2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION
Paper ID #50394

Learning to Self-Evaluate: How Students Calibrate Self-Assessment in Structured Engineering Studio Environments

Stephanie Fuchs, Cornell University

Dr. Stephanie Fuchs is an Active Learning Initiative (ALI) Postdoctoral Associate in the Department of Biomedical Engineering (BME) at Cornell University. She received her Ph.D. in Biological Engineering from Cornell University, where she focused on developing glucose-sensitive materials for electronics-free insulin delivery devices. As an ALI postdoc, her work focuses on developing and implementing engineering studio modules for core BME courses and developing tools to support student learning in the studios via active learning techniques. She is particularly interested in researching the impact of the engineering studio environment on student learning, engagement, and motivation, and investigating how the new studio curriculum impacts student's perception of their engineering identity.

Prof. Jonathan T. Butcher, Cornell University
Dr. Mridusmita Saikia, Cornell University

Dr. Mridusmita Saikia is a Lecturer at the Meinig School of Biomedical Engineering (BME), Cornell university. She is an interdisciplinary scientist with expertise in biochemistry, molecular biology, and genomics. Dr. Saikia completed her PhD at the University of Chicago, where she developed quantitative and high throughput biochemical assays to analyze RNA modification levels in biological systems. Her work was supported by a fellowship from the Burroughs-Wellcome Trust. Following her PhD, Dr. Saikia conducted postdoctoral research at Case Western Reserve University and Cornell University. Dr. Saikia used single cell RNA sequencing technology to study human immune cell function, as well as human pancreatic beta cell pathology that can lead to diabetes. As a teaching faculty at Cornell BME, Dr. Saikia teaches courses in the fields of Biomaterials & Drug Delivery (BMDD), and Molecular, Cellular, and Systems Engineering (MCSE).

Learning to Self-Evaluate: How Students Calibrate Self-Assessment in Structured Engineering Studio Environments

Abstract

Self-assessment practices have been widely recognized for their potential to deepen conceptual understanding, build metacognitive capacity, and enhance student self-efficacy. When paired with well-designed rubrics, these practices can help students take ownership of their learning by providing clear criteria for evaluating their own progress. Yet the success of these practices varies considerably across disciplines and instructional contexts. This study examines how biomedical engineering students develop self-assessment skills across four iterative design studios using a proficiency rubric for integrated skills measurement (PRISM), which accesses engineering design competencies across four proficiency levels. Following each studio, students individually evaluated their team's collaborative work using the same rubric instructors used for evaluation, providing parallel assessments for systematic comparison. We analyzed paired student-instructor ratings from 64 students in a junior-level BME course to investigate: (1) alignment between student self-assessments and instructor evaluations, (2) changes in alignment from Studio 1 to Studio 4, and (3) variation among team members evaluating the same deliverable. Results showed modest overall alignment (40.6% exact agreement; Gwet's AC1 = 0.208), with 86% of ratings within ±1 proficiency level. Contrary to expectations, calibration worsened over time: mean student-instructor differences increased from 0.15 in Studio 1 to 0.80 in Studio 4, driven by students' increasing tendency to overestimate competence in constraint analysis and connecting quantitative models to qualitative representations. These findings demonstrate that familiarity with assessment criteria alone does not produce calibration; students require explicit instruction in comparing their judgments to expert standards and justifying ratings with concrete evidence. The persistent within-team variation further suggests that shared deliverables do not automatically generate shared understanding—teammates attend to different evidence and interpret rubric criteria differently, particularly for competencies requiring translation between qualitative and quantitative representations. We discuss implications for designing calibration-focused interventions, including structured opportunities for students to compare self-ratings with instructor evaluations, team-based discussions of divergent ratings, and competency-specific scaffolds that guide evidence-based justification.

1. Introduction

Design is a foundational professional competency for engineering students [1], requiring them to navigate ill-defined problems, balance competing constraints, and iterate toward optimal solutions [2-4]. Central to effective design practice is the ability to self-evaluate: to critically evaluate whether one's work is meeting desired objectives, to identify areas for improvement, and to determine which feedback to prioritize and integrate [5, 6]. In engineering design, these capabilities manifest as engineers who can proactively solicit critiques on design proposals, interpret complex technical review comments accurately [7], apply insights to refine prototypes [8], offer constructive input to team members, and effectively manage emotional responses often associated with critical evaluation [9]. Without this capacity for self-evaluation and strategic engagement with feedback, engineers risk iterating inefficiently, failing to recognize when design

solutions fall short of performance requirements, or implementing feedback indiscriminately without exercising professional judgment.

The importance of self-evaluation extends beyond design practice to learning itself [10, 11]. Self-evaluation practices have been widely recognized for their potential to deepen conceptual understanding, build metacognitive capacity, and enhance student self-efficacy [11, 12]. When paired with well-designed rubrics that provide clear, specific criteria and distinct performance levels [13], these practices can help students take ownership of their learning by providing clear criteria for evaluating their own progress [14]. Research shows that rubrics improve both learning outcomes [15, 16] and performance [17-20], increase self-regulation compared to students without rubrics [21], enhance the accuracy of students' self-evaluations [22], and when combined with feedback, strengthen self-efficacy [16, 23].Yet the success of these practices varies considerably across disciplines and instructional contexts [24, 25].

Within engineering education, there is growing recognition of the value of integrating self-assessment into design courses. Structured self-assessment can support learning while improving instructional efficiency; rubric-based approaches requiring students to justify ratings with evidence have been reported to reduce grading time while shifting instructor effort toward coaching and formative feedback [26]. In design and capstone contexts, incorporating self-assessment into assessment cycles has been associated with improvements in student performance and greater independence in judging work quality [27].

At the same time, research cautions that self- and peer-assessments are not inherently accurate. Without calibration, ratings can be systematically biased, with lower-performing individuals or teams tending to overestimate and higher-performing peers sometimes underestimating; agreement with instructor judgments also varies depending on what is being evaluated [28]. These concerns have motivated efforts to improve rating quality through targeted instruction and structured processes, which have been shown to reduce inflated ratings and yield more discriminating peer feedback [29, 30]. Beyond rubric- and teamwork-focused approaches, survey-based self-assessments of design capabilities can be both reliable and sensitive to students' year level [31]. However, self-report data may show ceiling effects or fail to distinguish between students at different experience levels without complementary performance measures [32]. Additionally, students' perceptions of their own learning may not align with what they enjoy or what instructors intend [33]. Collectively, this literature supports the promise of self-assessment in design education while highlighting persistent challenges around accuracy, calibration, and validity.

