K., & Hickmont, J. (2009). Examining design and inter-rater reliability of a rubric measuring research quality across multiple disciplines. *Practical Assessment, Research & Evaluation*, 14(12), 1–7. <http://pareonline.net/getvn.asp?v=14&n=12>

- Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, 9(2004), 105–121. <https://doi.org/10.1016/j.asw.2001.07.001>
- Charney, D. (1984). The validity of using holistic scoring to evaluate writing: a critical overview. *Research in the Teaching of English*, 18, 65–81.
- Choi, H.-H., van Merriënboer, J. J. G., & Paas, F. (2014). Effects of the physical environment on cognitive load and learning: Towards a new model of cognitive load. *Educational Psychology Review*, 26(2), 225–244. <https://doi.org/10.1007/s10648-014-9262-6>
- Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37(2), 163–178. <https://doi.org/10.1111/j.1745-3984.2000.tb01081.x>
- Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment=in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation*, 13(5), 401–434. <https://doi.org/10.1080/13803610701728311>
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Sage Publications, Inc.
- Crisp, V. (2010). Judging the grade: Exploring the judgment processes involved in examination grading decisions. *Evaluation and Research in Education*, 23(1), 19–35. <https://doi.org/10.1080/09500790903572925>
- Crocker, J., Karpinski, A., Quinn, D. M., & Chase, S. K. (2003). When grades determine self-worth: Consequences of contingent self-worth for male and female engineering and psychology majors. *Journal of Personality and Social Psychology*, 85(3), 507–516. <https://doi.org/10.1037/0022-3514.85.3.507>
- Cumming, A. (1990) Expertise in evaluating second language compositions. *Language Testing*, 7, 31–51.
- Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., ... Steele, C. M. (2013). *Criteria for high-quality assessment*. Stanford Center for Opportunity Policy in Education.
- Diefes-Dux, H. A., Zawojewski, J. S., & Hjalmarson, M. A. (2010). Using educational research in the design of evaluation tools for open-ended problems. *International Journal of Engineering Education*. 26(4), 807–819. https://www.researchgate.net/profile/Margret-Hjalmarson/publication/228421234_Using_Educational_Research_in_the_Design_of_Evaluation_Tools_for_Open-Ended_Problems/links/00b7d52323601bc945000000/Using-Educational-Research-in-the-Design-of-Evaluation-Tools-for-Open-Ended-Problems.pdf

- Di Pasquale, V., Miranda, S., Iannone, R. & Riemma, S. (2015). A Simulator for Human Error Probability Analysis (SHERPA). *Reliability Engineering & System Safety*, 139(1), 17–32. <https://doi.org/10.1016/j.res.2015.02.003>
- Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing*, 11(2), 125–144. <https://doi.org/10.1177/026553229401100203>
- Douglas, E. P., Koro-Ljungberg, M., McNeill, N. J., Malcolm, Z. T., & Therriault, D. J. (2012). Moving beyond formulas and fixations: Solving open-ended engineering problems. *European Journal of Engineering Education*, 37(6), 627–651. <https://doi.org/10.1080/03043797.2012.738358>
- Douglas, K. A., Moore, T. J., Merzdorf, H. E., Lee, T., & Johnston, A. C. (2017, June). Content analysis of how engineering is assessed in published curricula. Proceedings in the 124th ASEE Annual Conference & Exposition, Columbus, OH.
- Douglas, K. A., & Purzer, Ş. (2015). Validity: Meaning and relevancy in assessment for engineering education research. *Journal of Engineering Education*, 104(2) 108–118. <https://doi.org/10.1002/jee.20070>
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. MIT Press.
- Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, 5(3), 178–186. https://doi.org/10.1207/215327884mca0503_3
- Forsythe, C., Liao, H., Trumbo, M. C. S., & Cardona-Rivera, R. E. (2015). *Cognitive neuroscience of human systems: Work and everyday life*. CRC Press.
- Garza, R. T., & Lipton, J. P. (1978). Culture, personality, and reactions to praise and criticism. *Journal of Personality*, 46(4), 743–761. <https://doi.org/10.1111/j.1467-6494.1978.tb00195.x>
- Ghaboussi, J., & Insana, M. F. (2018). *Understanding systems: A grand challenge for 21st century engineering*. World Scientific Publishing Co. Pte. Ltd.
- Gibbs, G. (2006). Why assessment in changing. In C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education* (pp. 11–22). Routledge.
- Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education*, 24, 355–392. <https://doi.org/10.2307/1167274>
- Greator, J., & Sütő, W. M. I. (2008). What do GCSE examiners think of ‘thinking aloud’? Findings from an exploratory study. *Educational Research*, 50(4), 319–331. <https://doi.org/10.1080/00131880802499720>

