2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION
Paper ID #50608

Toward Scalable Assessment of Undergraduate Reflective Practice: Comparing Multiple Reflection Quality Codebook Validation Approaches

Dr. Margaret Webb, Cornell University

Dr. Margaret (Maggie) Webb is postdoctoral associate at Cornell University. She holds a B.S. in Mechanical Engineering (Rice University, a M.S. in Civil Engineering, and a PhD in Engineering Education (Virginia Tech). Her research interests include STEM graduate and postdoctoral education, andragogy, understanding how academic systems influence researcher professional development, motivation, and agency.

Campbell James McColley, Cornell University

Dr. Campbell McColley is a NSF Postdoctoral STEM Ed Research Fellow in the Department of Biomedical Engineering (BME) at Cornell University in the Biomedical Engineering Education Assessment and Research (BEEAR) Group. He received his Ph.D. in Environmental Engineering from Oregon State University, where he investigated microplastics transformations and behavior in aquatic environments. His work focuses on developing and supporting college-industry partnerships in engineering curricula with the wastewater treatment sector.

Mohammed A. Alrizqi, Cornell University
Gabriel Azure Antonio Mendez-Sanders, Cornell University

Gabriel Mendez-Sanders is a first-year PhD student in Cornell University's Smith School of Chemical and Biomolecular Engineering. His research focuses on renewable energy education and how practicing engineers serve as educators in the energy industry.

Alexandra Werth, Cornell University

Alexandra Werth is an assistant professor at the Meinig School of Biomedical Engineering, specializing in Engineering Education Research (EER). She focuses on developing evidence-based teaching methodologies to foster authentic learning environments. Dr. Werth holds a Ph.D. in Electrical and Computer Engineering from Princeton University, where she developed a non-invasive mid-infrared glucose sensor. She later conducted postdoctoral research in physics education at the University of Colorado Boulder, where she helped develop the first large-enrollment introductory physics course-based research experience (CURE).

Toward Scalable Assessment of Undergraduate Reflective Practice: Comparing Multiple Reflection Quality Codebook Validation Approaches

Abstract

This methods-focused empirical research full paper advances a theory-driven approach to qualitative codebook development designed for Generative AI (GenAI)-enabled analysis at scale. We combine qualitative interrater agreement practices with multiple quantitative reliability calculations, and we compare the constraints and affordances of each across two iterations of a codebook for assessing undergraduate STEM student reflection quality. Findings suggest triangulating multiple interrater calculations, calculating them for multiple codebook versions across multiple student samples, and adapting validation processes based on code types support the scaling of reflection quality assessment with the enablement of GenAI. Our methods push the boundaries of validation and reliable application of qualitative codebooks in EER towards a reality where we can adapt codebooks for reflection quality to specific course contexts and automate their application. We argue theory-driven automation-ready codebooks demand validation processes that ensure both human and AI coders can reliably identify the same theoretically grounded constructs. This standard requires (i) integrating multiple quantitative coefficients, to address different statistical assumptions, (ii) with qualitative interrater agreement approaches, to ensure shared construct understanding and dataset limitations, and (iii) considerations for latent versus semantic meaning in data. Without this, automated reflection quality assessment risks scaling error rather than insight.

I. Introduction

Generative AI (GenAI) is making it possible to analyze qualitative data at unprecedented scale [1], [2], [3], [4], but it is also making weak codebooks dangerous, because the codebook functions as a measurement instrument that shapes what counts as evidence and what conclusions are defensible. In much of Engineering Education Research (EER), codebooks tend to be treated as locally useful artifacts; often justified through face validity and applied in bounded contexts [5]. That approach becomes a major vulnerability in GenAI-enabled qualitative research, where codebooks are increasingly used to support automated qualitative coding enabling large-scale inference.

In response, we critically examine what it takes to develop *automation-ready qualitative codebooks*, defined as codebooks intended not only for human interpretation but for scalable, GenAI-assisted application. Specifically, we analyze how different interrater agreement and reliability approaches, each with distinct assumptions and sensitivities, shape the development trajectory of a codebook and claims that can be made about its defensibility [6], [7], [8]. We focus on codebooks as measurement instruments and on validation evidence needed before scaling, therefore a full treatment of agent design and prompting is reserved for future work; we do however include examples showing how elements of the human-facing codebook were translated into instructions for a GenAI coding system.

We situate this methodological contribution in a larger study of reflection quality, using reflective writing as a deliberately challenging test case. Reflection quality is a historically difficult, latent construct to operationalize and assess with both human raters and automated NLP techniques due to inconsistent definitions, contextual variability, and the subjectivity of students' meaning making [9], [10]. Building on our prior work, we use a reflection quality codebook we previously developed that nests semantic ("situatedness") and latent ("abstraction") codes. This design makes it possible to probe how different codebook development approaches handle interpretive (latent) versus surface-level (semantic) coding, and what kinds of weaknesses each approach reveals, directly answering our research questions:

> **RQ1:** How do different interrater reliability approaches and assumptions shape conclusions about developing qualitative codebook quality, particularly across semantic versus latent codes?
> **RQ2:** What combination of validation approaches provides evidence for refining a codebook prior to scaling GenAI-assisted coding, especially for small, exploratory development datasets?

We make three methodological claims based on findings: (1) common agreement and reliability metrics embed assumptions about code prevalence, rater behavior, and error structure that can materially change

conclusions about a developing codebook; (2) latent coding introduces distinct challenges for both human and GenAI raters, and effective refinements depend on rater positionality (e.g., expertise and local, contextual knowledge); and (3) especially when smaller, exploratory datasets are used for codebook development, mixed-method agreement analyses are necessary to build defensible evidence before scaling automated or semi-automated coding. To support these claims, we combine qualitative interrater agreement practices (e.g., negotiated coding rationales and boundary-case analysis) with multiple quantitative reliability calculations [6], [7], [8], and we compare the constraints and affordances of each approach across the two codebook iterations. We argue that, because reflective writing varies widely by course discipline, topic, and instructional setting, validity evidence optimized for a single dataset can be misleading when the intended use is scale or automation. Our contribution is therefore methodological: we show how reliance on single interrater reliability (IRR) metrics can misrepresent codebook performance, and how triangulating across methods provides clearer guidance for refinement prior to GenAI-assisted deployment. To support EER scholars in selecting appropriate IRR metrics, we also share a decision-tree overviewing which metrics are appropriate based on codebook and dataset characteristics.

II. Background and Theory

A. Qualitative quality and codebook-based analysis in EER

Qualitative research in education research has long emphasized the need for transparent, well-justified analytical decisions and for credibility in interpretation. Frameworks for quality, including Lincoln and Guba's [11] trustworthiness criteria (e.g., credibility, dependability, confirmability, and transferability) and subsequent EER-specific guidance (e.g., Walther et al. [12], [13]; Sochacka et al. [14]), emphasize quality is established through documented procedures, reflexivity, and defensible links between data, interpretations, and claims. At the same time, qualitative traditions differ in the role they assign to coding reliability. For example, Braun et al.'s [15] reflexive thematic analysis explicitly cautions against treating interrater reliability (IRR) as a universal marker of quality, arguing instead for coherent interpretive logic developed through iteration as well as transparency via documentation.

In contrast, codebook-based approaches (especially those intended for reuse, comparison across contexts, or scaled application) often do require evidence that independent coders can apply the scheme in a consistent and interpretable way [7]. Recent work highlights inconsistencies of how EER reports and justifies codebook agreement. Martin et al.'s [5] systematic review found gaps in research articles' descriptions of reliability metrics and disagreement resolution processes, highlighting the need for clearly explained agreement using specific frameworks and documentation of analytical decisions used in resolutions . They also noted the importance of mixed method approaches to establishing agreement, as the "interpretive process of qualitative analysis" [5, p. 10] cannot be comprehensively described by quantitative measures like agreement coefficients, or perhaps worse, face validity assumptions alone. Additionally, single-metric or method validation can mislead scholars about codebook quality [5], [7].

A second (and often underappreciated) issue is that agreement and reliability statistics embed assumptions about code prevalence, rater behavior, and error structure. Gwet [7] and related methodological work show commonly used coefficients can behave counterintuitively under conditions typical of developing codebooks (e.g., rare codes, skewed distributions, evolving code definitions). In a single-dataset study, rare codes may never appear in sufficient quantity to evaluate systematically, and researchers often can justify interpretations of those few instances through reflexive team discussions and transparent memoing. However, when a codebook is intended for scaled GenAI use, operating conditions can quickly change: the same codes may be common in one population and rare in another and shifting base rates can affect both agreement statistics and the practical consequences of misclassification. As a result of these shifting base rates, relying on a single metrics alone like Cohen's Kappa by itself, without corroboration with additional metrics like Krippendorf's Alpha, can lead researchers to overestimate or underestimate codebook quality [7]. Multiple varied IRR metrics' have varied statistical assumptions and meanings that, when combined, can account for errors in single metrics alone due to use on unbalanced datasets, for example (discussed in more detail later). These concerns motivated our focus on triangulating agreement

evidence with multiple complementary approaches, particularly important for codebooks mixing semantic (surface) and latent codes (interpretive).

B. GenAI for qualitative research: automation changes evidence required for quality

EER scholars have recently emphasized that GenAI may streamline parts of qualitative workflows, but only if researchers maintain human oversight, ensure privacy and data control, and make analytic assumptions transparent [2], [5]. We align with this call to uphold established qualitative standards, but we also argue that automation increases the standard of evidence required. When qualitative coding is scaled through GenAI-assisted protocols, the claims supported by the codebook typically extend beyond a single local analysis: codes may be applied to larger datasets, compared across settings, or used to generate aggregated inferences more quickly than traditional qualitative workflows allow. Under these conditions, "standard" reporting of agreement processes is necessary but may not be sufficient.