Despite these advances, less is known about how team members vary in their individual assessments of shared design work, or how these student evaluations align with instructor judgments. This gap is particularly important in biomedical engineering (BME) design education. BME problems require teams to integrate and balance physiological, clinical, and engineering considerations. When team members individually assess shared work, they may prioritize different aspects based on what they value or what aligns with their expertise, potentially influencing which improvements the team pursues in subsequent iterations. Understanding these patterns matters because collaborative design decisions about what to refine and how to iterate emerge from how individual team members evaluate and prioritize competing design requirements. This work examines self-assessment practices in a junior-level BME studio course where students complete

four iterative design challenges. We ask: as students gain experience with proficiency-based assessment frameworks, how do they assess their collaborative design work, and how does this assessment evolve with practice? Specifically, we investigate

1. **RQ1:** How do student's self-assessments of their work in BME studios align with instructor evaluations of the same work?
2. **RQ2:** How does the alignment between students and instructors shift (if at all) going from Studio 1 to Studio 4 as students gain familiarity with studio expectations?
3. **RQ3**: How do individual assessments vary amongst team members evaluating the same studio deliverable?

By analyzing patterns of (mis)alignment, we aim to reveal how students' self-evaluation skills develop and how they calibrate their self-assessments as they gain experience with self-evaluation in BME design.

2. Theoretical Framework

This study draws on two complementary frameworks to assess how students evaluate their own work in BME studios: self-assessment and self-regulated learning (SRL). SRL describes how students take charge of their own learning by setting goals and then actively managing their thinking, motivation, and actions to achieve those goals [34, 35]. Self-assessment, in contrast, involves students examining their completed work, comparing it against specific standards or criteria, and making revisions based on that evaluation [10, 36]. Both practices serve a common purpose: they generate internal feedback that students can use to strengthen their understanding and enhance their performance [6].

2.1 Self-Assessment

Self-assessment encompasses various approaches through which students evaluate and judge the quality of their own learning processes and academic work [37]. It differs from self-grading, where students simply assign themselves scores, and from general self-reflection about strengths and weaknesses. Effective self-assessment is formative in nature and requires three elements: clear articulation of performance expectations, systematic critique of work against those expectations with specific evidence, and opportunities to revise and improve based on insights gained [6].

Goodrich identifies eight conditions necessary for self-assessment to work: students need to (1) understand the purpose and value of self-assessment, (2) have access to clear criteria, (3) have a specific task to assess, (4) see models of good self-assessment, (5) receive direct instruction and feedback on how to do it, (6) get practice, (7) know when to do it, and (8) have opportunities to revise their work [38]. When these conditions are met, self-assessment helps students identify gaps between their current work and the target, then figure out how to close those gaps [39].

However, students' self-assessments are not automatically accurate. Calibration develops when students repeatedly self-assess and then compare their judgments to expert feedback; they need to see how their evaluations match up with instructor standards to internalize what good work looks like. Moreover, self-assessment is more likely to be accurate when students must justify their ratings with specific evidence rather than just assigning numbers [18], and when they receive feedback on the accuracy of their self-assessments, not just on their work [6, 40].

2.2 Self-Regulated Learning (SRL)

SRL describes the dynamic processes by which students actively manage their own learning by setting goals, monitoring their progress, and adjusting their strategies and behaviors to achieve desired outcomes [34, 41]. In Zimmerman's cyclical SRL model, self-regulation involves three interrelated phases: forethought, performance, and self-reflection [42]. During *forethought*, learners analyze tasks and set goals; in the *performance* phase, they apply strategies and monitor progress; and in the *self-reflection* phase, they evaluate outcomes and adapt future approaches. These phases are dynamic and recursive—feedback from reflection shapes new goals, and the cycle continues as learners refine their strategies over time [35].

Self-assessment connects to all three phases of self-regulated learning, not just reflection [6, 36]. Clear assessment criteria help students set realistic goals and plan better strategies during forethought. Students can monitor their work more accurately during performance when they understand the standards they're working toward. Self-evaluation during reflection is the most obvious form of self-assessment, but it depends on the quality of goal-setting and monitoring that came before. Panadero and Romero [13] demonstrate that using explicit assessment criteria significantly impacts all three SRL phases. Likewise, research by Dignath and colleagues (44) shows that interventions combining planning with monitoring and evaluation are significantly more effective for enhancing strategy use and academic performance than interventions targeting these processes separately, suggesting that self-assessment works best when integrated throughout the entire learning cycle.

2.3 Applications To This Study

Engineering design studios, such as Cornell University's BME's PRIME studios [43] (see section 3.3), embody this SRL cycle in authentic, open-ended learning environments. In these studios, students work on complex, real-world problems that require iterative reasoning, creativity, and collaboration. The PRIME pedagogy scaffolds each SRL phase: students engage in analytical reasoning and goal setting as they model physiological systems; they apply design and problem-solving strategies while collaborating in teams; and they reflect on their growth using the PRISM (Proficiency Rubric for Integrated Skills Measurement) tool. PRISM requires students to assign numerical ratings to their competencies and justify those ratings in writing, while instructors use the same rubrics for group-level feedback. This parallel structure allows for the systematic study of calibration, or the alignment between students' perceived and demonstrated competence. The PRIME studio thus offers an ideal context for investigating how SRL processes, reflection, and calibration develop longitudinally in authentic engineering education settings. While self-assessment and self-regulated learning are deeply intertwined—with self-assessment serving as a mechanism that supports SRL processes across all three phases—this study focuses specifically on self-assessment practices and the calibration that develops through repeated cycles of student and instructor evaluation.

3. Methods

3.1 Institutional Context and Participants

This study took place in an upper-level undergraduate biomedical engineering course at a large, private university in the Northeastern United States. The course is open to all students who have completed introductory biology, biochemistry, and mathematics through differential equations, with priority enrollment given to BME majors. This enrollment priority reflects the course's role

as a required third-year experience for BME students, focusing on understanding cellular systems essential to BME practice. During Fall 2025, 64 students enrolled in the course, supported by two graduate teaching assistants and three undergraduate teaching assistants. Beyond three weekly 50-minute lectures, students engaged in hands-on laboratory work through three lab sessions and worked through open-ended BME problems in five three-hour design studio sessions. Laboratory and studio components were each led by dedicated instructors, providing students with specialized guidance in experimental techniques and design problem-solving.

