Paper ID #27732

Understanding Grader Reliability through the Lens of Cognitive Modeling

Nathan M. Hicks, Purdue University, West Lafayette

Nathan M. Hicks is a Ph.D. student in Engineering Education at Purdue University. He received his B.S. and M.S. degrees in Materials Science and Engineering at the University of Florida and taught high school math and science for three years.

Dr. Kerrie A. Douglas, Purdue University, West Lafayette

Dr. Douglas is an Assistant Professor in the Purdue School of Engineering Education. Her research is focused on improving methods of assessment in large learning environments to foster high-quality learning opportunities. Additionally, she studies techniques to validate findings from machine-generated educational data.

Prof. Heidi A. Diefes-Dux, University of Nebraska, Lincoln

Heidi A. Diefes-Dux is a Professor in Biological Systems Engineering at the University of Nebraska - Lincoln. She received her B.S. and M.S. in Food Science from Cornell University and her Ph.D. in Food Process Engineering from the Department of Agricultural and Biological Engineering at Purdue University. She was an inaugural faculty member of the School of Engineering Education at Purdue University. Her research focuses on the development, implementation, and assessment of modeling and design activities with authentic engineering contexts. She also focuses on the implementation of learning objective-based grading and teaching assistant training.

WIP: Understanding Grader Reliability through the Lens of Cognitive Modeling

Introduction

The evaluation of student learning, whether formal or informal, is essential to the educational process as we know it. The results of such evaluation can communicate valuable information that may inform highly consequential decisions for individual students, instructors, curriculum designers, administrators, and even policy makers. With such significant consequences, it is imperative that evaluation data, often communicated through letter grades (i.e., discrete, ordinal classifications of student performance [1]), be as meaningful and trustworthy as possible. Unfortunately, the components of grades and corresponding standards of success often vary so extensively across instructors that grades are rendered effectively meaningless [2].

The meaning and trustworthiness of an assessment score or grade, often referred to as its validity, absolutely requires reliability—that is, the consistency of a score, regardless of when the assessment is conducted, when it is scored, or by whom it is scored [3], [4]. Unfortunately, attaining reliable assessment scores can be challenging in the many engineering courses that utilize open-ended performance tasks to authentically assess the engineering competencies called for by ABET and the Engineer of 2020 [5] – [7]. This challenge increases with larger class sizes, which are often encountered in first-year engineering programs [8]. The strain on resources imposed by large courses frequently necessitates grading by less expensive—and less experienced—graduate and undergraduate teaching assistants, whose inexperience often leads to a variety of grading inconsistencies [9], [10].

Based in the context of a large-scale engineering course, this study frames the grading process as a large, complex socio-technical system consisting of elements that are both human (i.e., graders, students, and content designers) and non-human (i.e., rubrics, assignments, and student work). As the underlying goal is to achieve a grading system that is both consistent and accurate, this study seeks to develop a deep understanding of how the system functions and which aspects of the system contribute the most to variable performance. Thus, the work-in-progress study described in this paper involves a Human Reliability Analysis (HRA) of grading open-ended engineering problems by many graders. More specifically, it explores how the grading system can be modeled and what contributes to model variability.

Background

Previous studies exploring the reliability of grading tend to be completely quantitative in nature, focusing on various measures of inter-rater reliability and consistency [11] – [13]. Some look at grading schemes, e.g., [14] – [16], while others focus on grading tools like rubrics, e.g., [17] – [19], or the graders themselves and how they think, e.g., [10], [20] – [23]. None of these treatments, however, fully captures the complexity of the entire grading system, particularly when many graders are involved. Recognizing that variable outputs of the grading system initiate from variable decision making on the part of the systems' human components makes HRA an appropriate analysis technique to explore grading variability.

There are three generations of HRA techniques: the first generation focused on quantification of success and failure probabilities but mostly ignored underlying causes; the second generation shifted the focus to the underlying causes of human error; and the third generation evolved beyond the static systems of the first two generations to better handle more dynamic, socio-technical systems [24], [25]. These analysis techniques derived from heavy industrial settings, primarily studying systems in which "human error" was rare but extremely costly. The educational setting of a grading system differs: errors are relatively common but can be fixed easily and have minimal consequences if caught. Still, in a large course with many graders grading a wide range of problems, the system is highly complex and dynamic.