- Griswold, P. A. (2010). Beliefs and influences about grading elicited from student performance sketches. *Educational Assessment*, 1(4), 311–328. https://doi.org/10.1207/s15326977ea0104_2
- Guskey, T. R. (2001). Helping standards make the grade. *Educational Leadership*, 59(1), 20–27. https://uknowledge.uky.edu/edp_facepub/8
- Hansen, E. J. (2011). *Idea-based learning: A course design process to promote conceptual understanding*. Stylus Publishing, LLC.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–333. https://doi.org/10.1207/S15324814AME1503_5
- Hicks, N. M., & Diefes-Dux, H. A. (2017, June). Grader consistency using learning objective-based rubrics. Proceedings in the 124th ASEE Annual Conference & Exposition, Columbus, OH.
- Hoc, J.-M. (2001). Towards ecological validity of research in cognitive ergonomics, *Theoretical Issues in Ergonomics Science*, 2(3), 278–288, <https://doi.org/10.1080/14639220110104970>
- Hollnagel, E. (1998). *Cognitive reliability and error analysis method*. Elsevier Science Inc.
- Hollnagel, E. (2012). FRAM: the Functional Resonance Analysis Method: Modeling complex socio-technical systems. CRC Press.
- Hollnagel, E. (n.d.). *The ETTO principle – Efficiency-Thoroughness Trade-Off*. <http://erikhollnagel.com/ideas/etto-principle/index.html>
- Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring A mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy & Practice*, 18(3), 239–258. <https://doi.org/10.1080/0969594X.2011.577408>
- Johnson, M. (2008). Grading in competence-based qualifications – is it desirable and how might it affect validity? *Journal of Further & Higher Education*, 32(2), 175–184. <https://doi.org/10.1080/03098770801979183>
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. *Educational Research Review*, 2(2), 130–144. <https://doi.org/10.1016/j.edurev.2007.05.002>
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 49–81). Cambridge University Press.

- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535. <https://doi.org/10.1037/0033-2909.112.3.527>
- Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Karimi, A. (2015, November). *Bringing uniformity in topic coverage and grading fairness in multiple sections of an engineering course*. In Proceedings of the ASME 2015 International Mechanical Engineering Congress and Exposition, Houston, TX, USA. <http://proceedings.asmedigitalcollection.asme.org/>
- Knight, P. T. (2002). Summative assessment in higher education: Practices in disarray. *Studies in Higher Education*, 27(3), 275–286. <https://doi.org/10.1080/03075070220000662>
- Kohn, A. (2011). The case against grades. *Educational Leadership*, 69(3), 28–33. <https://doi.org/10.1136/bmj.314.7079.503>
- Krahmer, E. & Ummelen, N. (2004). Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication*, 47(2), 105–117. <https://doi.org/10.1109/tpc.2004.828205>
- Kurdziel, J., Turner, J., Luft, J., & Roehrig, G. (2003). Graduate teaching assistants and inquiry-based instruction: Implications for graduate teaching assistant training. *Journal of Chemical Education*, 80(10), 1206–1210. <https://doi.org/10.1021/ed080p1206>
- Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgments. *Quarterly Journal of Experimental Psychology*, 42A, 239–254. - English, mathematics, and physics higher ed essay examination
- Lengh, C. J. (2010). Generalizability theory: Measuring the dependability of selected methods for scoring classroom assessments (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3438129)
- Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, 66(1), 799–823. <https://doi.org/10.1146/annurev-psych-010213-115043>
- Liu, P., Lyu, X., Qiu, Y., Hu, J., Tong, J., & Li, Z. (2017). Identifying macrocognitive function failures from accident reports: A case study. In S. Cetiner, P. Fechtelkötter, & M. Legatt (Eds.), *Advances in human factors in energy: Oil, gas, nuclear and electric power industries* (pp. 29–40). https://doi.org/10.1007/978-3-319-41950-3_3
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496. <https://doi.org/10.1075/ijcl.15.4.02lu>

- Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1), 36–62. <https://doi.org/10.5054/tq.2011.240859>
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2), 190–208. <https://doi.org/10.1111/j.1540-4781.2011.01232.1.x>
- Lu, X. & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, 29(1), 16–27. <https://doi.org/10.1016/j.jslw.2015.06.003>
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19(3), 246–276. <https://doi.org/10.1191/0265532202lt230oa>
- Main, J. B., Mumford, K. J., & Ohland, M. W. (2015). Examining the influence of engineering students' course grades on major choice and major switching behavior. *International Journal of Engineering Education*, 31(6A), 1468–1475. <https://www.krannert.purdue.edu/faculty/kjmumfor/papers/Engineering%20Grades%20and%20Majors.pdf>
- Malicky, D. (2003, June). A literature review on the under-representation of women in undergraduate engineering: Ability, self-efficacy, and the “chilly climate”. Proceedings in the 110th ASEE Annual Conference & Exposition, Nashville, TN. <https://peer.asee.org/a-literature-review-on-the-underrepresentation-of-women-in-undergraduate-engineering-ability-self-efficacy-and-the-chilly-climate.pdf>
- Marzano, R. J. (2000). *Transforming classroom grading*. Association for Supervision and Curriculum Development.
- Marzano, R. J. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education*, 15(3), 249–268. https://doi.org/10.1207/S15324818AME1503_2
- Meier, S. L., Rich, B. S., & Cady, J. (2006). Teachers' use of rubrics to score non-traditional tasks: Factors related to discrepancies in scoring. *Assessment in Education*, 13(1), 69–95. <https://doi.org/10.1080/09695940600563512>
- Menéndez-Varela, J.-L., Gregori-Giralt, E. (2018). The reliability and sources of error of using rubrics-based assessment for student projects. *Assessment & Evaluation in Higher Education*, 43(3), 488–499. <https://doi.org/10.1080/02602938.2017.1360838>
- Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037//0003-066X.50.9.741>
- Milanovic, M., Saville, N. & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Studies in language testing 3*:

Performance testing, cognition and assessment—selected papers from the 15th Language Testing Research Colloquium. Cambridge University Press.

- Moskal, B. M., & Leydens, J. A. (2000). Scoring rubrics development: Validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10), 1–6. <http://PAREonline.net/getvn.asp?v=7&n=10>
- Muñoz, M. A., & Guskey, T. R. (2015). Standards-based grading and reporting will improve education. *Phi Delta Kappan*, 96(7), 64–68. <https://doi.org/10.1177/0031721715579043>
- National Academies of Sciences, Engineering, and Medicine, (2016). *Barriers and opportunities for 2-year and 4-year STEM degrees: Systemic change to support students' diverse pathways*. The National Academies Press. <https://doi.org/10.17226/21739>
- National Academies of Sciences, Engineering, and Medicine. (2018). *Indicators for monitoring undergraduate STEM education*. The National Academies Press. <https://doi.org/10.17226/24943>
- National Academy of Engineering. (2004). *The engineer of 2020: visions of engineering in the new century*. The National Academies Press. <https://doi.org/10.17226/10999>
- National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2007). *Beyond bias and barriers: Fulfilling and potential of women in academic science and engineering*. The National Academies Press. <https://doi.org/10.17226/11741>
- National Council on Measurement in Education. (2019). *Classroom assessment standards*. <https://www.ncme.org/about/classroom-assessment/task-force-standards>
- National Research Council. (1993). *Measuring what counts: A policy brief*. National Academy Press
- National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. The National Academies Press. <https://doi.org/10.17226/18409>
- Newton, P. E. (2005). The public understanding of measurement inaccuracy. *British Journal of Educational Research*, 31(4), 419–442. <https://doi.org/10.1080/01411920500148648>
- Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Merrill/Prentice Hall.
- Oakleaf, M. (2009). Using rubrics to assess information literacy: An examination of methodology and interrater reliability. *Journal of the American Society for Information Science and Technology*, 60(5), 969–983. <https://doi.org/10.1002/asi.21030>
- Pantzare, A. L. (2015). Interrater reliability in large-scale assessments – Can teachers score national tests reliably without external controls? *Practical Assessment, Research and Evaluation*, 20(9). <http://pareonline.net/getvn.asp?v=20&n=9>

- Patriarca, R., Bergström, J., & Di Gravio, G. (2017). Defining the functional resonance analysis space: Combining Abstraction Hierarchy and FRAM. *Reliability Engineering and System Safety*, 165(1), 34–46. <https://doi.org/10.1016/j.ress.2017.03.032>
- Porter, A. C., & Smithson, J. L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. *Yearbook-National Society for the Study of Education*, 2, 60–80.
- Price, M., & Rust, C. (1999). The experience of introducing a common criteria assessment grid across academic departments. *Quality in Higher Education*, 5(2), 133–144. <https://doi.org/10.1080/1353832990050204>
- Rasmussen, J. (1985). Trends in human reliability analysis. *Ergonomics*, 28(8), 1185–1195. <https://doi.org/10.1080/00140138508963241>
- 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>
- Ricco, G. D., Salzman, N., Long, R. A., & Ohland, M. W. (2012). Sectionality or why section determines grades: An exploration of engineering core course section grades using a hierarchical linear model and the Multiple-Institution Database for Investigating Engineering Longitudinal Development (Paper 20). Retrieved from School of Engineering Education Graduate Student Series: <http://docs.lib.purdue.edu/enegs/20>
- Rittel, H. W. J., & Webber, M. M. (1984). Planning problems are wicked problems. In N. Cross (Ed.), *Developments in design methodology* (pp. 135–144). John Wiley & Sons.
- Russell, J., Van Horne, S., Ward, A. S., Bettis, E. A., & Gikonyo, J. (2017). Variability in students' evaluating processing in peer assessment with calibrated peer review. *Journal of Computer Assisted Learning*, 33(2), 178–190. <https://doi.org/10.1111/jcal.12176>
- Rust, C. (2011). The unscholarly use of numbers in our assessment practices: What will make us change? *International Journal for Scholarship of Teaching and Learning*, 5(1), 1–6. <https://doi.org/10.20429/ijsotl.2011.050104>
- Sadler, D. R. (2009a). Grade integrity and the representation of academic achievement. *Studies in Higher Education*, 34(7), 807–826. <https://doi.org/10.1080/03075070802706553>
- Sadler, D. R. (2009b). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 34(2), 159–179. <https://doi.org/10.1080/02602930801956059>
- Sadler, D. R. (2010). Fidelity as a precondition for integrity in grading academic achievement. *Assessment & Evaluation in Higher Education*, 35(6), 727–743. <https://doi.org/10.1080/02602930902977756>
- Saldaña, J. (2016). *The coding manual for qualitative researchers* (3rd ed.). SAGE Publications, Inc.