3.2 Course Design

The course develops students' understanding of cellular systems through three modules, each paired with design studio challenges and hands on lab experiences. Module 1 examines how cells signal, grow, and propagate, with students culturing mammalian cells and exploring bioreactor applications in the lab. Module 2 investigates how cells organize into tissues and organs, focusing on different cell types and stem cell applications in tissue engineering. Module 3 explores how immune cells respond to infection and injury, with lab work examining how different stimuli trigger immune cell migration. While lecture and laboratory components of the course were assessed using traditional letter grades, all studio-related work was evaluated using proficiency-based rubrics that articulated performance expectations across multiple competency levels.

Each module's concepts were applied in studio sessions where teams tackled authentic BME challenges. In Studio 1, students designed oral drug delivery systems to disrupt cancer cell signaling. Studio 2 asked teams to engineer systems for producing red blood cells *in vitro* using stem cells. Studio 3 challenged students to model glucose regulation using organ-on-chip technology connecting pancreas, liver, and muscle tissues. Studio 4 focused on modeling T cell activation for vaccine development. Following Studios 1-4, each student completed individual reflections in which they used the PRISM proficiency rubric to assess their team's work, justify their ratings with specific evidence, and identify concrete strategies for improvement in subsequent studios. This iterative self-assessment process was designed to help students develop calibration and metacognitive awareness across the studio sequence. To support students in managing their workload, they were permitted to use one "flexible reflection" across the four studios, which allowed them to submit a shortened version focusing only on improvement strategies without completing the full self-assessment table.

Studio 5 represented a final design review experience where teams returned to one of their earlier studio challenges with added complexity requiring quantitative modeling and computational simulation. Teams selected a scenario from Studios 1-4, then presented how they applied the PRIME studio framework to address this additional design challenge. Because Studio 5 served as a final summative assessment without the accompanying self-reflection component, it was excluded from this analysis, which focuses on the iterative self-assessment practices in Studios 1-4.

3.3 PRIME BME Studio Model and PRISM Assessment Framework

The PRIME studio model integrates three-hour team-based learning sessions into core BME curriculum courses from sophomore through senior year [43]. PRIME *employs an engineering before design* framework that systematically guides student teams through five analytical steps: (P) identifying the relevant Physiological system and dysfunctions, (R) establishing Ranked

engineering constraints, (I) developing block diagrams to vIsualize system interactions, (M) applying equations to quantitatively Model system behaviors, and (E) proposing Evidence-based solutions. Studios operate as deliberately low-stakes learning environments where students receive full credit for achieving developing-level proficiency, allowing them to take intellectual risks without grade penalty. Students receive feedback mirroring professional engineering practice: real-time guidance during collaborative work, peer critique through pin-ups, and objective evaluation from expert panels [44].

To provide clear progression markers for design competency development, we developed PRISM (Proficiency Rubric for Integrated Skills Measurement) based on the Informed Design Teaching and Learning Matrix [45]. PRISM tailors these design principles for studio-based BME education, evaluating six core competencies: Understanding the BME Challenge (decomposing problems, establishing team ownership), Building Knowledge (systematic research, translating needs into measurable constraints), Qualitative Process Representation (visualizing system relationships through block diagrams), Quantitative Process Representation (mathematical modeling and optimization), Weighing Options (systematic decision-making and exploring alternatives), and Making and Communicating Engineering Judgments (metacognitive reflection, incorporating feedback, professional communication). Each competency is assessed using four proficiency levels: Emerging (superficial understanding with minimal technical vocabulary), Developing (growing independence but requiring guidance on complex applications), Proficient (systematic problem-framing and clear technical communication—our target undergraduate level), and Exemplary (exceeding expectations through leadership and innovation). Both students and instructors use PRISM to evaluate the same teamwork, with students assessing their team's performance individually and instructors providing independent evaluations, creating parallel assessments that allow for systematic comparison of student and instructor judgments. For more detailed information on the studio model and preliminary results, please refer to these prior works [43, 44, 46-48].

3.4 Data Collection

Studio worksheets (Studios 1-4) documented teams' collaborative work, including problem analysis, constraint identification and prioritization, system representations, and proposed solutions. We assessed all studio worksheets and final design review presentations using the PRISM rubric previously described. After Studios 1-4, teams received criterion-specific scores and feedback to guide their improvement. Students' post-studio reflections provided self-assessment data, where each student individually evaluated their team's work using the same PRISM rubric and justified their ratings with specific evidence. Reflections in which students used their "flexible reflection" option—which did not include the complete self-assessment table—were excluded from analysis, as these shortened submissions focused only on improvement strategies without numerical ratings. All data procedures were approved by the Cornell University Institutional Review Board under protocol IRB (0146842).

To ensure consistent application of the PRISM rubric, instructors first collaboratively scored a subset of student teams to establish shared understanding of the PRISM rubric criteria, then independently scored all teams from one complete studio session. Two instructors (S.F. and M.S) independently scored all teams ($n = 170$ team-criterion observations). We used Gwet's AC1 statistic with linear weighting for inter-rater reliability analysis because it provides more stable

estimates than Cohen's kappa when score distributions are skewed or certain categories are used infrequently [49, 50]. Results showed substantial to excellent agreement between raters (AC1 = 0.81, 95% CI [0.73, 0.90]), supporting consistent application of the PRISM rubric.

3.5 Statical Analysis

To address RQ1–RQ2, we paired each student self-assessment with the corresponding instructor rating on the 0–3 proficiency scale and summarized agreement using a confusion matrix, exact-match rates, and the distribution of disagreement magnitudes (exact, ±1, ±2, ±3). Agreement beyond chance was quantified with Gwet's AC1, computed overall and separately by competency, and re-computed by studio (Studios 1–4) to evaluate changes over time. We also quantified calibration using a difference score (student − instructor), summarized as means by studio and by competency (and grouped by PRISM studio category where applicable), with 95% Confidence Intervals (CI) used for visualization. To address RQ3, we computed the within-team standard deviation (SD) and range (max–min) for each team × competency combination using available numeric responses; team–competency observations were excluded when fewer than two numeric ratings were available (e.g., flexible reflections). Differences in within-team variation across studios and competencies were tested using Kruskal–Wallis tests (η^2 reported as the omnibus effect size). To characterize competencies with relatively higher or lower variation, we computed standardized deviations of each competency's mean within-team SD from the grand mean, scaled by the across-studio standard deviation of competency means (reported as Cohen's d) and tested these deviations for significance. Associations between within-team variation and team characteristics/outcomes (team size, mean self-reported score, and student–instructor agreement metrics) were assessed using Pearson correlations, with Spearman correlations reported as a robustness check for mean score–variation relationships. All analyses were conducted in MATLAB (two-sided $\alpha = 0.05$).