Grading open-ended performance tasks is inherently subjective and the process can vary significantly depending on the task or level of performance; thus, this study employs a relatively new approach called the Functional Resonance Analysis Method (FRAM) due to its strength with dynamic systems and its emphasis on variable human performance rather than the more traditional interpretation of "human error" [25]. Unlike older techniques that consider a static process and calculate error probabilities at several points throughout the process, the FRAM provides greater flexibility for variable systems [25]. Rather than creating a static model, the FRAM creates separate instantiations for every iteration of the process. Depending on the specific circumstances under which the process occurs, each instantiation might contain different elements occurring in any order. Therefore, developing a model of the grading system first involves exploration of all the elements (referred to as "functions") that occur in the foreground of grading, as well as in the background.

Methods

The overall research study has four stages. The first stage has already been completed and involved conducting direct observation of the grading process through a set of think-aloud interviews with undergraduate graders grading actual student work. The second stage, which is currently in progress, involves the qualitative analysis of the think-aloud data to develop a model of the grading process using the FRAM. Following this analysis, a third stage will use quantitative grading data from the course to determine the extent to which the models generalize to more realistic settings and a wider range of problems, rubrics, and student work. In the final stage, these analyses will be processed to create a set of recommendations to reduce variability.

Context. This study is contextualized in a first-year engineering program at a large public university. Analysis is centered in the second course of a two-semester sequence of courses. This course typically has over a dozen sections of over 100 students each spring semester. Each section usually employs an instructor, a graduate teaching assistant, four undergraduate peer teachers, and two undergraduate graders. For any given assessment, each grader typically grades one-third of the section responses and the peer teachers split the remaining responses. All of the grading is generally overseen by the graduate teaching assistant, though specifics vary by section.

Throughout the semester, students complete a collection of near-weekly problem sets. These problem sets are graded by the undergraduate peer teachers and graders using extensive rubrics based on specific learning objectives and evidence items of proficiency. Students are only told the learning objectives covered and cannot see the detailed rubrics [26]. Prior to the grading of each problem set, all first-time undergraduate peer teachers and graders are expected to complete training modules for each new learning objective in the upcoming problem set.

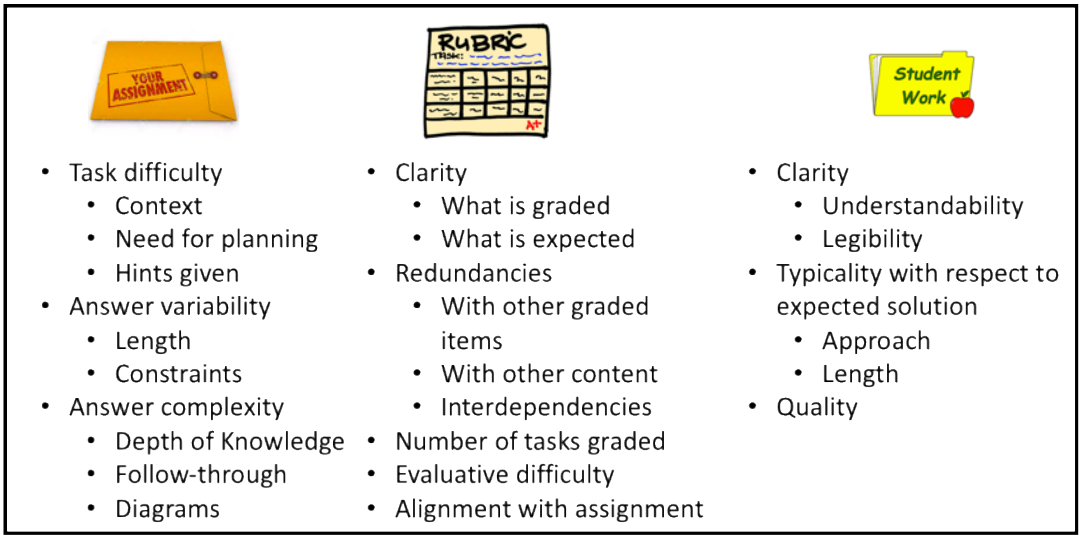
Study participants. After contacting all 76 undergraduate peer teachers and graders, 17 agreed to participate in the think-aloud studies. The participants ranged between second- and fifth-year engineering students from various majors and had between two and eight semesters of experience assisting with the course. They had each passed the course with a grade of B or higher. Each participant was given $20 at the conclusion of the interview.