Efforts to scale qualitative analysis have a long history in natural language processing (NLP), including bag-of-words and dictionary methods, supervised classification, and topic modeling. These approaches can be powerful, but they struggle with context dependence and latent meaning [3]. More recent transformer-based language models have expanded what is technically feasible, including zero-/few-shot classification and instruction-following systems that can apply codes across large datasets [16], [17], [18]. These developments are able to leverage greater context and integrate into existing qualitative research workflows [19], [20], and they have prompted growing experimentation in education research and EER with AI-supported qualitative coding, summarization, codebook development, etc. [1], [2], [21]. At the same time, model outputs remain sensitive to prompt framing, category ambiguity, and hidden assumptions in the coding scheme itself [22], [23], [24], [25] making the codebook comparisons a central determinant of what automated systems can do reliably [4]. For example, in their 2025 study, Fuchs et al. [23] found adding synthetic data improved NLP classification accuracy from 68% to 81%, but that models produced transcripts with extraneous information and failed to capture discourse as instructor-dominated (an analysis which would require uncovering assumptions in the data).

C. Reflection quality as a motivating case and testbed

We situate this methodological contribution in our larger study of reflection quality in engineering education. Existing theory for students' development of reflective writing and thinking provide assessment scales for reflection quality in general education contexts, but they tend to focus solely on the content *or* the process of reflection [26], [27], [28], [29], [30], [31], neglecting the ways in which they interact as reflective practices deepen [27], [32]. Models are typically implemented through custom rubrics in single courses, with modest transferability to specific STEM disciplines' technical, context-bound tasks. To connect content to process, our codebook is nested for *situatedness* and *abstraction* elements of reflection quality, grounded in Ryan & Ryan's 4R model [32]. Situatedness is "the content of student reflection responses that may include various elements related to their personal or academic experiences, [and] their broader discipline or field" and abstraction is "the process by which students distill and generalize experiences and knowledge in reflections" [33, p. 1]. In both codebooks (V2 and V3), situatedness codes included course contextual factors expected to surface in reflections on learning (e.g., activities A, learning objectives L, academic disciplines D, personal connections P, connections with others O, group experiences G). Providing the nested nature of the codebook, these codes were each stratified with sub-codes spanning responding R1, relating R2, reasoning R3, and reconstruction R4 levels of abstraction. See Appendices A and B for the codebook versions at the center of this study.

To illustrate the value of our nested codebook, consider how reflection assessment involves not just analyzing "what" is being discussed, but also "how" the "what" is mentioned [28], [32]. We want to know if students are reflecting on content situated to their specific course (i.e., learning activities, outcomes, interactions with others) or not, but we also want to know if they are displaying a change in thinking, or abstracting that content. For example, there are differences between students changing their thinking, indicating reconstruction (R4), reasoning (R3) about a given situation, and simply repeating back a response to specific course content based on a given question (R1).

Coding not just content (via situatedness codes) but also abstraction level of that content is desirable for researchers and educators alike because of the interplay between these elements of reflection [9], [33]. For instance, as biomedical engineering (BME) students develop skills to reflect on course-specific content, they develop *situated* reflective practice, developing not just as general humans in general education, but as biomedical engineers in BME. And, as they develop that situated practice, their abilities to abstract within it co-develop [34]. We therefore argue that reflection quality operationalized in our codebook is an important case for studying qualitative coding automation with GenAI because it has been historically difficult to do, and our tiered codebook challenges IRR for human researchers and AI agents with combined content (semantic) and abstract (latent) codes [15].

Drawing on Braun & Clarke's [35] book on thematic analysis, qualitative coding in general can be understood along a continuum from semantic to latent meaning. Semantic codes capture ideas that are explicitly stated in participants' data, even if not in participants' exact words, while latent codes require coders to interpret meaning concealed or abstracted from what the data immediately presents. This is not a binary distinction, and neither type is inherently more sophisticated; latent coding simply tends to be less immediately evident and requires coders to move further from the data toward theoretical inference. Importantly, latent coding does not imply unconscious meaning – students can and do demonstrate metacognition explicitly – but it does require interpretive work to surface meaning not fully recoverable from text alone. In this study, these two types are operationalized through the nested structure of our codebook: situatedness codes (A, L, D, P, O, G) function primarily as semantic codes, capturing content explicitly referenced in student reflections that can often be identified through relatively direct reading of the text, while abstraction codes (R1–R4) are latent, requiring coders to interpret not just what a student mentions but how that content is engaged. This distinction is made for GenAI-supported analysis, as current large language models are more sensitive to surface cues than to the kind of interpretive, theory-grounded judgment latent coding of student reflection data demands [22], [23], [24], [25].

III. Methods

To answer our research question, we present our GenAI-enabled codebook development process from one version of our codebook to the next, first describing two human researchers' effort coding 20 student reflection sessions (92 responses in all) with Version 2 (V2) of our codebook (see Appendix A). Then, we present our qualitative interrater agreement validation along with varying quantitative IRR analyses of that coding. Qualitative validation aimed toward representativeness, parsimony, mutual exclusivity, clarity of code definitions, and addressing instances of potential coder bias [6], [8]. Quantitative measures included simple percent agreement Cohen's Kappa, Krippendorf's Alpha, and Gwet's AC1 [7]. Results from these varied agreement analyses on V2 coding informed the development of codebook Version 3 (V3) (see Appendix B), which our two human researchers used to code an additional 10 reflection sessions (44 responses in all). We share the results of these analyses and discuss how the merits and limitations of varying codebook validation efforts separately (and together) informed codebook versioning. We also employed multi-agent GenAI-enabled coding, enabled by GPT-4.1 and based on system prompt and multi-agent flow design recommendations from OpenAI, of the same excerpts for both codebooks (V2 and V3), though they are not the focus of this paper. Details on quantitative IRR methods and justifications are in Appendix C.

A. Research Participants and Data Collection

This study sampled from a broader dataset of over 3,000 reflection assignments tied to in-class activities in undergraduate STEM courses across multiple semesters. We focused on a set of 20 and 10 reflection sessions (with 92 and 44 individual responses each) using both V2 and V3 codebooks respectively. Student participants were enrolled in a Physics lab (Fall 2024), Biomedical Engineering seminar (Spring 2024) or a summer research experience (Summer 2025), and were required to respond to administered reflection questions associated with participation in course activities such as labs, discussions, exams, and quizzes. Sometimes assignments featured a single reflection question, and other times sessions involved multiple.

At the start of a sampled course, students were given the option to provide IRB-approved (#0148748) informed consent to allow us to use their responses for research.

The 92 responses used for V2 coding were drawn from the Physics lab and Biomedical Engineering seminar, while the 44 responses used for V3 coding were drawn from both these classes and a summer research experience, providing cross-context variability across codebook iterations. Reflection prompts varied in specificity and focus across sessions — some targeting technical course content (e.g., lab procedures, exam review) such as *"Describe how your team dynamics influenced the outcomes of your desert island activity this week compared to last week,"* others targeting broader professional or personal development, like *"Describe how one's own biases may impact engineering design in biomedical engineering."* This meant our dataset captured a range of situatedness and abstraction patterns likely to surface in practice. Individual responses ranged in length and depth, from a single sentence to several paragraphs.

These dataset characteristics influenced coding for both human and GenAI raters: human raters' direct familiarity with courses, instructors, and activities students reflect on provided contextual grounding that supported human interpretation of more ambiguous responses. For instance, one BME seminar activity had students role-play as clinical team members treating a simulated patient, an orange named "Leah." When a student wrote, *"the information I had on Leah eating the bad shrimp was important in determining a treatment plan that would cover all the bases and get her healthy again... the kindergarten teacher's skills in first aid also helped with the balance of things to do to treat Leah,"* our research team immediately recognized "Leah" was an orange used as a prop in a structured class activity, and we could interpret the reflection accordingly. However, a GenAI rater encountering the same excerpt has no context, which could distort coding beyond what the student is describing to the degree of abstraction the student is actually demonstrating. As such, for the GenAI rater, this variability across contexts and prompt types challenged not just identification of surface-level or semantic content in student reflections, but also the more latent, inferential judgments about abstraction level.

B. Limitations

Importantly, the larger corpus (3000+ student reflections) associated with our study stemmed from collection in only three classes, though these courses did vary widely in both content and instructional format and reflections were collected throughout multiple semesters across a variety of activities. Only smaller samples from this corpus (136 reflections) were used for this portion of codebook validation and development, though previous versions of our codebook had been used on 400+ reflections, limiting our study's external validity in important ways. First, two coders from the same institutional context may share implicit assumptions about student reflection that would not hold across research teams, disciplines, or institutions, meaning our interrater findings may underestimate the disagreement that would emerge with greater rater diversity. Second, the specific patterns we observed, and recommendations made based on them, may be characteristic of our particular dataset's distribution rather than reflective of broader trends. In future studies we plan to extend validation beyond currently sampled courses and to researchers outside of our group to continue expanding the representativeness of our dataset.

C. Positionality Statement

Three authors were involved in all data collection and analysis, and all team members are familiar with the institutions, courses, and activities reflection data stem from. Our contextual understanding was critical to supporting development of codes expected to surface in data as well as the interpretation of results.

IV. Findings and Discussion

GenAI is changing the scale and speed at which researchers can apply qualitative codebooks, but the field still lacks clear guidance on what constitutes defensible agreement evidence when a codebook is intended for semi-automated or automated use. Prior work has emphasized transparent reporting and limitations of single-metric approaches [2], [5], [7], yet a practical question remains: *how do different agreement methods shape conclusions researchers draw while a codebook is being developed?* We address this by integrating quantitative agreement indices, qualitative interrater debriefing (rationales, boundary cases, disagreement

resolution), and comparison across codebook versions. We present results and discussion as claims to synthesize this multi-method record into transferable insight, highlighting assumptions each method brings, failure modes revealed or obscured, and specific code refinement decisions it supports.

Claim #1: Agreement metrics are not neutral; they have embedded assumptions. Under skew and rare codes, metrics can yield sharply different conclusions about the same codebook.

Table C.0 in Appendix C summarizes agreement evidence for V2 vs. V3 codebooks and two human raters' coding. Table 1 below highlights a few results for discussion. Across most codes, V3 showed improvement relative to V2, consistent with our intentions and goals for codebook improvement. We present detailed calculations in Appendix C.

Table 1. Individual Code Reliability Metrics: V2 vs V3 Codebook Comparison. We report out Cohen's Kappa (κ) for V2 and V3 codebooks, the change in Cohen's Kappa from V2 to V3 (Δκ), the percent agreement (Agr%), Krippendorf's Alpha (α).