3.6 Researcher Positionality Statements

The research team in the BME Department at Cornell University includes three members with different areas of expertise in BME education and assessment. Faculty members M.S. and J.B both have extensive experience teaching engineering courses, with technical backgrounds in Chemistry and Mechanical Engineering, respectively. S.F., an Active Learning Initiative Postdoctoral Associate in BME, leads the analysis and serves as first author, bringing expertise in Biological Engineering and Engineering Education Research.

As both researchers and studio instructors, our dual role directly shaped how we approached this study of student self-assessment. Our experience providing proficiency-based feedback to students motivated us to find ways to engage students more rigorously in self-assessment. We began asking students to reflect on their studio work using the same PRISM rubric we used for evaluation. As we reviewed these reflections, we noticed discrepancies—sometimes student assessments diverged significantly from instructor evaluations, and even team members evaluating the same collaborative work assigned different scores. These patterns raised important questions about how students develop the ability to evaluate their own work accurately. We recognized that to effectively support students in developing self-assessment skills, we needed to move beyond anecdotal observations to systematic analysis. This study represents our effort to produce evidence-based insights about how students learn to assess their design work, how their judgments align with expert evaluations over time, and how we can better scaffold their self-assessment

development. We are committed to helping all students develop these critical capabilities, recognizing that the ability to evaluate one's own work is as essential for engineering practice as technical skills.

4. Results

4.1 Overall alignment between student and instructor assessments (RQ1)

Across all studios and competencies, student self-assessments and instructor ratings showed modest alignment (**Figure 1**). Among 2,183 paired ratings, 40.6% were exact matches, and overall agreement beyond chance was slight (Gwet's AC1 = 0.208). The confusion matrix (Figure 1, left) indicates that both students and instructors most frequently assigned scores of 2 (Developing) and 3 (Proficient)**.** Although exact agreement was limited, most discrepancies were small: 86% of student ratings were within ±1 proficiency level of the instructor score, and only 14% differed by two or more levels. Agreement also varied meaningfully by PRISM competency (Figure 1, right), with Gwet's AC1 ranging from 0.19 to 0.43. Most competencies (10 of 12) fell in the fair range (0.21–0.40), one reached moderate agreement **(**Creating system equations**),** and one showed poor agreement (Connecting equations to block diagram), suggesting that student–instructor alignment depends strongly on the specific design practice being assessed.

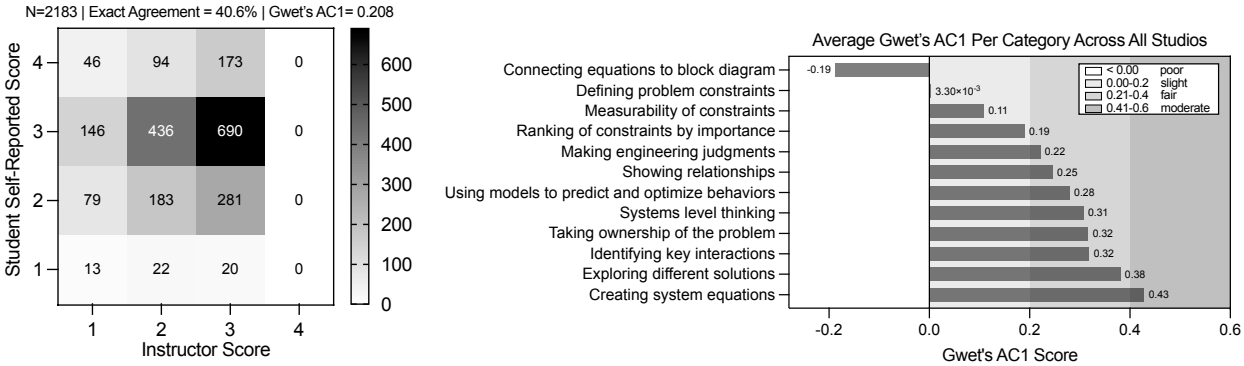

Figure 1. ***Student–instructor agreement on proficiency ratings across all studios.*** (Left) Matrix comparing student self-reported scores (rows) to instructor scores (columns) pooled across all competencies and studios (N = 2183). Cell shading and labels indicate the number of observations in each score pairing. Overall exact agreement was 40.6%, with Gwet's AC1 = 0.208. (Right) Mean Gwet's AC1 by competency across all studios; bars are labeled with AC1 values and categorized as poor (0–0.20), fair (0.21–0.40), or moderate (0.41–0.60), with lower values indicating weaker agreement.

4.2 Changes in Student–Instructor Alignment Across Studios 1-4 (RQ2)

Alignment between student and instructor ratings shifted noticeably from Studio 1 to Studio 4, with students showing an increasing tendency to rate themselves higher than instructors (**Figure 2**). The mean calibration gap (student − instructor) increased from 0.15 in Studio 1 to 0.80 in Studio 4 (**Figure 2, left**). In parallel, agreement beyond chance declined from Gwet's AC1 = 0.28 (fair) in Studio 1 to 0.09 (poor) in Studio 4 (**Figure 2, middle**). Exact matches also became less common, dropping from 20.9% to 5.3% across the same span. Despite this decline in exact agreement, most ratings remained close: the share of student scores within ±1 proficiency level of the instructor score decreased only modestly, from 67.0% in Studio 1 to 61.6% in Studio 4 (Figure 2, right). Taken together, these results indicate that while misalignment increased over time, disagreements were typically incremental rather than extreme.