Think-aloud interviews. Think-aloud interviews (i.e., interviews in which participants verbalize their thought processes while performing tasks) were conducted in the spring of 2017 following the suggestions of Boren and Ramey [27]. The interviews lasted approximately one hour and consisted of grading three real, de-identified student responses for each learning objective in one of the course's problem sets. The problem set used in the interview was that which had the lowest average accuracy with respect to the "definitive grades" during training. This decision was made following the assumption that lower accuracy in training corresponded to items that were harder to grade consistently, thereby being more likely to demonstrate greater variability in cognitive processes. The three student response samples for each learning objective included in the interview documents were purposefully selected to represent a range of solution approaches and levels of achievement as well as prompt alternative cognitive grading processes on the part of the graders. Data was collected using the Notability App on an iPad, which records both audio and notations made on the interview documents (i.e., the sample responses and rubrics). Audio recordings were transcribed and checked for accuracy and notes were taken by the interviewer at the time of the interviews.

Qualitative modeling with the FRAM. The FRAM consists of four steps: (1) function identification and description, (2) variability identification, (3) variability aggregation, and (4) control mechanism identification [25]. The functions that comprise each model, identified and defined in the first step, represent all actions that occur within the system. Each function is characterized by up to six factors: input(s), output(s), precondition(s), resource(s) or executive condition(s), control(s), and time. A function may be a foreground function if it is the primary process of concern or a background function if it affects the process but is not directly involved. The first three steps of the FRAM in this work are achieved through the coding of the interview documents (i.e., the assignment problems, rubrics, and sample responses) and the think-aloud interviews.

The background functions were identified and described by qualitative coding of the assignments, the rubrics, and the sample responses. These documents were coded by the primary author using a thorough codebook based on the literature regarding aspects of assignments, rubrics, and student work that have been shown to affect grading accuracy (i.e., [21], [22], and [28]), in conjunction with an open-coding option for new, emergent functions. Figure 1 summarizes the factors identified in the literature. Each variable aspect associated with these documents corresponds to a function decision function associated with the design of that document. The second and third authors provided a check on coding quality and trustworthiness through discussion and agreement seeking regarding coding decisions.

The foreground functions, along with new potential background functions that may have been missed in the literature, are currently being identified through coding of the cognitive processes demonstrated by the participants during the think-aloud interviews. The cognitive strategies and grading stages identified in [22] and [28], namely the cognitive strategies of identifying whether or not there is a response, scanning, matching, scrutinizing, or evaluating, served as initial *a priori* codes, but many new codes have been added to capture additional processes and greater nuance. The appendix includes a list and short description of all functions identified at the time of this publication.



- Task difficulty
  - Context
  - Need for planning
  - Hints given
- Answer variability
  - Length
  - Constraints
- Answer complexity
  - Depth of Knowledge
  - Follow-through
  - Diagrams

- Clarity
  - What is graded
  - What is expected
- Redundancies
  - With other graded items
  - With other content
  - Interdependencies
- Number of tasks graded
- Evaluative difficulty
- Alignment with assignment

- Clarity
  - Understandability
  - Legibility
- Typicality with respect to expected solution
  - Approach
  - Length
- Quality

Figure 1. Coded aspects of assignments, rubrics, and student work.

After all of the interviews have been coded, each interview will consist of many instantiations of the grading process model. Figure 2 shows a simplified example of such an instantiation, where the letters in the circles around each function represent the factors characterizing the functions (inputs, time, controls, outputs, resources, and preconditions). In the example, all "assignment," "rubric," and "student work," functions are aggregated into single functions, but in a full model, these would be composed of multiple background functions. For any given learning objective and student response, the "assignment," "student work," and "rubric" functions would be static, but the grading process functions (the foreground functions of "no response," "scanning," and "matching" in Figure 2) might vary from grader to grader. Thus, there will be separate foreground model instantiations for each of the over 500 grades assigned during the think-aloud interviews. It is expected, however, that many of these will be similarly structured, but will help to illustrate how functions can vary in practice (FRAM step 2) and how that variability can aggregate through the system (FRAM step 3).