Code	V2 κ	V3 κ	Δκ	V2 Agr%	V3 Agr%	V3 α
O.R1	0.224	0.550	+0.326	73.9	90.7	0.311
O.R2	0.000	—	—	91.3	100.0	—
D.R1	0.383	0.239	−0.144	91.3	81.4	0.205
D.R2	0.000	—	—	95.7	100.0	—

Note: Kappa (κ) measures interrater agreement beyond chance; Δκ is the change in Kappa from V2 to V3 with positive values indicating improvement; percent agreement (Agr%) captures the proportion of excerpts coded identically by raters without accounting for chance; and Alpha (α) assesses agreement while accommodating unbalanced code distributions.

Between V2 and V3, we saw improvements in metrics for almost all codes (except L.R2), suggesting changes made between versions helped improve reliability. That said, the appropriateness of each metric and changes made to codebooks based on them differed depending on the nature of the codes and data used for validation. For example, since simple percent agreement calculations do not account for percent chance, Cohen's Kappa is considered more accurate; however, Kappa suffers from are commonly termed paradoxes of Kappa [7]. The first paradox is if percent chance agreement is large, Kappa calculations can convert a relatively high value percent agreement to a relatively low Kappa; the second is that unbalanced totals (or marginal application of codes) produce higher Kappa values than balanced totals [7].

Our dataset, though relatively diverse across different student populations, exhibited both paradoxes. For instance, during our interrater discussions, we found one rater was consistently applying R4 more often than the other (confirmed by IRR data), and this was not the only source of imbalance. Some codes, like O.R2 and D.R2, showed high percent agreement (91.3%, 95.7%) with kappa values near zero, affecting interpretability of results because they were used both sparsely and by one rater more than the other. It was important to calculate both metrics across quant and qual, as one interpretation alone could have misled us or given no direction forward. Because we computed multiple coefficients understanding their assumptions [7] and because we also simultaneously qualitatively analyzed codes (as per [5]), we realized O and D codes were experiencing a complex error due to both sparse usage and unclear definitions, known limitations of Kappa [7], [10], [36].

Upon inspection of the O codes and specifically places where one rater coded an excerpt as P and the other O, excerpts tended to actually emphasize students' own experiences of observing teammate's actions (personal connections) or working with others in action (something other than personal connections or connections with others) rather than describing what a teammate did (connections with others). In these cases, students often were writing using first-person plural language to describe collective experience, rather than attributing actions or perspectives to a teammate. A physics student reflecting on a lab wrote:

"None of my group members this week [were] Physics major[s]. This made the lab a little more challenging because we didn't fully understand some of the principles behind this experiment. We struggled a lot to start … because we weren't sure what to set as our research question."

A different student noted, *"We did a lot of our discussion and decisions as a group in person, then finalized the coding and the writing of the report throughout the week remotely."* Across these and similar excerpts, the recurring issue was that the V2 codebook lacked a category for collective or shared group experience, or the "we" voice showing up in the data that was neither personal nor attributable to other individuals. Therefore, during qualitative inter agreement assessing representativeness, parsimony, mutual exclusivity, and clarity [6], [8], we determined that group-oriented reflections represented a distinct construct worthy of a separate code when we identified notable overlaps between "Connection with Others" (O codes) with "Personal Connection" (P codes), particularly when students used first-person plural language ("we") to describe group experiences. The new situatedness code in V3: "Group Experience" (G.R1–G.R4) captured shared team perspectives, collective decisions, and joint challenges conceptually distinct from either personal connections in "we" observations or observations about teammates' behaviors in connection with others. We also clarified codebook definitions so that P codes focused on students' own lived experiences and perspectives, while O codes focused on others' actions, thoughts, or contributions explicitly.

These findings emphasize the importance of triangulating multiple interrater agreement analyses and of reflection on their assumptions and the balance of code application in relevant test datasets. If the assumptions and common issues with Kappa were unknown to us, and if we were unfamiliar with the possibility our dataset might not have a balanced code distribution, we would not have been able make improvements in the same strategic way (i.e., calculating Alpha alongside Kappa for handling unbalanced data). Claim 1 therefore suggests human interpretation is still particularly necessary for establishing construct validity, the extent to which codes capture the concepts investigators intend to assess [37], *and* content validity, the extent to which codes represent the entire domain intended [38], of codebooks and associated datasets for automation purposes. To support EER researchers selecting appropriate metrics, the decision-tree below overviews appropriate metrics based on codebook and dataset characteristics.

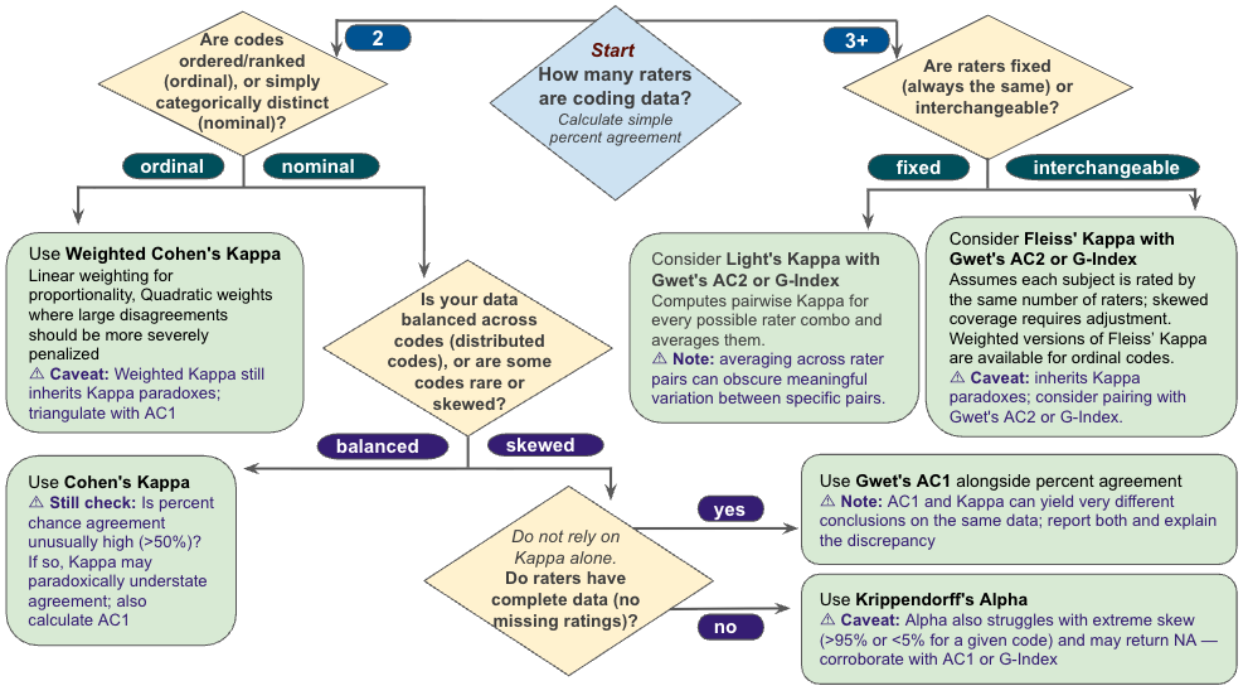

Figure 1: Decision-Tree for Common IRR Metrics in EER Based on Codebook and Dataset Characteristics

Claim #2: Latent coding (vs semantic coding) is challenging for both humans and GenAI agents; improvements to address these challenges differ based on raters positionality.

Tables 2 and 3 show how humans and GenAI raters were both generally more likely to agree on situatedness codes (semantic, content-level meaning) over more complex abstraction ones (latent, conceptual meaning). This finding confirms arguments for complexity based on code type [15] and suggests the automation of latent codes requires more human oversight and interpretation than semantic counterparts.

Table 2: Percent Agreement Summary: V3 Codebook Situatedness for 2 and 3 Raters

Situatedness Group	Avg % Agreement (2-Rater)	Avg % Agreement (3-Rater)
A Assignment	87.21%	72.68%
D Discipline	91.28%	88.95%
G Group Experience	92.44%	79.07%
L Learning Outcome	83.72%	70.35%
O Connection with Others	97.68%	89.71%
P Personal Connection	85.47%	73.86%

Table 3: Percent Agreement Summary: V3 Codebook Abstraction Level for 2 and 3 Raters

Abstraction Level	Avg % Agreement (2-Rater)	Avg % Agreement (3-Rater)
R1 Responding	79.77%	55.05%
R2 Relating	98.39%	96.28%
R3 Reasoning	83.37%	69.86%
R4 Reconstructing	97.18%	94.64%

To illustrate, during peer debriefing discussions, the abstraction codes, with higher percent agreement overall, were simply augmented with the addition of statements to contrast codes from one another (i.e., disconfirming evidence or exclusion criteria) and the provision of example keywords. In contrast, for latent codes, addressing disagreement (and therefore construct validity [37]) was not as simple as adding a new code category (e.g., G codes mentioned in Claim 1), keywords, or contrasting phrases to distinguish codes. Instead, we had to provide full example excerpts and justification statements for the application of abstraction constructs to our data to improve performance. Keywords were still provided, but example excerpts and justification statements established through convergent interpretation were also necessary. To see these codebook changes in more detail, see Appendix A (V2 codebook) and B (V3 codebook). These Claim 2 findings corroborate previous research suggesting semantic coding augmented by GenAI in higher education is currently more automatable and scalable than latent coding [3], but also reveal approaches for addressing challenges with automating latent codes worth future consideration.

Claim #3: Multiple and mixed-method agreement analyses are key to reliable scaled deployment, especially when smaller, exploratory datasets are used for codebook development.

Table 4 overviews the results of calculating unweighted, linear, and quadratic weighted versions of Cohen's Kappa results for the V3 codebook's abstraction codes. They provide insight into the benefits of applying multiple weighted calculations of Kappa for certain code types (i.e., ordinal rather than nominal) [7] in

combination with qualitative (and distinctly human) understanding. Details on assumptions of and calculations for these Kappas are in Appendix C. *Note:* Claim 3 differs from Claim 1 primarily for its focus on multimeric interrater analyses and the nature of datasets used for codebook development; Claim 1 centers on the embedded assumptions of interrater reliability calculations and their implications, and Claim 3 is about how performing multiple calculations helps the interpretation of embedded assumptions.