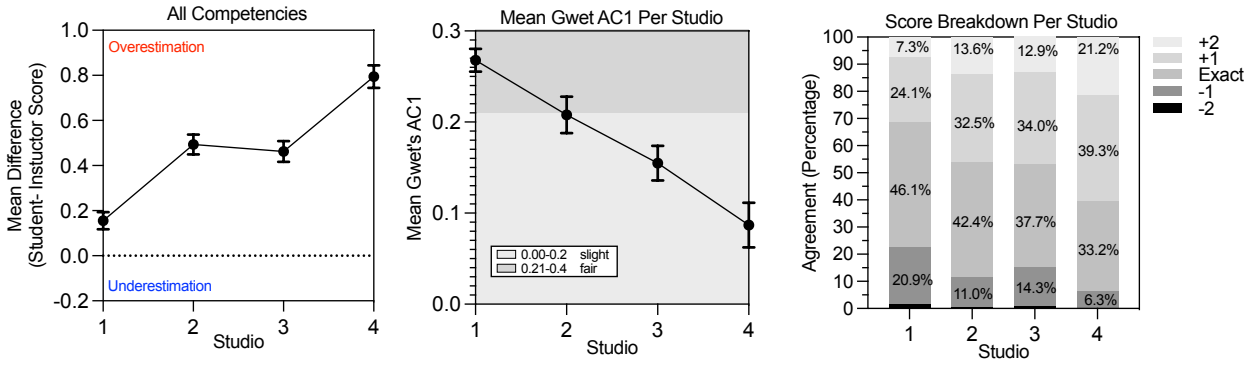

Figure 2. ***Trends in student–instructor calibration and agreement across studios (all competencies pooled).*** (Left) Mean difference between student self-reported and instructor scores (student − instructor) by studio; the dashed line at 0 indicates perfect agreement, with positive values reflecting overestimation and negative values reflecting underestimation. Points show means with 95% confidence intervals. (Middle) Mean Gwet's AC1 agreement by studio (±95% CI), with shaded bands indicating interpretive ranges (slight: 0.00–0.20; fair: 0.21–0.40). (Right) Distribution of agreement by studio showing the percentage of ratings that were an exact match and those differing by ±1 or ±2 points.

This increasing overestimation was not uniform across competency types (**Figure 3**). The most consistent upward shifts occurred in *Building Knowledge* competencies—defining problem constraints, ranking constraints by importance, and measurability of constraints—which moved from near-zero differences in Studio 1 to overestimations by Studio 4. *Qualitatively Representing Processes* showed more heterogeneous patterns: some competencies (e.g., identifying key interactions) trended upward, whereas others remained comparatively stable. *Quantitatively Representing Processes* likewise varied, with connecting equations to block diagrams showing a pronounced increase in overestimation, while creating system equations and using models to predict behaviors remained fairly aligned with instructor ratings. Overall, the data suggest that the widening student–instructor gap was driven most strongly by competencies tied to constraint analysis and connecting equations to block diagrams whereas calibration for some process-representation competencies changed less across the studio sequence. One interpretation is that as students gained experience with the studio format and PRISM language, they became more confident and more willing to rate their work highly, even as instructor ratings did not increase at the same pace—producing a growing calibration gap.

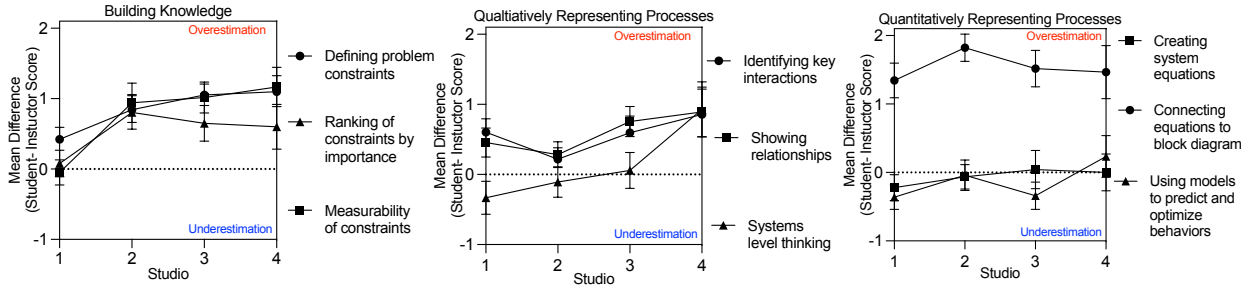

Figure 3. ***Student–instructor calibration by PRISM studio category across studios.*** Mean difference scores (student self-report − instructor score) are shown across Studios 1–4 for competencies grouped into Building Knowledge (left), Qualitatively Representing Processes (middle), and Quantitatively Representing Processes (right). Points represent studio-level means for each competency and error bars indicate 95% confidence intervals. The dotted line at 0 denotes agreement; positive values indicate student overestimation and negative values indicate underestimation.

*4.3 Within-Team Variation in Self-*Assessments *(RQ3)*

Team members evaluating the same collaborative deliverable often differed in how they rated their own proficiency (**Figure 4**). For each team–competency combination, we computed the within-team score range (max–min) on the 0–3 scale. Ranges spanned from 0 (perfect agreement) to 3 (maximum disagreement), although most team–competency cells fell between 0 and 2, indicating that disagreements were typically present but bounded. The heatmaps also highlight meaningful team-to-team differences: some teams showed consistently tight alignment across competencies (lighter cells), while others exhibited broader spreads across multiple competencies (darker cells). Blank cells indicate items for which a team range could not be computed because one or more members used the flexible reflection option (i.e., did not provide a numerical rating). These missing data were more frequent in later studios—especially Studio 4—so comparisons of within-team variation across studios should be interpreted cautiously.

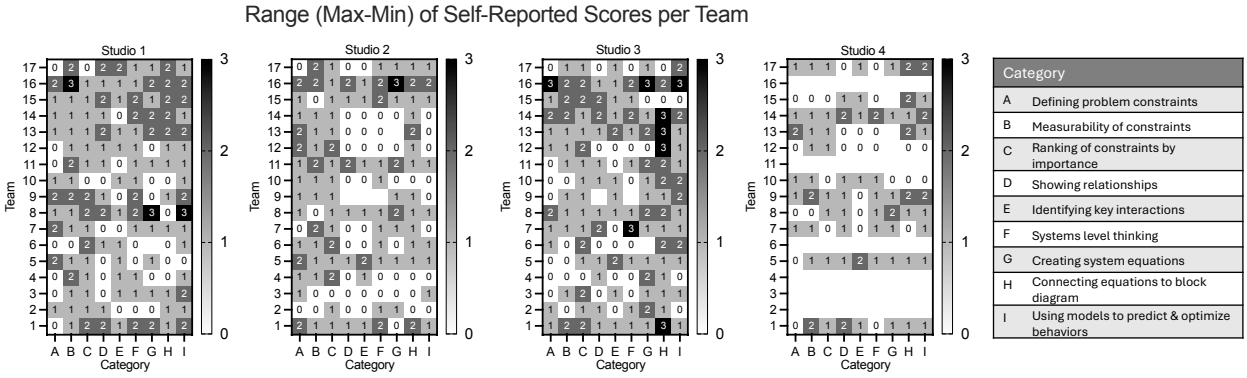

Figure 4. ***Within-team spread of self-reported proficiency scores by studio and competency.*** Heatmaps display the within-team range of self-reported scores (maximum − minimum) for each competency (Categories A–I) within Studios 1–4 for each team. Darker shading indicates a larger within-team range (greater disagreement), and cell labels report the range on the 0–3 scale. Blank cells indicate that one or more team members used the flexible reflection option (no numerical rating), so a team range could not be computed.