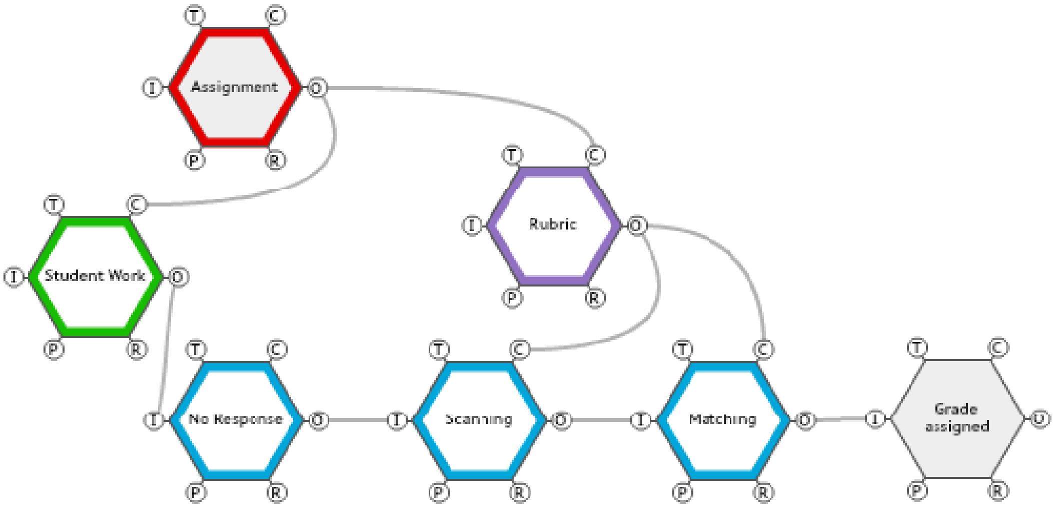


Figure 2. A simplified example of a grading process model instantiation with the FRAM.

Future Steps

Ultimately, the collection of all of these instantiations, along with the secondary quantitative analyses mentioned previously, will allow for the final step of the FRAM: identification of possible control mechanisms. Through this process, common trends will be explored to identify which functions lead to the most variability in the system and the outcome. This analysis should give indications of design decisions that can be made to reduce the corresponding variability. For instance, if variability originating in an assignment function leads to large variability of outcome, that can inform aspects of assignment design. While many of the specific aspects of this project are rooted in the context of the study, the final recommendations will be stated as generally as the analysis will allow.

References

[1] M. Johnson, "Grading in competence-based qualifications – is it desirable and how might it affect validity?," *J. Further High. Educ.*, vol. 32, no. 2, pp. 175–184, Apr. 2008. doi: 10.1080/03098770801979183. [Accessed: May 22, 2018].

[2] R. J. Marzano, *Transforming Classroom Grading*. Alexandria, VA, USA: Association for Supervision and Curriculum Development, 2000.

[3] AERA, APA, & NCME, *Standards for Educational and Psychological Testing*. Washington, DC, USA: American Educational Research Association, 2014.

[4] S. Messick, "Validity of psychological assessment," *Am. Psychol*, vol. 50, no. 9, pp. 741–749, Sept. 1995. doi: 10.1037//0003-066X.50.9.741. [Accessed: Apr. 6, 2018].

[5] I. Arffman, "Threats to validity when using open-ended items in international achievement studies: Coding responses to the PISA 2012 problem solving test in Finland," *Scand. J. Educ. Res.*, vol. 60, no. 6, pp. 609–625, Sept. 2015 doi: 10.1080/00313831.2015.1066429. [Accessed: Jun. 21, 2018].

[6] ABET. (2016, Oct. 29). *Criteria for Accrediting Engineering Programs, 2017–2018 [Online]*. Available: http://www.abet.org/wp-content/uploads/2016/12/E001-17-18-EAC-Criteria-10-29-16-1.pdf. [Accessed: Apr. 8, 2018].

[7] NAE, *The Engineer of 2020: Visions of Engineering in the New Century*. Washington, DC: The National Academies Press, 2004. doi: 10.17226/10999.

[8] C. Gipps, "Socio-cultural aspects of assessment," *Rev. Res. Educ.*, vol. 24, pp. 355–392, 1999. doi: 10.2307/1167274. [Accessed: May 22, 2018].

[9] H. I. Braun, "Understanding scoring reliability: Experiments in calibrating essay readers," *J. Educ. Stat.*, vol. 13, no. 1, pp. 10–18, 1988. doi: 10.2307/1164948. [Accessed: Jul. 12, 2017].

[10] V. Crisp, "Judging the grade: Exploring the judgment processes involved in examination grading decisions," *Eval. Res. Educ.*, vol. 23, no. 1, pp. 19–35, Mar. 2010. doi: 10.1080/09500790903572925. [Accessed: Jul. 12, 2017].

[11] M. Oakleaf, “Using rubrics to assess information literacy: An examination of methodology and interrater reliability,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 5, pp. 969–983, Feb. 2009. doi:10.1002/asi.21030. [Accessed: Jul. 12, 2017].