Table 4: Weighted Cohen's Kappa Unweighted, Linear, and Quadratic Weighted Abstraction Code Group Reliability: V3 Codebook Only

Level	Kappa, Unweighted	Kappa, Linear	Kappa, Quadratic
R1 Responding	0.448	0.372	0.31
R2 Relating	0.587	0.585	0.566
R3 Reasoning	0.375	0.307	0.308
R4 Reconstructing	0.541	0.463	0.352

For abstraction codes (R1-R4), we considered there to be more agreement when raters categorized excerpts as "reasoning R3" and "reconstructing R4" than when we categorized that same excerpt as "responding R1" and "reconstructing R4," making these codes ordinal and dependent (in contrast to nominal and independent situatedness codes that are not ordered). This dependency is part of the latent nature of abstraction, which is also related to these codes abductive derivation based on theory and connections between situatedness (content) and abstraction (process) in reflection [33]. Because of the nature of these codes, the weighted Kappa calculations shown above allowed us to examine whether raters confused theoretically adjacent versus distant abstraction levels and revealed key limitations of our dataset for validating our codebook for further automatability.

As shown, for R1 (responding/reporting), we found weighted Kappas were consistently lower than unweighted (unweighted K = 0.448 vs quadratic K = 0.31), with similar patterns for R3 (reasoning) and R4 (reconstruction). This pattern suggests human and AI raters were disagreeing across distant categories (R1 vs R3/R4) less than adjacent ones, struggling to distinguish abstraction levels in a way that runs contrary to expected theory [27], [28], [32]. We would have expected there to be more disagreement across distant categories (i.e., weighted kappa values would be above unweighted), as Ryan & Ryan's framework suggests R1-R4 are ordered 1, 2, 3, 4, where relating R2 is "closer" to responding R1 than R3 reasoning is; however, the weighted Kappa's that adopt this assumption are coming out lower than unweighted calculations, suggesting either these constructs are either not ordered at all or they need to be reordered. This perceived incongruence suggests potential limitations of either the theory, our codebook, or our dataset (which would require another empirical study to parse). Comparing our data to Ryan & Ryan's [32] framework for student reflection, this finding suggests R2 (relating) and R3 may be flipped for the STEM students sampled, raising theoretical questions about whether STEM learning environments surveyed emphasized responding and reasoning over relating or reconstructing. These additional questions would not have been revealed without comparing multiple versions of weighted and unweighted Kappa, so it points to the complexity of validating latent abstraction codes due to nuance and *contextual* evidence required. This kind of conclusion relies on local knowledge of datasets and study sites, our experiences in STEM learning environments, as well as our understanding of the strengths and limitations of Kappa. Further investigation of this phenomenon deserves additional data collection and research attention outside the scope of this paper. See Appendix C for more details on the weighting schemes used in Kappa calculations. Linear weighting was selected because Ryan & Ryan's 4R framework theoretically proposes equal-interval progression across R1-R4 levels, making proportional distance penalties conceptually appropriate; quadratic weighting was added too, applying heavier penalties for distant disagreements (e.g., R1 vs. R4) than adjacent ones so we could probe further.

Key dataset limitations future researchers must consider include addressing both the content validity, or representativeness [38], of their dataset as well as its construct validity, or how much defined codes capture their intended concepts [37]. For exploratory research where theory may be lacking or need adjustment, or where diversity in student responses across specific contexts is expected to vary (such as student reflection quality specified to individual course and disciplinary contexts), large datasets with diverse population sampling are necessary, though typically not a feature of qualitative research in EER that is not purposively seeking scale.

In all, because code types, interrater agreement assumptions, and the interpretation of their results inherently impact validation conclusions for qualitative codebooks enabled by GenAI, we argue human interpretation is still key to scaling and deployment. EER researchers must demonstrate understanding and alignment of IRR measures and code types and address issues of skew or unbalanced data. Ideally, careful consideration for both codes and datasets associated with them can result in exploratory theoretical contributions (e.g., STEM environments emphasizing responding and reasoning over relating reconstruction as per Claim 3), like smaller scale qualitative analysis, but with scalability impossible without GenAI.

V. Conclusions

Overall, findings remind us that qualitative validity of codebooks in EER should not stop at high interrater agreement coefficients or rounds of coding performed alone; finding also provide validation design principles for GenAI-enabled qualitative research going forward. EER researchers need proficiency with both latent and semantic code types [15] for reliable GenAI automation of reflection quality assessment. Latent codes require interpretive convergence via consensus-based validation triangulated with multiple quantitative measures, and they are difficult for both humans and AI to code reliably. Mixing multiple validity and reliability methods adds rigor and insights undiscoverable otherwise.

The automation of qualitative coding should therefore be preceded by both qualitative and triangulated quantitative validations of codebooks, and human interpretation is particularly necessary for establishing content validity and external validity [37] of codebooks and associated datasets. Human researchers distinctly familiar with data collection sites support validity in ways GenAI performing internal analysis cannot. By understanding the limitations and transfer potential of data upon which codebooks are developed, human researchers provide intervention necessary for addressing confounding contextual factors and interpreting them while also reporting and justifying codebook validation measures.

VI. Acknowledgements

This material is based on work supported by the National Science Foundation under Grant No. 2430496. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] A. Johri, A. S. Katz, J. Qadir, and A. Hingle, "Generative artificial intelligence and engineering education," *J. Eng. Educ.*, vol. 112, no. 3, pp. 572–577, 2023, doi: 10.1002/jee.20537.

[2] A. J. Magana, R. Watkins, and C. Vieira, "Recommendations for the integration of generative artificial intelligence in support of engineering education research workflows," *J. Eng. Educ.*, vol. 115, no. 1, p. e70043, 2026, doi: 10.1002/jee.70043.

[3] M. Perkins and J. Roe, "Generative AI Tools in Academic Research: Applications and Implications fo Qualitative and Quantitative Research Methodologies," in *Handbook of Artificial Intelligence in Higher Education*, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2408.06872

[4] H. Schroeder, M. A. L. Quéré, C. Randazzo, D. Mimno, and S. Schoenebeck, "Large Language Models in Qualitative Research: Uses, Tensions, and Intentions," Feb. 27, 2025, *arXiv*: arXiv:2410.07362. doi: 10.48550/arXiv.2410.07362.

[5] D. Martin, M. de Andrade, S. Hunt, V. Ramachandran, and L. Dokter, "Reporting Reliability in Qualitative Engineering Education Research: A Literature Review," in *53rd Annual Conference of the European Society for Engineering Education (SEFI)*, Tampere University in Tampere, Finland: Zenodo, 2025. doi: 10.5281/ZENODO.17631625.

[6] C. Geisler and J. Swarts, *Coding Streams of Language: Techniques for the Systematic Coding of Text, Talk, and Other Verbal Data*. The WAC Clearinghouse; University Press of Colorado, 2019. doi: 10.37514/PRA-B.2019.0230.

[7] K. L. Gwet, *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters*, Fourth edition. Gaithersburg, Md: Advances Analytics, LLC, 2014.

[8] M. B. Miles, A. M. Huberman, and J. Saldaña, *Qualitative data analysis: a methods sourcebook*, 3rd ed. Thousand Oaks, CA: Sage, 2014. [Online]. Available: https://eric.ed.gov/?id=ED565763

[9] S. Koole *et al.*, "Factors confounding the assessment of reflection: a critical review," *BMC Med. Educ.*, vol. 11, no. 1, p. 104, Dec. 2011, doi: 10.1186/1472-6920-11-104.

[10] T. Moniz, S. Arntfield, K. Miller, L. Lingard, C. Watling, and G. Regehr, "Considerations in the use of reflective writing for student assessment: issues of reliability and validity," *Med. Educ.*, vol. 49, no. 9, pp. 901–908, Sep. 2015, doi: 10.1111/medu.12771.

[11] Y. S. Lincoln and E. G. Guba, "Naturalistic inquiry Thousand Oaks," *Calif Sage*, 1985.

[12] J. Walther, N. W. Sochacka, and N. N. Kellam, "Quality in Interpretive Engineering Education Research: Reflections on an Example Study," *J. Eng. Educ.*, vol. 102, no. 4, pp. 626–659, Oct. 2013, doi: 10.1002/jee.20029.

[13] J. Walther *et al.*, "Qualitative Research Quality: A Collaborative Inquiry Across Multiple Methodological Perspectives," *J. Eng. Educ.*, vol. 106, no. 3, pp. 398–430, Jul. 2017, doi: 10.1002/jee.20170.

[14] N. W. Sochacka, J. Walther, and A. L. Pawley, "Ethical Validation: Reframing Research Ethics in Engineering Education Research To Improve Research Quality," *J. Eng. Educ.*, vol. 107, no. 3, pp. 362–379, Jul. 2018, doi: 10.1002/jee.20222.

[15] V. Braun, V. Clarke, N. Hayfield, and G. Terry, "Thematic Analysis," in *Handbook of Research Methods in Health Social Sciences*, P. Liamputtong, Ed., Singapore: Springer, 2019, pp. 843–860. doi: 10.1007/978-981-10-5251-4_103.

[16] A. Katz, M. Gerhardt, and M. Soledad, "Using Generative Text Models to Create Qualitative Codebooks for Student Evaluations of Teaching," Mar. 18, 2024, *arXiv*: arXiv:2403.11984. doi: 10.48550/arXiv.2403.11984.

[17] A. Katz, U. Shakir, and Benjamin Chambers, "The Utility of Large Language Models and Generative AI for Education Research," *arXiv:2305.18125v1*, vol. Human-Computer Interaction, 2023, doi: https://doi.org/10.48550/arXiv.2305.18125.