To summarize within-team disagreement more continuously, we next examined within-team standard deviations (SDs) (**Figure 5**). Average within-team SDs were relatively similar across Studios 1–4 (approximately 0.45–0.55), and a Kruskal–Wallis test indicated no statistically detectable difference by studio ($p = 0.097$, $\eta^2 = 0.0043$). In other words, teams' tendency to disagree when rating the same work did not systematically increase or decrease across the studio sequence. In contrast, within-team variation differed significantly by competency ($H = 41.75$, $p < 0.001$; $\eta^2 = 0.039$), indicating that some competencies elicited more consistent teammate ratings than others (**Figure 5, right**). Specifically, *Showing relationships* exhibited the lowest within-team variation (mean SD ≈ 0.39; $p = 0.0012$, Cohen's $d = -0.39$), indicating comparatively stronger consistency among teammates when rating this competency.

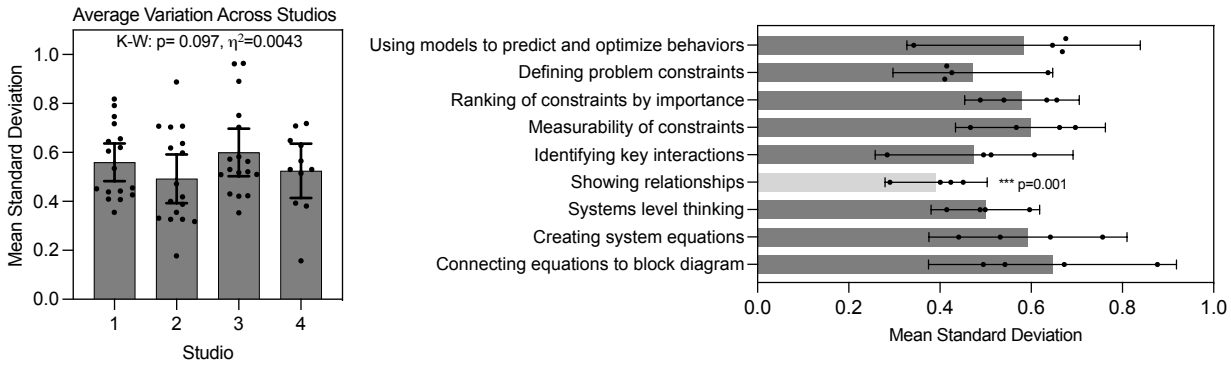

Figure 5*. **Within-team variation in competency scores across studios and proficiencies.*** Left: mean within-team standard deviation (SD) by studio (points = team values; bars/error bars = mean ± 95% Confidence Intervals (CI)). Within-team variation did not differ significantly across studios (Kruskal–Wallis H = 6.32, $p = 0.097$; $\eta^2 = 0.0043$). Right: mean within-team SD by competency across all studios (points = team values; bars/error bars = mean ± 95% CI). Within-team variation differed significantly across competencies (Kruskal–Wallis H = 41.75, $p < 0.001$; $\eta^2 = 0.039$). Teams showed significantly lower within-team variation on Showing relationships (i.e., team members' self-ratings were more consistent/aligned) relative to the overall baseline used in this analysis (p = 0.0012, Cohen's d = −0.39).

Finally, we examined whether within-team variation was associated with team size or with student–instructor agreement quality (**Figure 6**). Team size showed no relationship with within-team variation (Pearson $r = 0.044$, $p = 0.22$), indicating that adding members did not systematically increase disagreement. In addition to team size, we tested whether within-team variation was related to the *level* of teams' self-assessed performance. Across team–competency observations, within-team variation was negatively associated with mean self-reported score, indicating that teams reporting higher average proficiency tended to show greater consistency in their self-ratings (Pearson $r = -0.1695$, $p = 0.000002$). This pattern was also supported using a rank-based correlation (Spearman $\rho = -0.1253$, $p = 0.000452$), suggesting the relationship is not driven solely by distributional assumptions. Finally, competencies with higher within-team variation were not necessarily those with poorer student–instructor alignment: the association between within-team variation (Avg SD) and exact-match agreement (%) across competencies was weak and non-significant (Pearson $r = -0.18$, $p = 0.57$). Together, these findings suggest that improving student–instructor calibration may not automatically reduce within-team disagreement, and vice versa.