[12] A. L. Pantzare, “Interrater reliability in large-scale assessments – Can teachers score national tests reliably without external controls?,” *Pract. Assess. Res. Eval.*, vol. 20, no. 9, Apr. 2015. Available: http://pareonline.net/getvn.asp?v=20&n=9. [Accessed: Jul. 12, 2017].

[13] S. E. Stemler, “A comparison of consensus, consistency, and measurement approaches to estimate interrater reliability,” *Pract. Assess. Res. Eval*, vol. 9, no. 4, Mar. 2004. Available: http://pareonline.net/getvn.asp?v=9&n=4. [Accessed: Jul. 12, 2017].

[14] C. J. Lengh, “Generalizability theory: Measuring the dependability of selected methods for scoring classroom assessments,” Ph. E. dissertation, Coll. Educ. Lead., Card. Stritch Univ., Milwaukee, WI. 2010. [Online]. Available: ProQuest Dissertations and Theses database. (UMI No. 3438129). [Accessed: Jun. 28, 2018].

[15] R. J. Marzano, “A comparison of selected methods of scoring classroom assessments,” *Appl. Meas. Educ.*, vol. 15, no. 3, pp. 249–268, 2002. doi: 10.1207/S15324818AME1503_2. [Accessed: Jun. 28, 2018].

[16] M. K. Thompson, L. H. Clemmensen and B.-U. Ahn, “Effect of rubric rating scale on the evaluation of engineering design projects,” *Int. J. Eng. Educ.*, vol. 29, no. 6, pp. 1490–1502, 2013. Available: http://orbit.dtu.dk/files/60489369/Effect_of_Rubric_Rating_Scale.pdf. [Accessed: Jun. 28, 2018].

[17] J. A. Baird., M. Meadows, G. Leckie, and D. Caro, “Rater accuracy and training group effects in Expert- and Supervisor-based monitoring systems,” *Assess. Educ.: Princ. Policy Pract.*, vol. 24, no. 1, pp. 44–59, Dec. 2015. doi: 10.1080/0969594X.2015.1108283. [Accessed: Jul. 16, 2018].

[18] K. Barkaoui, “Effects of marking method and rater experience on ESL essay scores and rater performance,” *Assess. Educ.: Princ. Policy Pract.*, vol. 18, no. 3, pp. 279–293, Aug. 2011. doi: 10.1080/0969594X.2010.526585. [Accessed: Jun. 22, 2018].

[19] Y. M. Reddy, and H. Andrade, “A review of rubric use in higher education,” *Assess. Eval. Higher Educ*, vol. 35, no. 4, pp. 435–448 Jul. 2010. doi:10.1080/02602930902862859. [Accessed: Jun. 22, 2018].

[20] R. W. Cooksey, P. Freebody, and C. Wyatt-Smith, “Assessment as judgment-in-context: Analysing how teachers evaluate students’ writing,” *Educ. Res. Eval.*, vol. 13, no. 5, pp. 401–434 Oct. 2007. doi:10.1080/13803610701728311. [Accessed: Jul. 12, 2017].

[21] B. Black, I. Suto, and T. Bramley, “The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement,” *Assess. Educ.: Princ. Policy Pract.*, vol. 18, no. 3, pp. 295–318, Aug. 2011. doi: 10.1080/0969594X.2011.555328. [Accessed: Jun. 22, 2018].

[22] W. M. I. Suto, and J. Greatorex, “What goes through an examiner’s mind? Using verbal protocols to gauge insights into the GCSE marking process,” *Br. Educ. Res. J.*, vol. 34, no. 2, 213–233, Apr. 2008. doi:10.1080/01411920701492050. [Accessed: Aug. 7, 2018].

[23] T. Lumley, “Assessment criteria in a large-scale writing test: What do they really mean to the raters?” *Lang. Test.*, vol. 19, no. 3, pp. 246–276, Jul. 2002. doi: 10.1191/0265532202lt230oa. [Accessed: Jun. 22, 2018].

[24] V. Di Pasquale, S. Miranda, R. Iannone, and S. Riemma, “An overview of human reliability analysis techniques in manufacturing operations,” in *Operations Management*, M. Schiraldi, Eds., pp. 221–240, Mar. 2013. doi:10.5772/55065. [Accessed: Jun. 28, 2018].

[25] E. Hollnagel, *FRAM: the Functional Resonance Analysis Method: Modeling complex socio-technical systems*. Boca Raton, FL, USA: CRC Press, 2012.