- Schunn, C., Godley, A., & DeMartino, S. (2016). The reliability and validity of peer review of writing in high school AP English classes. *Journal of Adolescent and Adult Literacy*, 60(1), 13–23. <https://doi.org/10.1002/jaal.525>
- Sharit, J. (2012). Human error and human reliability analysis. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (4th ed., pp. 734–800). <https://doi.org/10.1002/9781118131350.ch26>
- Smith, M. W., & Stein, M. K. (1998). Reflections on practice: Selecting and creating mathematical tasks: From research to practice. *Mathematics Teaching in the Middle School*, 3(5), 344–350. <https://www.jstor.org/stable/41180423>
- 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>
- Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimate interrater reliability. *Practical Assessment, Research and Evaluation*, 9(4), 111. <http://pareonline.net/getvn.asp?v=9&n=4>
- Suto, W. M. I., Greatorex, J., & Nádas, R. (2009). Thinking about making the right mark: Using cognitive strategy research to explore examiner training. *Research Matters: A Cambridge Assessment Publication* (8), 23–32.
- Suto, I., & Nádas, R. (2007, September). The ‘Marking Expertise’ projects: Empirical investigations of some popular assumptions. Paper presented at the annual conference of the International Association for Educational Assessment, Baku, Azerbaijan.
- Suto, W. M. I., & Nádas, R. (2009). Why are some GSCE examination questions harder to mark accurately than others? Using Kelly’s Repertory Grid technique to identify relevant question features. *Research Papers in Education*, 24(3), 335–377. <https://doi.org/10.1080/02671520801945925>
- Suto, I., & Nádas, R. (2010). Investigating examiners’ thinking: Using Kelly’s Repertory Grid technique to explore cognitive marking strategies. *Research in Education*, 84(1), 38–53. <https://doi.org/10.7227/rie.84.3>
- Suto, I., Nádas, R., & Bell, J. (2011). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*, 26(1), 21–51. <https://doi.org/10.1080/02671520902721837>
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295–312. <https://doi.org/10.1016%2F0959-4752%2894%2990003-5>

- Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, 22(2), 123–138. <https://doi.org/10.1007/s10648-010-9128-5>
- Tekkumru-Kisa, M. Stein, M. K., & Schunn, C. (2015). A framework for analyzing cognitive demand and content-practices integration: Task Analysis Guide in Science. *Journal of Research in Science Teaching*, 52(5), 659–685. <https://doi.org/10.1002/tea.21208>
- Thorndike, R. M. (1997). The early history of intelligence testing. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 3–16). Guilford Press.
- Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Pearson Education, Inc.
- Vaughan, C. (1992). Holistic assessment: what goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*. Ablex.
- Webster, F., Pepper, D. & Jenkins, A. (2000). Assessing the undergraduate dissertation. *Assessment and Evaluation in Higher Education*, 25(1), 71–80. <https://doi.org/10.1080/02602930050025042>
- Wiggins, G. & McTighe, J. (2005). *Understanding by design* (2nd ed.). Association for Supervision and Curriculum Development.
- Woodley, J., Yeaton, K., & Hutchins, T. (2017). Rubric scales: How a 'zero' scoring option may alter rater choices. *Journal of Higher Education Theory and Practice*, 17(8), 81–88. http://www.na-businesspress.com/JHETP/JHETP17-8/WoodkleyJ_17_8_.pdf
- Wyatt-Smith, C., & Klenowski, V. (2013). Explicit, latent and meta-criteria: Types of criteria at plan in professional judgement practice. *Assessment in Education: Principles, Policy & Practice*, 20(1), 35–52. <https://doi.org/10.1080/0969594X.2012.725030>