[18] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can Large Language Models Transform Computational Social Science?," Apr. 12, 2023, *arXiv*: arXiv:2305.03514. Accessed: Sep. 06, 2023. [Online]. Available: http://arxiv.org/abs/2305.03514

[19] S. D. Mehta *et al.*, "Evaluation of large language models within GenAI in qualitative research," *Sci. Rep.*, vol. 15, no. 1, p. 34993, Oct. 2025, doi: 10.1038/s41598-025-18969-w.
[20] M. Nicmanis and H. Spurrier, "Getting Started with Artificial Intelligence Assisted Qualitative Analysis: An Introductory Guide to Qualitative Research Approaches with Exploratory Examples from Reflexive Content Analysis," *Int. J. Qual. Methods*, vol. 24, p. 16094069251354863, Apr. 2025, doi: 10.1177/16094069251354863.
[21] D. Reeping, C. Hampton, and D. Özkan, "Interrogating the Use of Large Language Models in Qualitative Research Using the Qualifying Qualitative Research Quality Framework," *Stud. Eng. Educ.*, vol. 6, no. 2, Jul. 2025, doi: 10.21061/see.174.
[22] B. Ahtisham, K. Vanacore, J. Lee, Z. Zhou, D. Pietrzak, and R. F. Kizilcec, "AI Annotation Orchestration: Evaluating LLM verifiers to Improve the Quality of LLM Annotations in Learning Analytics," Nov. 12, 2025, *arXiv*: arXiv:2511.09785. doi: 10.48550/arXiv.2511.09785.
[23] S. Fuchs, A. Werth, C. Méndez, and J. Butcher, "Leveraging AI-generated synthetic data to train natural language processing models for qualitative feedback analysis," *J. Eng. Educ.*, vol. 114, no. 4, p. e70033, 2025, doi: 10.1002/jee.70033.
[24] W. Holmes, F. Miao, and others, *Guidance for generative AI in education and research*. Unesco Publishing, 2023.
[25] J. Berryman and A. Ziegler, *Prompt Engineering for LLMs: The Art and Science of Building Large Language Model–Based Applications*. O'Reilly Media, Inc., 2024.
[26] Á. Alsina *et al.*, "Improving and evaluating reflective narratives: A rubric for higher education students," *Teach. Teach. Educ.*, vol. 63, pp. 148–158, 2017, doi: https://doi.org/10.1016/j.tate.2016.12.015.
[27] J. D. Bain, R. Ballantyne, C. Mills, and N. Lester, *Reflecting on practice: student teachers' perspectives*. Flaxton, Qld.: Post Pressed, 2002.
[28] J. Bain, R. Ballantyne, C. Mills, and N. Lester, "Reflecting on Practice: Student Teachers' Perspectives." Accessed: Apr. 03, 2025. [Online]. Available: https://www.researchgate.net/publication/43505901_Reflecting_on_Practice_Student_Teachers'_Perspectives
[29] K. R. Csavina, C. R. Nethken, and A. R. Carberry, "Assessing Student Understanding of Reflection in Engineering Education," in *2016 ASEE Annual Conference & Exposition*, Jun. 2016. [Online]. Available: https://peer.asee.org/assessing-student-understanding-of-reflectionin-engineering-education
[30] R. R. Rogers, "Reflection in Higher Education: A Concept Analysis," *Innov. High. Educ.*, vol. 26, no. 1, pp. 37–57, Sep. 2001, doi: 10.1023/A:1010986404527.
[31] M. Ryan, "The pedagogical balancing act: Teaching reflection in higher education," *Teach. High. Educ.*, vol. 18, no. 2, pp. 144–155, 2013.
[32] M. E. Ryan and M. Ryan, "A Model for Reflection in the Pedagogic Field of Higher Education," in *Teaching Reflective Learning in Higher Education: A Systematic Approach Using Pedagogic Patterns*, M. E. Ryan, Ed., Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-09271-3.
[33] C. J. McColley, M. Alrizqi, and A. Werth, "WIP: Development of a Qualitative Codebook to Study Reflective Depth in Engineering Courses," *Front. Eng. Conf.*, 2025.
[34] A. J. Brockman, D. E. Naphan-Kingery, and R. N. Pitt, "When talent goes unrecognized: racial discrimination, community recognition, and STEM postdocs' science identities," *Stud. Grad. Postdr. Educ.*, vol. 13, no. 2, pp. 221–241, May 2022, doi: 10.1108/SGPE-12-2020-0079.
[35] V. Braun and V. Clarke, *Thematic Analysis: A Practical Guide*. Sage, 2022.
[36] W. Vach and O. Gerke, "Gwet's AC1 is not a substitute for Cohen's kappa – A comparison of basic properties," *MethodsX*, vol. 10, p. 102212, 2023, doi: 10.1016/j.mex.2023.102212.
[37] E. C. Matthay and M. M. Glymour, "A Graphical Catalog of Threats to Validity: Linking Social Science with Epidemiology," *Epidemiology*, vol. 31, no. 3, pp. 376–384, May 2020, doi: 10.1097/EDE.0000000000001161.

[38] H. K. Ro, D. Merson, L. R. Lattuca, and P. T. Terenzini, “Validity of the Contextual Competence Scale for Engineering Students,” *J. Eng. Educ.*, vol. 104, no. 1, pp. 35–54, Jan. 2015, doi: 10.1002/jee.20062.
[39] C. Honda and T. Ohyama, “Homogeneity score test of AC1 statistics and estimation of common AC1 in multiple or stratified inter-rater agreement studies,” *BMC Med. Res. Methodol.*, vol. 20, p. 20, Feb. 2020, doi: 10.1186/s12874-019-0887-5.
[40] T. Jowsey, V. Braun, V. Clarke, D. Lupton, and M. Fine, “We reject the use of generative artificial intelligence for reflexive qualitative research,” *Qual. Inq.*, p. 10778004251401851, 2025.

Appendix A

Table A.1: Version 2 of Codebook

Code	Definition
Personal Connection Reporting/Responding (P.R1)	Students notice and speak about personal experience, thoughts, beliefs, challenges, or opinions.
Personal Connection Relating (P.R2)	Students reflect on the issue in terms of their prior experiences with the issue, a related issue, or in a similar setting. They make connections and comparisons with their prior experiences, thoughts, beliefs, challenges or opinions, and consider how these relate to the reflection prompt or experience
Personal Connection Reasoning (P.R3)	Students go beyond responding by comparing experience to personal beliefs, experiences to analyzing the context, issue, or possible factors related to their experience. The reflection critically examines the reasons behind an event.
Personal Connection Reconstructing (P.R4)	Student demonstrates new ideas and ways of thinking related to personal experience, thoughts, beliefs, challenges, or opinions compared to how they thought/believed previously. The reflection synthesizes disparate pieces of information to form a cohesive understanding or new perspective. Students may utilize these lessons or patterns and apply them to situations beyond immediate context or reflection.
Alignment with Learning Objectives Reporting/Responding (L.R1)	Student response includes mention of intended specific learning objectives.
Alignment with Learning Objectives Relating (L.R2)	Students reflect on the experience in terms of their prior experiences with the issue, a related issue, or in a similar setting. They make connections and comparisons with their prior experiences and consider how these relate to the reflection prompt or experience centered around intended learning objectives.
Alignment with Learning Objectives Reasoning (L.R3)	Students go beyond responding by comparing experience to intended learning objectives. They also analyze the context, issue, or possible factors related to their experience. The reflection critically examines the reasons behind an event.
Alignment with Learning Objectives Reconstructing (L.R4)	Students demonstrate new ideas and ways of thinking related to intended learning objectives compared to how they thought/believed previously. The reflection synthesizes disparate pieces of information to form a cohesive understanding or new perspective. Students may utilize these lessons or patterns and apply them to situations beyond immediate context or reflection.
Discipline Connection Reporting/Responding (D.R1)	Student response includes mention of larger discipline.
Discipline Connection Relating (D.R2)	Students reflect on the experience in terms of their prior experiences with the issue, a related issue, or in a similar setting. They make connections and comparisons with their prior experiences and consider how these relate to the larger discipline.
Discipline Connection Reasoning (D.R3)	Students go beyond responding by comparing experience to larger discipline. They also analyze the context, issue, or possible factors related to their experience. The reflection critically examines the reasons behind an event.
Discipline Connection Reconstructing (D.R4)	Students demonstrate new ideas and ways of thinking related to the larger discipline compared to how they thought/believed previously. The reflection synthesizes disparate pieces of information to form a cohesive understanding or new perspective. Students may utilize these lessons or patterns and apply them to situations beyond immediate context or reflection.
Connection with Others Reporting/Responding (O.R1)	Student response includes mention of perspectives and/or contributions of others not of their own. Simply stating, “We did this,” is not sufficient while, “X person did/said this,” would fall under this code.
Connection with Others Relating (O.R2)	Students reflect on the experience in terms of the prior experiences of others with the issue, a related issue, or in a similar setting. They make connections and comparisons with others’ prior experiences and consider how these relate to the current context.
Connection with Others Reasoning (O.R3)	Students go beyond responding by comparing experience to others’ experiences not of their own. They also analyze the context, issue, or possible factors related to their experience. The reflection critically examines the reasons behind an event.
Connection with Others Reconstructing (O.R4)	Students demonstrate new ideas and ways of thinking related to others’ experiences and perspectives compared to how they thought/believed previously. The reflection synthesizes disparate pieces of information to form a cohesive understanding or new perspective. Students may utilize these lessons or patterns and apply them to situations beyond immediate context or reflection.

Assignment Connection Reporting/Responding (A.R1)	Student responses notice, mentions, or speaks to specific assignments and/or course activities.
Assignment Connection Relating (A.R2)	Students reflect on the course assignment/activities in terms of prior experiences and/or course assignment, a related issue, or in a similar setting. They make connections and comparisons and consider how these relate to the current context.
Assignment Connection Reasoning (A.R3)	Students go beyond responding by comparing experience to specific course assignments or activities. They also analyze the context, issue, or possible factors related to their experience. The reflection critically examines the reasons behind an event.
Assignment Connection Reconstructing (A.R4)	Students demonstrate new ideas and ways of thinking related to course assignments/activities compared to how they thought/believed previously. The reflection synthesizes disparate pieces of information to form a cohesive understanding or new perspective. Students may utilize these lessons or patterns and apply them to situations beyond immediate context or reflection.

Appendix B

Table B.1: Version 3 Codebook, with examples provided for personal connection codes

Code Category	Definition
Situatedness	The *content* of student reflection responses that are contextually grounded, incorporating various elements that relate to their personal and academic experiences, as well as the broader discipline or field. Interested in what topics, specific beliefs, experiences are included in reflection responses.
Abstraction	The process by which students distill and generalize experiences and knowledge in reflections. May include responding to the reflection prompt, relating personal experiences or knowledge about the subject, providing some evidence to reason their response to the reflection prompts, reconstructing previous views and beliefs based on new knowledge gained or experienced and/or apply them beyond the reflection prompt or learning environment being presented to the student.

Note: Text that is **bolded** below represents changes made to codebook V3 as compared to codebook V2 in Appendix A. To save space, we do not include example excerpts for all codes, only Personal Connection (P) codes.