[26] Diefes-Dux, H. A., & Ebrahiminejad, H. (2018). Standards-based grading derived data to monitor grading and student learning. *Proceedings of the 125th ASEE Annual Conference and Exposition, Salt Lake City, UT.*

[27] M. T. Boren, and J. Ramey, "Thinking aloud: Reconciling theory and practice," *IEEE Trans. Prof. Comm.*, vol. 43, no. 3, pp. 261–278, Sept. 2000. doi:10.1109/47.867942. [Accessed: Jul. 12, 2017].

[28] W. M. I. Suto, & R. Nadas, "Why are some GSCE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features," *Res. Pap. Educ.*, vol. 24, no. 3, pp. 335–377, Sept. 2009. doi: 10.1080/02671520801945925. [Accessed: Jul. 27, 2018].

Appendix

This appendix includes tables for functions associated with the actual grading process (the foreground functions) as well as the functions associated with production of the documents involved in the grading process (i.e., the background functions for the production of the assignment, the rubric, and the response). At the time of submission, these functions may still be subject to further analysis and potential refinement.

Table A.1. Cognitive functions (foreground) identified in the grading process based on observation and [22] and [23]

Cognitive Functions	Description
Orienting	Tasks to orient the grader regarding the task, expected performances, or specified portions of a response.
Questioning	Expressing confusion regarding one of the documents or part of one of the documents.
Translating	Stating element of rubric in simpler language to assist interpretation.
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.
Shifting	Switching attention from one document (i.e., problem set, solution, rubric, sample response) to another.
Error spotting	Finding an unexpected part of a response.
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 criterion 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.
Documenting	Making an actual physical annotation of criterion achievement or making a mental note.

Table A.2. Assignment design (background) functions based on [21], [27], and additional analysis

Assignment Functions	Description
Requiring context	Designing extent a problem will require understanding of context in order to produce an adequate response.
Distancing context	Designing the familiarity and concreteness of the problem's context.
Writing directions	Designing the level of detail included in the problem directions.
Designing task complexity	Designing the task's complexity (i.e., how many steps or how much advanced planning is needed for an adequate response).
Providing scaffolds	Designing the way a problem is broken into sub-tasks or the provision of extra guidance or hints.
Expecting length	Expecting responses to be an approximate length.
Expecting openness	Expecting a range of acceptable answers.
Expecting task dependence	Designing the dependence of separate tasks within a problem.
Expecting interpretability	Designing the extent to which students will likely need to provide explanation to interpret their responses.
Expecting depth of knowledge	Designing tasks that will require a certain level of knowledge to successfully complete.
Expecting diagrams	Expecting responses to include visuals, diagrams, tables, or charts.
Aligning with instruction	Designing the extent to which problems align with instructional materials.

Table A.3. Response (background) functions based on [21] and additional analysis

Response Functions	Description
Employing expected approach	Student employing the approach (or one of the approaches) expected by the assignment and rubric designers.
Communicating responses clearly	Student communicating ideas or approach to the problem in a way that is reasonably understandable or interpretable for a grader.
Writing legibly	Student providing a response that can be read.
Meeting requirements	Student producing a response that meets all the requirements specifically stated in the assignment.

Table A.4. Rubric (background) functions based on [27] and additional analysis

Rubric Functions	Description
Defining range of acceptable responses	Addressing within the rubric how to handle different possible student responses to the task.
Indicating what to grade	Indicating within the rubric the specific aspect of a response that is to be graded for a particular rubric item.
Communicating criteria clearly	Selecting the language for communicating the criterion to be evaluated.
Communicating criteria concisely	Writing the criterion in concise language.
Defining quantity graded in a criterion	Designing how much of a response (or, how many parts of a response) are graded simultaneously by a single criterion.
Grouping criteria	Deciding how many criteria constitute a single rubric item (i.e., learning objective).
Evaluating excerpt of response	Deciding to focus on a specified portion of an expected response for multiple criteria.
Repeating performance tasks	Deciding to grade a criterion multiple times across an assignment.
Including dependencies	Making criteria interrelated such that success on one criterion is dependent upon or directly tied to success on one or more other criteria.
Designing evaluation difficulty	Designing the level of difficulty involved in evaluating a criterion (i.e., having a specific number-valued response versus deciding a paragraph clearly communicates an idea).
Aligning with problems	Matching the expectations communicated within the rubric to the constraints placed on the students as communicated through the assignment.