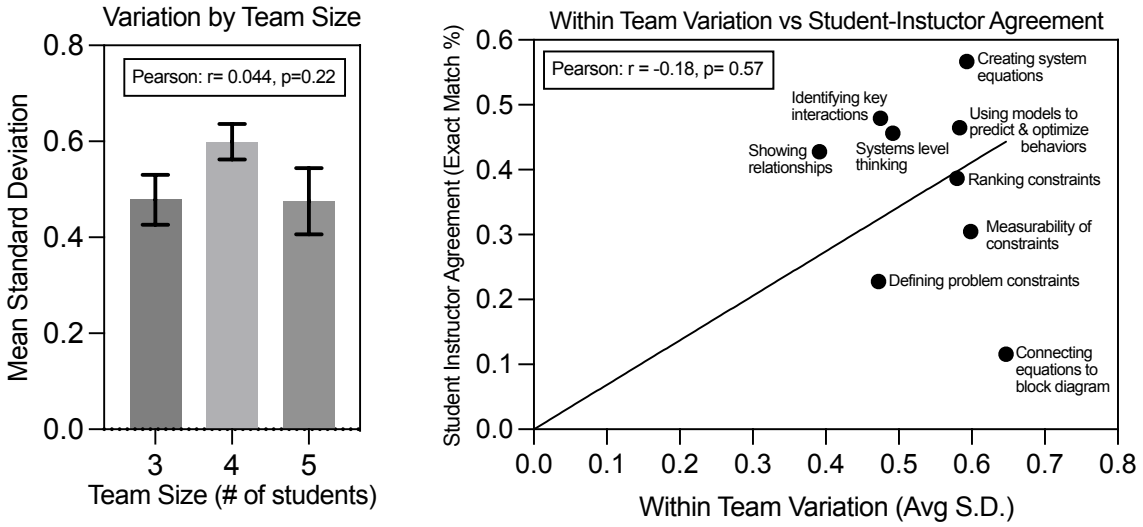

Figure 6. ***Relationships between within-team variation, team size, and student–instructor agreement.*** (Left) Mean within-team variation (average within-team standard deviation) by team size (3–5 students); bars show means with 95% confidence intervals. Variation was not significantly correlated with team size (Pearson $r = 0.044$, $p = 0.22$). (Right) Association between within-team variation (Avg S.D.) and student–instructor agreement (exact match %) across competencies; each point represents a competency and the line shows the linear fit. No significant relationship was observed (Pearson $r = -0.18$, $p = 0.57$).

5. Discussion

The results paint an encouraging picture of students actively engaging with a proficiency-based language for design—even if their judgments do not fully converge with instructors' standards over time. Across all studios, most student ratings stayed *close* to instructor evaluations (the large majority within ±1 level), suggesting that students generally recognized the intended performance bands and were not making wildly inconsistent claims. At the same time, the calibration gap widened from Studio 1 to Studio 4, driven by an increasing tendency for students to rate their team's work higher than instructors did. Rather than viewing this solely as "inaccuracy," we interpret it as a signal that experience alone does not automatically produce calibration: repeated exposure to the PRISM framework and studio workflow may build confidence and fluency in the rubric language, but students may still be forming their own internal standards for what "proficient" looks like—standards that drift from (or lag behind) instructor expectations as tasks grow more complex. In other words, students appear to become *more willing* to make strong claims about their competence, but without structured calibration opportunities, that confidence does not translate into better alignment. This supports prior self-assessment work emphasizing that calibration develops through deliberate comparison to expert judgments and feedback on the accuracy of the self-assessment itself, not just feedback on the work product.

Within teams, we also see a productive—if imperfect—signal: teammates evaluating the same shared deliverable often disagreed, and the amount of disagreement depended on the competency being rated. That variation suggests that PRISM is functioning as a meaningful interpretive tool rather than a "fill-in-the-blank" checklist: different students attend to different evidence, prioritize different aspects of the deliverable, or interpret the rubric levels differently—especially for competencies that require translating qualitative work into quantitative judgments. Importantly,

some competencies elicited relatively tighter within-team alignment, indicating that teams *can* converge when the evidence is more visible and the performance cues are easier to interpret. The finding that within-team variation was largely stable across studios and only weakly related to team size reinforces the idea that shared understanding is not a simple function of time or group composition; instead, it likely depends on the clarity of evidence, how teams negotiate meaning around the rubric, and the extent to which students are coached to justify ratings with concrete artifacts. The small but reliable negative relationship between mean self-reported score and within-team variation further suggests a positive pathway: when teams perceive their work as stronger (and perhaps produce clearer evidence), teammates' interpretations converge. Taken together, these findings position PRISM-based reflection as a valuable diagnostic: it not only reveals where students diverge from instructor standards, but also where teammates diverge from one another—highlighting specific competencies where shared meaning-making around "good design work" may still be developing.

These findings also suggest several design implications for instruction. Calibration should be treated as an explicit learning objective rather than an assumed byproduct of repetition; students need structured chances to compare their self-ratings to instructor or exemplar ratings and to discuss why discrepancies occur. Within-team disagreement, similarly, can be leveraged as a learning resource: brief, structured conversations about divergent ratings and the evidence supporting them may help students surface assumptions and build shared standards for quality work. Finally, competencies that consistently elicited greater disagreement or misalignment may benefit from targeted scaffolds—such as annotated exemplars, clearer "search-points" tied to deliverable features, and prompts that require students to cite specific artifacts when assigning a rating. Taken together, these strategies position self-assessment not as a solitary reflective act, but as a coached, evidence-based, and socially mediated practice embedded within the design process.

6. Limitations and Future Directions

This study's retrospective design limits our ability to draw causal conclusions about how self-assessment can be improved. Because we analyzed ratings and scores only after studios were completed—and did not introduce calibration activities or other interventions between studios—we can describe how alignment and within-team variation emerged under typical course conditions, but not determine which instructional supports would most effectively enhance self-assessment accuracy or whether targeted interventions would meaningfully alter these patterns.

A second limitation concerns our emphasis on numerical ratings rather than on the quality of students' reflective reasoning. Although students were required to justify their ratings with specific evidence, we did not systematically evaluate the depth, specificity, or evidentiary strength of these written justifications. As a result, students may appear mis-calibrated numerically while engaging in thoughtful, criteria-based reasoning, or conversely, numerically aligned with instructor scores while offering superficial or weakly supported explanations. Without analyzing justification quality, we cannot distinguish among qualitatively different forms of calibration—for example, students who are numerically misaligned but conceptually sophisticated versus those who are numerically aligned but poorly reasoned. Finally, missing data from students who opted for the flexible reflection option reduced the amount of comparable data available in later studios, particularly Studio 4. This attrition limits our ability to draw strong conclusions about trends in

within-team variation across the studio sequence and may obscure patterns that would be more visible in a complete dataset.

Future work will address these limitations by integrating quantitative analyses of rating patterns with systematic evaluation of justification quality. By coding students' written justifications for evidentiary strength, specificity, and alignment with PRISM criteria, we can develop a more nuanced account of self-assessment development that captures both rating accuracy and reasoning depth. This integrated approach would allow us to identify distinct support needs: some students may require assistance calibrating numerical judgments, while others may need scaffolding to articulate evidence-based reasoning more effectively.

With larger, multi-semester datasets, we plan to identify clusters of self-assessment profiles and examine whether these patterns differ by student demographics or prior experiences. Understanding how calibration and reasoning quality vary across student populations would inform the design of more equitable and personalized instructional supports. Building on these analyses, we aim to develop real-time learning analytics that provide students with actionable feedback on both their ratings and their reasoning. Such a system could indicate whether a student's self-assessment demonstrates strong evidence use, clear connections to PRISM criteria, and realistic goal-setting, and could be paired with targeted scaffolds such as exemplar-based calibration exercises, structured evidence prompts, and team-level norming discussions prior to individual reflection.