Situatedness Code	Definition and Notes	Abstracted Subcode		Definition	Example
Personal Connection		Reporting/ Responding	P.R1	Students notice and speak about personal experience, thoughts, beliefs, challenges, or opinions.	"More coordinated in that each person seemed to do what we were assigned to do. Shane was collecting data, Simon was recording it, and I was writing summaries in the lab notes as to what was going on." In this example, the exercept, "...and I was writing summaries in the lab notes as to what was going on," would be P.R1
	Student response includes personal experiences, thoughts, beliefs or opinions. This reflection is closely tied to the student's actual experiences rather than generalizations Keywords or phrases include: "I believe...", "Based on my experience with...", "I think..." References real challenges, interactions, or moments from the students' personal experiences	Relating	P.R2	Students reflect on the issue in terms of their prior experiences with the issue, a related issue, or in a similar setting. They make connections and comparisons with their prior experiences, thoughts, beliefs, challenges or opinions.	"I felt like I'm getting more creative. I came up with the idea of connecting the straw to the spray bottle, and I also suggested to use safety pins and tooth floss to sew the gut in the patients if bleeding cannot be stopped." In this example, the student has a comparison of their creativity evidenced by the use of "more." This reflection would NOT be considered P.R3 as they are not providing reasoning for their ideas or decisions, they are two separate thoughts.
		Reasoning	P.R3	Students go beyond responding by comparing experience to personal beliefs, experiences to analyzing the context, issue, or possible factors related to their experience. The reflection critically examines the reasons behind an event.	"Three regulatory agencies I believe are important in deciding standards and safety requirements in biomedical engineering are the FDA, the ISO, and OSHA. I think the FDA is important because they make sure that a medical device is able to perform as intended and does so by providing various assessments of said device." In this example, the student shares their thoughts as why they believe the FDA is important. They provide a justification that"...they make sure that a medical device is able to perform as intended and does so by providing various assessments of said device." They is also an indication of reasoning where they say X statement BECAUSE Y statement.

Situatedness Code	Definition and Notes	Abstracted Subcode		Definition	Example
		Reconstructing	P.R4	Student demonstrates new ideas and ways of thinking related to personal experience, thoughts, beliefs, challenges, or opinions compared to how they thought/believed previously. The reflection synthesizes disparate pieces of information to form a cohesive understanding or new perspective. Students may utilize these lessons or patterns and apply them to situations beyond immediate context or reflection.	"Focusing on quantity over qualirty made me more creative as I forced myself to think through the process and to consider various factors which I overlooked during previous study. While for sure more ideas have been brought up out of this process, I also notice that it was hard to develop the first few ones and that more creative I had become as I made the first few steps." In this example, P.R4 is applied because the student is demonstrating a new way of thinking/perception/experience that is born from a previously held way of thinking. The student recognizes that they overlooked aspects of a process in a previous study but through this experiences and "focusing on quantity over quality" made them more creative, forcing themselves into a new way of thinking and consideration of various factors.
Alignment with learning outcomes	Student response includes application of specific learning objectives related to the course and/or activity [Insert LOs here] **This is different from Assignments, which refer to experiences in specific course activities. Learning Objectives refer to the display of specific intended outcomes of the course assignments.** Keywords or phrases depends on learning objectives	Reporting/ Responding	L.R1	Student response includes application of intended specific learning objectives.	
		Relating	L.R2	Students reflect on the experience in terms of their prior experiences with the issue, a related issue, or in a similar setting. They make connections and comparisons with their prior experiences and consider how these relate to the reflection prompt or experience centered around intended learning objectives.	
		Reasoning	L.R3	Students go beyond responding by comparing experience to intended learning objectives. They also analyze the context, issue, or possible factors related to their experience. The reflection critically examines the reasons behind an event.	
		Reconstructing	L.R4	Students demonstrate new ideas and ways of thinking related to intended learning objectives compared to how they thought/believed previously. The reflection synthesizes disparate pieces of information to form a cohesive understanding or new perspective. Students may utilize these lessons or patterns and apply them to situations beyond immediate context or reflection.	
Discipline Connection	Student response includes relation and impact to the related discipline or field (e.g., engineering, physics) and/or field related principles, theories, knowledge in a broader scope outside of the course.	Reporting/ Responding	D.R1	Student response includes mention of larger discipline **or field (e.g., engineering, physics) and/or field related principles, theories, knowledge in a broader scope outside of the course.**	
		Relating	D.R2	Students reflect on the experience in terms of their prior experiences with the issue, a related issue, or in a similar setting. They make connections and comparisons with their prior experiences and consider how these relate to the larger discipline.	

Situatedness Code	Definition and Notes	Abstracted Subcode		Definition	Example
	Keywords or phrases include reference to specific field or discipline (e.g., "Because this is important to biomedical engineering..."	Reasoning	D.R3	Students go beyond responding by comparing experience to larger discipline. They also analyze the context, issue, or possible factors related to their experience. The reflection critically examines the reasons behind an event.	
		Reconstructing	D.R4	Students demonstrate new ideas and ways of thinking related to the larger discipline compared to how they thought/believed previously. The reflection synthesizes disparate pieces of information to form a cohesive understanding or new perspective. Students may utilize these lessons or patterns and apply them to situations beyond immediate context or reflection.	
Peer Behavior	Student response includes experiences of other people, not themselves Keywords or phrases include: "My teammate believed..." , "My [insert other person(s)] thought or experienced..."	Reporting/ Responding	O.R1	Student response includes mention of perspectives and/or contributions of others not of their own. Simply stating, "We did this," is not sufficient while, "X person did/said this," would fall under this code.	
		Relating	O.R2	Students reflect on the experience in terms of the prior experiences of others with the issue, a related issue, or in a similar setting. They make connections and comparisons with others' prior experiences and consider how these relate to the current context.	
		Reasoning	O.R3	Students go beyond responding by comparing experience to others' experiences not of their own. They also analyze the context, issue, or possible factors related to their experience. The reflection critically examines the reasons behind an event.	
		Reconstructing	O.R4	Students demonstrate new ideas and ways of thinking related to others' experiences and perspectives compared to how they thought/believed previously. The reflection synthesizes disparate pieces of information to form a cohesive understanding or new perspective. Students may utilize these lessons or patterns and apply them to situations beyond immediate context or reflection.	
Assignment Connection	Student response includes or references specific assignments or course activities **(e.g., reflections, dissections, test, lab, quiz, problem set, coding).** **This is different from Learning Objective; Assignments refer to experiences in specific course activities, whereas Learning Objectives refer to the display of specific intended outcomes of the course assignments.** Keywords or phrases include: "This assignment..." , "[Named	Reporting/ Responding	A.R1	Student responses notice, mentions, or speaks to specific assignments and/or course activities (e.g., reflections, dissections, test, lab, quiz, problem set, coding).	
		Relating	A.R2	Students reflect on the course assignment/activities (e.g., reflections, dissections, test, lab, quiz, problem set, coding) in terms of prior experiences and/or course assignment, a related issue, or in a similar setting. They make connections and comparisons and consider how these relate to the current context.	
		Reasoning	A.R3	Students go beyond responding by comparing experience to specific course assignments or activities (e.g., reflections, dissections, test, lab, quiz, problem set, coding). They also analyze the context, issue, or possible factors related to their experience. The reflection critically examines the reasons behind an event.	

Situatedness Code	Definition and Notes	Abstracted Subcode		Definition	Example
	activity]...", "While working on this in class..." Learning environment and experiences within environment is mentioned **(e.g., reflections, dissections, test, lab, quiz, problem set, coding)**	Reconstructing	A.R4	Students demonstrate new ideas and ways of thinking related to course assignments/activities (e.g., reflections, dissections, test, lab, quiz, problem set, coding) compared to how they thought/believed previously. The reflection synthesizes disparate pieces of information to form a cohesive understanding or new perspective. Students may utilize these lessons or patterns and apply them to situations beyond immediate context or reflection.	
Group Experience	Student response includes ***shared or group*** experiences, thoughts, beliefs or opinions. **"Keywords or phrases include: ""We believe..."", ""Based on our..."""** **References real challenges, interactions, or moments from the students' shared or group experiences within the classroom context**	Reporting/ Responding	G.R1	Student response includes mention of shared perspectives, decisions, and/or contributions of their entire group or multiple group members.	
		Relating	G.R2	Student reflects on their shared team experience in terms of their team's prior experiences with the issue, a related issue, or a similar setting. They make connections and comparisons with their shared prior experiences and how they relate to the current context.	
		Reasoning	G.R3	Students go beyond responding by comparing shared experiences to others' experiences not their own. They also analyze the context, issue, or possible factors related to their own experience. The reflection critically examines reasons behind an event.	
		Reconstructing	G.R4	Students demonstrate new ideas and ways of thinking related shared experiences compared to how they thought/believe previously. The reflection synthesizes disparate pieces of information to form a cohesive understanding or new perspective. Students may utilize these lessons or patterns and apply them to situations beyond immediate context or reflection.	

Appendix C

Table C.0: Individual Code Reliability Metrics: V2 vs V3 Codebook Comparison

Code	V2 κ	V3 κ	Δκ	V2 Agr%	V3 Agr%	V3 α
A.R1	0.412	0.635	+0.223	70.7	81.4	0.277
A.R2	0.417	1.000	+0.583	88.0	100.0	—
A.R3	0.232	0.371	+0.139	63.0	72.1	0.175
A.R4	0.167	0.645	+0.478	82.6	95.4	—
L.R1	0.216	0.398	+0.182	63.0	69.8	0.138
L.R2	0.465	0.000	−0.465	91.3	97.7	—
L.R3	0.204	0.269	+0.065	64.1	72.1	0.035
L.R4	0.157	0.645	+0.488	81.5	95.4	0.550
O.R1	0.224	0.550	+0.326	73.9	90.7	0.311
O.R2	0.000	—	—	91.3	100.0	—
O.R3	0.044	1.000	+0.956	83.7	100.0	—
O.R4	0.000	—	—	91.3	100.0	—
P.R1	0.393	0.389	−0.004	70.7	69.8	0.345
P.R2	0.225	0.000	−0.225	93.5	97.7	—
P.R3	0.356	0.538	+0.182	73.9	81.4	0.326
P.R4	0.194	0.534	+0.340	79.3	93.0	0.471
D.R1	0.383	0.239	−0.144	91.3	81.4	0.205
D.R2	0.000	—	—	95.7	100.0	—
D.R3	0.419	0.173	−0.246	94.6	86.0	0.170

Note. κ = Cohen's Kappa; Δκ = Change in Kappa from V2 to V3; Agr% = Percent Agreement. Dashes (—) indicate undefined values due to perfect agreement or insufficient variance. Positive changes indicate improvement.

Quantitative Interrater Agreement Coefficient Calculations Based on Contingency Tables

Gwet's (2014) [7] Handbook of interrater reliability (IRR) bases all rater agreement calculations in equations and 'contingency tables' [7, p. 31]. The following sections detail the varied quantitative interrater agreement coefficients we employed using this methods, detailing how they are calculated and what they represent with an example throughout. The basis example provided relates to our "Assignment" situatedness codes in our Version 2 codebook, but we ran calculations for all codes present in both V2 and V3 and present those in the Results. The base example contingency table of observed frequencies for the set of Assignment codes in our codebook for Author 1 and 2's V2 coding is presented in Table C.1 below. Here, the numbers along the diagonal of the matrix (i.e., [34, 5, 17, 2] below) represent the number of student reflection responses both authors coded as A.R1, A.R2, A.R3, and A.R4, respectively. The numbers outside of the diagonal represent the number of excerpts observed where one author coded them one way, and the other, differently. To illustrate, 13 excerpts were observed as being coded as A.R1 by Author 1 and as A.R3 by Author 2.

Table C.1: Observed Frequencies of Assignment Codes Contingency Table for Author 1 and 2 Interrater Agreement

		Author 1			
		A.R1	*A.R2*	*A.R3*	*A.R4*
Author 2	*A.R1*	34	4	12	8
	A.R2	0	5	2	0
	A.R3	13	7	17	10
	A.R4	2	0	0	2

(1) Simple percent agreement

Our simplest interrater reliability study [7] consisted of evaluating the 'percent agreement' between our two raters coding the same set. Percent agreement (p_a) for n subjects where there are four optional codes to be applied (as per our example) is given by the general form with the following generalized table. Fewer or more codes expand both the equation for percent agreement and its reference contingency table.

$$Pa = \frac{(n_{11} + n_{22} + n_{33} + n_{44})}{n}$$

Table C.2: Observed Frequencies Generalized Table

		Author 1				
		A.R1	***A.R2***	***A.R3***	***A.R4***	
Author 2	***A.R1***	n11	n12	n13	n14	*n+1*
	A.R2	n21	n22	n23	n24	*n+2*
	A.R3	n31	n32	n33	n34	*n+3*
	A.R4	n41	n42	n43	n44	*n+4*
Total		*n+1*	*n+2*	*n+3*	*n+4*	n

Percent agreement for Authors 1 and 2 coding of n=92 excerpts with V2 is thus 63%, which is considered reasonably high agreement.

$$p_a = \frac{(34 + 5 + 17 + 2)}{92} = 0.63 \; or \; 63\,\%$$

(2) Cohen's Kappa and addressing chance agreement

Percent agreement however is known to overstate IRR [7] because the calculation does not account for chance agreement. In other words, what if one of our coders was ignoring a particular excerpt's characteristics and chose to categorize it at random? With a small enough number of codes like our four Assignment codes, Author 1 could still end up categorizing (with a 25% chance) an excerpt in the same group as Author 2 regardless of their random choice. To account for this potential for "lucky agreement" [7, p. 32], Gwet proposes Cohen's Kappa (κ) for measuring beyond chance. We also calculated Cohen's Kappa for our data, which is given by:

$$\kappa = \frac{p_a - p_c}{1 - p_c}$$

To get the term for regular percent chance agreement (p_c) in the equation above, first we sum expected probabilities that raters classify a subject into a given category. This time, we use a contingency table for expected probabilities [7] as opposed to expected frequencies for regular percent agreement. To calculate expected probabilities that our two Authors classified an excerpt, we use our observed frequencies table (Table C.2) and sum our observed frequencies along each column and row (as well as along the diagonal) divided by the total number of excerpts (in our case, n=92). This calculation of expected probabilities from our observed frequences necessary for calculating percent change agreement (p_c) shown in Table C.3 below.

$$p_c = \frac{n_{1+}}{n} \times \frac{n_{+1}}{n} + \frac{n_{2+}}{n} \times \frac{n_{+2}}{n} + \frac{n_{3+}}{n} \times \frac{n_{+3}}{n} + \frac{n_{4+}}{n} \times \frac{n_{+4}}{n} \ldots$$

Table C.3: Expected Probabilities of Assignment Codes for Author 1 and 2 Interrater Agreement Calculated from Observed Frequencies

		Author 1				
		A.R1	***A.R2***	***A.R3***	***A.R4***	***Probability***
Author 2	***A.R1***	34	4	12	8	*= 68/92* ***= 63%***
	A.R2	0	5	2	0	*= 7/92* ***= 8%***
	A.R3	13	7	17	10	*= 47/92* ***= 51%***
	A.R4	2	0	0	2	*= 4/92* ***= 4%***
	Probability	*= 49/92* ***= 53%***	*= 16/92* ***= 17%***	*= 31/92* ***= 34%***	*= 20/92* ***= 22%***	

Percent chance agreement (p_c) for this example is thus equal to:

$$p_c = \frac{49}{92} \times \frac{68}{92} + \frac{16}{92} \times \frac{7}{92} + \frac{31}{92} \times \frac{47}{92} + \frac{20}{92} \times \frac{4}{92} = 0.58894 \text{ or } \sim 59\%.$$

Plugging this percent chance calculation into the equation for Cohen's Kappa along with our previously calculated regular percent agreement gives:

$$\kappa = \frac{p_a - p_c}{1 - p_c} = \frac{0.63 - 0.59}{1 - 0.59} = \sim 0.0976$$

Pushing EER Beyond Cohen's Kappa

Martin et al.'s (2025) [5] systematic review of reporting reliability in EER found that 57% of surveyed articles made no mention to IRR, 91% did not report any reliability metrics and the other 9% used Cohen's Kappa. They therefore highlighted the need for calculations *at least* as far as we have thus discussed, and recommend scholars provide these statistics with appropriate justifications for use considering datatype and number of raters.

To push STEM Education Research further in qualitative coding reliability reporting further, we also calculated and present three additional interrater agreement coefficients beyond this simple form of Cohen's Kappa.

1. **Weighted Cohen's Kappa** to address codebooks with dependent ratings, or codes that are simply categorical in nature and independent of each other as opposed to being ranked or ordered (i.e., ordinal) [7, p. 35].
2. **Krippendorff's Alpha** [7, p. 39] for addressing missing data, assessing reliability based solely on excerpts where both raters have coded them to avoid Kappa sensitivity to imbalanced code assignments.
3. **Gwet's AC1,** which is based on based on both regular percent agreement and a Gwet [7] percent chance agreement calculation addressing a Cohens Kappa limitation where it can "yield a low value when the raters show high agreement" [7, p. 104] and instances of unbalanced data.

(3) Weighted Cohen's Kappa and dealing with ordinal ratings

Interrater agreement coefficient calculations presented thus far have assumed our V2 codebook features nominal ratings, where our codes for both situatedness and abstraction in the Appendix A Table A.1 are considered distinguishable but not sortable to a meaningful order [7]. The problem with this assumption is that our abstraction codes are indeed ordered – reporting/responding is considered a lower level of abstraction in our codebook compared to relating or reconstructing. Additionally, our codes are dependent of each other in every version of our codebook; we considered there to be more agreement when our two raters categorized an excerpt as "reasoning" and "reconstructing" than when we categorized that same excerpt as "responding" and "reconstructing."

In the unweighted Cohen's Kappa shown thus far, the expression for Percent chance agreement (p_c)

$$p_c \quad where \quad \frac{n_{1+}}{n} \times \frac{n_{+1}}{n} + \frac{n_{2+}}{n} \times \frac{n_{+2}}{n} \ldots$$

as shown earlier represents probability of agreement between two raters if the codes are known to be independent. When this is not the case, this calculation "does not have a particular meaning and does not represent a measure of agreement. Using it in the Kappa equation may yield unpredictable results" [7, p. 35].

Returning to our contingency tables, we also calculated weighted versions of Cohen's Kappa according to the following equation for Weighted Cohen's Kappa (κ_w), where w_{ij} refers to weighting factors in a weighting matrix W, and p_{ij} and e_{ij} refer to individual elements of the observed frequencies contingency table (example in Table C.2 above) and a new expected frequences contingency Table (example Table C.5 below) derived from expected probabilities calculated in Table C.4. In other words, the expected frequences contingency table as shown (Table C.5) is calculated by first constructing a new expected probabilities contingency table (Table C.4). Pulling the expected probabilities of Assignment codes from Table C.2, each cell of the new expected probabilities contingency table is calculated by multiplying a row and associated column probability together as shown.

$$\kappa_w = 1 - \frac{\sum w_{ij} \cdot p_{ij}}{\sum w_{ij} \cdot e_{ij}}$$

Table C.4: Expected Probabilities Contingency Table

		Author 1			
		A.R1	*A.R2*	*A.R3*	*A.R4*
Author 2	*A.R1*	0.53 * 0.63 = ***0.34***	0.17 * 0.63 = ***0.11***	0.34*0.63 = ***0.21***	0.22*0.63 = ***0.14***
	A.R2	0.53*0.08 = ***0.04***	0.17*0.08 = ***0.01***	0.34*0.08 = ***0.03***	0.22*0.08 = ***0.02***
	A.R3	0.53*0.51 = ***0.27***	0.17*0.51 = ***0.09***	0.34*0.51 = ***0.17***	0.22*0.51 = ***0.11***
	A.R4	0.53*0.04 = ***0.02***	0.17*0.04 = ***0.01***	0.34*0.04 = ***0.01***	0.22*0.04 = ***0.01***

Multiplying each of these expected probabilities by the number of excerpts coded (n=92) in Table C.4 above, we get the expected frequencies contingency table (Table C.5 below) with cells e_{ij}.

Table C.5: Expected Frequences Contingency Table

		Author 1			
		A.R1	*A.R2*	*A.R3*	*A.R4*
Author 2	*A.R1*	30.89	10.09	19.54	12.61
	A.R2	3.73	1.22	2.36	1.52
	A.R3	25.03	8.17	15.84	10.22
	A.R4	2.13	0.70	1.35	0.87

Once these elements of the Weighted Cohen's Kappa equation are calculated, weighting matrices W are chosen based on how "close together" the codebook's codes are considered to each other. Linear and quadratic weightings are common, and some researchers weight based on the number of excerpts one rater has rated compared to the other to address differences in association with the dataset [7].

Linear weighting reflects disagreement where a 0 penalty is assigned for perfect agreement and penalties for disagreement increase linearly with the distance between ratings (with a maximum penalty for largest possible disagreement). Think of linear weighting as saying: "Not all disagreements are equal. A small disagreement counts as partly correct; a big disagreement counts as fully wrong, with equal intervals in between." Quadratic weighting works similarly, but it penalizes larger disagreements much more strongly than linear weighting. Where linear weighting increases penalties proportionally to the distance between ratings, quadratic weighting increases penalties with the square of the distance.

The choice of linear and quadratic weighting schemes was grounded in the theoretical structure of Ryan & Ryan's 4R [32] framework, which proposes that reflection quality progresses through four ordered levels: Responding (R1), Relating (R2), Reasoning (R3), and Reconstructing (R4). This hierarchy implies

that adjacent levels are more conceptually similar than distant ones – R1 and R2 share more surface-level engagement with content, while R1 and R4 represent more qualitatively distinct cognitive acts. . Linear weighting was selected to test whether rater disagreements followed this equal-interval theoretical ordering, assigning proportionally larger penalties as the distance between assigned codes increases. Quadratic weighting was added as a sensitivity check, applying disproportionately heavier penalties for disagreements spanning distant levels (e.g., R1 vs. R4) than adjacent ones (e.g., R1 vs. R2), thus probing whether the theoretical hierarchy was reflected in rater behavior more strongly at the extremes. Together, these two schemes allowed us to examine not just whether raters disagreed, but how their disagreements mapped onto or diverged from the progression that Ryan & Ryan proposed.

Here are example weighting matrices W for linear and quadratic calculations of Cohen's Kappa. In each weighted matrix, diagonal values of 1 represent perfect agreement between raters with no disagreement penalty applied; off-diagonal values reflect the proportional (linear) or squared (quadratic) distance penalty between abstraction levels, where values closer to 0 indicate larger penalties for greater disagreement.

$$W\ (for\ unweighted\ \kappa) = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

$$Linear\ W\ (for\ weighted\ \kappa) = \begin{bmatrix} 1 & 0.667 & 0.333 & 0 \\ 0.667 & 1 & 0.667 & 0.333 \\ 0.333 & 0.667 & 1 & 0.667 \\ 0 & 0.333 & 0.667 & 1 \end{bmatrix}$$

$$Quadratic\ W\ (for\ weighted\ \kappa) = \begin{bmatrix} 1 & 0.889 & 0.556 & 0 \\ 0.889 & 1 & 0.889 & 0.556 \\ 0.556 & 0.889 & 1 & 0.889 \\ 0 & 0.556 & 0.889 & 1 \end{bmatrix}$$

Once the Weighting Matrix W, Observed Frequencies of Contingency Table p_{ij}, and Expected Frequences Contingency Table e_{ij} are prepared, each individual cell of the Weighting Matrix W is multiplied with each associated cell in Observed Frequencies of Contingency Table p_{ij}. Then they are all summed together ($\sum w_{ij} \cdot p_{ij}$), and the same operation is performed between each individual cell of the Weighting Matrix W and the cells of the Expected Frequences Contingency Table e_{ij} to give $\sum w_{ij} \cdot e_{ij}$, the numerator and denominator of the fraction in the equation for Weighted Cohen's κ_w.

$$\kappa_w = 1 - \frac{\sum w_{ij} \cdot p_{ij}}{\sum w_{ij} \cdot e_{ij}}$$

Our Linear Weighted Cohen's Kappa interrater agreement coefficient for Assignment coding was $\kappa_w = 0.076$.

(4) Krippendorff's Alpha for dealing with missing data cleanly

Krippendorff's α is designed to handle incomplete Observed Frequency Contingency Tables, where every coder has not coded all excerpts or there is imbalanced category use (e.g., one coder used the "reconstruction" code much more than any other). In calculating α, reliability is we compute using only shared ratings, making the coefficient more stable and interpretable when codes do not use all codes equally. In contrast, Cohen's Kappa calculations can behave unreliably if raters use categories unevenly because it is sensitive to rare coding assignments or skewed coding distributions [7].

In the case of our dataset, we did not calculate Cohen's Kappa based on any incomplete data (i.e., both our coders coded the 20 and 10 reflection sessions mentioned thus far as opposed to us needing to remove single rater data), so the first step of calculating Krippendorff's α is taken care of – "all subjects rated by a single

rater must be eliminated upfront" [7, p. 39]. For data with all subjects rated by both raters (i.e., all subjects rated by a single coder removed) like ours, Krippendorff's α is calculated as follows (for the two-rater case):

$$\alpha_\kappa = \frac{(p_a - p_e)}{(1 - p_e)},\ where\left\{p_{e=}\ (\pi_{hat})^2 + (1-\pi_{hat})^2\right\} and\left\{p_a = (1-\varepsilon_n)p_a + \varepsilon_n\right\}$$

Where, for P_{1+} and P_{+1} being raters' propensity for classifying a subject into category "1," π_{hat} is given by the following equation:

$$\pi_{hat} = \frac{(p_{1+} + p_{+1})}{2}$$

It is important to note that our dataset exhibited extreme skewness (>95% or <5% for a given code like D.R2 that was quite rare) and insufficient variance for multiple codes in our codebook, making Krippendorff's Alpha unideal and worth corroborating with other IRR coefficients. For Assignment code A.R1 in the V3 codebook, Krippendorff's Alpha came out to 0.277, but many of our Alpha values resulted as NA due to assumption failures.

(5) Gwet's AC1 and a new percent chance agreement

Gwet's AC1 is similar to Cohen's Kappa in that both analyses compare observed agreement rate, but Gwet's AC1 uses expected disagreement and Kappa uses expected agreement. This fundamental different in assumptions means that AC1 can be both positive or negative in cases of no association between raters, but Kappa always gives 0 for these cases. Vach & Gerke [36] suggest Gwet's AC1 should not be seen as a substitute for Cohen's kappa [36, p. 1], though AC1 responds to two key "Paradoxes of Kappa" [7, p. 58] originally described by Feinstein & Cicchetti (1990). The first is that if percent chance agreement is large, Kappa calculations can convert a relatively high value of percent agreement to a relatively low Kappa. The second is that unbalanced totals (or marginal application of codes) produce higher values of Kappa than more balanced totals [7]. Our dataset, though relatively diverse across different student populations, exhibits both of these paradoxes upon review of percent chance agreement alongside Kappa results and the integration of quantitative and qualitative codebook validation (see Results section). These errors are likely more common in EER outside of our case alone based on known challenges with reliably assessing student reflections in medical education research [9], [10]; Kappa paradoxes require oversight in EER codebooks related to undergraduate student reflection as well.

Gwet's AC1, developed to overcome these paradoxes, is calculated as follows. AC1 is based on the same percent agreement Pa as first described in this Appendix C, but it also depends on a new percent agreement. The new percent agreement, is equal to: $p_{c\,new} = 2 * \pi_{hat}(1 - \pi_{hat})$ and Gwet's AC1 is then given by:

$$GAC1 = \frac{p_a - p_{cnew}}{1 - p_{cnew}}$$

(6) Multiple Interrater Percent Agreement Calculations (simple percent agreement, Gwet's AC1, and the Holley-Guilford G-Index)

As discussed earlier in this Appendix, Cohen's Kappa addresses simple percent agreements' lack of consideration for chance agreement by including percent chance agreement in its calculation. We have also covered how Gwet's AC1 was proposed to address paradoxes of Cohen's Kappa. To provide one more comparative understanding of percent agreement in this study, we also calculated the Holley-Guilford G-Index for our codebooks. Here, instead of Kappa's percent chance agreement or AC1's calculation for it, G-Index operates on a simple percent chance of 1/n, where n is the number of categories used in a given round of coding (for us, this would be our total number of codes in our V2 and V3 codebooks), but it uses the same percent agreement p_a as the aforementioned approaches.

$$G - Index = \frac{p_a - \left(\frac{1}{n}\right)}{1 - \left(\frac{1}{n}\right)} \; where\; n = number\; of\; categories\; coded$$

The benefit of using the G-Index calculation as shown is the magnitude of this reliability metric is considered more reasonable than simple percent agreement, which is known to inflate results compared to Cohen's Kappa due to chance agreement even though Kappa still suffers from challenges with unbalanced data and overestimation of probability based on chance [7], [39]. Additionally, Gwet's AC1 only performs better than G-index for the case of two raters, as discussed in medical education literature [39], though Gwet does offer an AC2 calculation for more than two rater cases [7].

With our V3 codebook and desire to compare GenAI coding to human coding, we needed an interrater reliability metric appropriate for our 3-rater (two humans and one GenAI multi-agent model) analyses. G-index was therefore a strong contender for addressing limitations of Gwet's AC1 and other prior calculations by providing a more generalized percent agreement calculation for cases with more than two raters. Based on the structure of our V3 codebook, our n values and therefore equal probabilities for the G-index calculations as shown above were: (i) for the 4-level abstraction codes (R1, R2, R3, R4) = 4 categories = 1/4 probability each, and (ii) for R-levels with categories {0, A, L, O, P, D, G} = 7 categories = 1/7 probability each. Note: Gwet's AC2 does provide a calculation for 2+ rater cases, but to provide variety in this paper we focused on G-index instead. AC2 calculations merit comparison with G-index metrics in our future reliability analyses, especially because the G-index, while valuable for addressing some limitations of other interrater reliability metrics, presents its own challenges in assuming equal percent chance for all codes in a given codebook. Qualitative scholars know that codebooks are not a prior designed to include an equal distribution like this; it is an assumption fundamentally and paradigmatically at odds with some forms of qualitative research, in fact (see Braun & Clarke, 2022 [35] and ongoing work on reflexive thematic analysis [40]).