Ultimately, our goal is to move from retrospective measurement toward instructional interventions that actively support the development of self-assessment. We envision a learning environment in which students not only assign ratings to their work but learn to ground those judgments in evidence, recognize gaps between current performance and target standards, and translate those insights into concrete improvement strategies. By understanding how students currently engage in self-assessment—across both their successes and their challenges—we can design instruction that cultivates the metacognitive capacities essential for independent professional practice.

7. Acknowledgements

This work was supported by the Cornell University Active Learning Initiatives and the NSF EF-2222434 grant. The authors would like to thank the students for their participation in the reflection experiences and for sharing their thoughtful insights with the instructional team.

8. References

[1] G. C. Fleming, M. Klopfer, A. Katz, and D. Knight, "What engineering employers want: An analysis of technical and professional skills in engineering job advertisements," *Journal of Engineering Education,* vol. 113, no. 2, pp. 251-279, 2024.

[2] C. J. Atman, O. Eris, J. McDonnell, M. E. Cardella, and J. L. Borgford-Parnell, "Engineering design education," *Cambridge handbook of engineering education research,* pp. 201-225, 2014.

[3] C. J. Atman, O. Eris, J. McDonnell, M. E. Cardella, and J. L. Borgford-Parnell, "Engineering Design Education: Research, Practice, and Examples that Link the Two," in

Cambridge Handbook of Engineering Education Research, A. Johri and B. M. Olds Eds. Cambridge: Cambridge University Press, 2014, pp. 201-226.

[4] C. L. Dym, A. M. Agogino, O. Eris, D. D. Frey, and L. J. Leifer, "Engineering design thinking, teaching, and learning," *Journal of engineering education,* vol. 94, no. 1, pp. 103-120, 2005.

[5] Z. Yan and D. Carless, "Self-assessment is about more than self: the enabling role of feedback literacy," *Assessment & Evaluation in Higher Education,* vol. 47, no. 7, pp. 1116-1128, 2022.

[6] H. L. Andrade, "Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning," in *Handbook of formative assessment*: Routledge, 2010, pp. 90-105.

[7] K. J. Åström and R. Murray, *Feedback systems: an introduction for scientists and engineers*. Princeton university press, 2021.

[8] I. Mohedas, K. H. Sienko, S. R. Daly, and G. L. Cravens, "Students' perceptions of the value of stakeholder engagement during engineering design," *Journal of Engineering Education,* vol. 109, no. 4, pp. 760-779, 2020.

[9] J. H. Nieminen and D. Carless, "Feedback literacy: A critical review of an emerging concept," *Higher Education,* vol. 85, no. 6, pp. 1381-1400, 2023.

[10] H. L. Andrade, "A critical review of research on student self-assessment," in *Frontiers in education*, 2019, vol. 4: Frontiers Media SA, p. 87.

[11] G. T. Brown and L. R. Harris, "Student Self-Assessment 1," in *SAGE handbook of research on classroom assessment*: SAGE Publications, Inc., 2013, pp. 367-393.

[12] K. Topping, "Self and peer assessment in school and university: Reliability, validity and utility," in *Optimising new modes of assessment: In search of qualities and standards*: Springer, 2003, pp. 55-87.

[13] E. Panadero and M. Romero, "To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy," *Assessment in Education: Principles, Policy & Practice,* vol. 21, no. 2, pp. 133-148, 2014.

[14] Y. M. Reddy and H. Andrade, "A review of rubric use in higher education," *Assessment & evaluation in higher education,* vol. 35, no. 4, pp. 435-448, 2010.

[15] J. Hafner and P. Hafner, "Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer-group rating," *Int. J. Sci. Educ.,* vol. 25, no. 12, pp. 1509-1528, 2003.

[16] E. Panadero, J. A. Tapia, and J. A. Huertas, "Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education," *Learning and individual differences,* vol. 22, no. 6, pp. 806-813, 2012.

[17] H. L. Andrade, Y. Du, and K. Mycek, "Rubric-referenced self-assessment and middle school students' writing," *Assessment in Education: Principles, Policy & Practice,* vol. 17, no. 2, pp. 199-214, 2010.

[18] H. L. Andrade, Y. Du, and X. Wang, "Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing," *Educational Measurement: Issues and Practice,* vol. 27, no. 2, pp. 3-13, 2008.

[19] H. G. Andrade, "The effects of instructional rubrics on learning to write," *Current issues in education,* vol. 4, 2001.

[20] M. J. McCormick, K. E. Dooley, J. R. Lindner, and R. L. Cummins, "Perceived Growth versus Actual Growth in Executive Leadership Competencies: An Application of the

Stair-Step Behaviorally Anchored Evaluation Approach," *Journal of Agricultural Education,* vol. 48, no. 2, pp. 23-35, 2007.
[21] E. Panadero and A. Jonsson, "The use of scoring rubrics for formative assessment purposes revisited: A review," *Educational research review,* vol. 9, pp. 129-144, 2013.
[22] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," *Educational research review,* vol. 2, no. 2, pp. 130-144, 2007.
[23] H. L. Andrade, X. Wang, Y. Du, and R. L. Akawi, "Rubric-referenced self-assessment and self-efficacy for writing," *The Journal of Educational Research,* vol. 102, no. 4, pp. 287-302, 2009.
[24] M. Van Zundert, D. Sluijsmans, and J. Van Merriënboer, "Effective peer assessment processes: Research findings and future directions," *Learning and instruction,* vol. 20, no. 4, pp. 270-279, 2010.
[25] K. Topping, "Peer assessment between students in colleges and universities," *Review of educational Research,* vol. 68, no. 3, pp. 249-276, 1998.
[26] J. Trevelyan, "Incremental self-assessment rubrics for capstone design courses," in *2015 ASEE Annual Conference & Exposition*, 2015, pp. 26.951. 1-26.951. 17.
[27] T. El-Maaddawy, "Enhancing learning of engineering students through self-assessment," in *2017 IEEE Global Engineering Education Conference (EDUCON)*, 2017: IEEE, pp. 86-91.
[28] L. W. Lackey, R. Mines, and H. Jenkins, "Engineering Student Self-assessment in a Capstone Design Course," in *ASEE Annual Conference and Exposition*, 2006.
[29] F. C. Berry, W. Huang, and M. Exter, "Improving accuracy of self-and-peer assessment in engineering technology capstone," *IEEE Transactions on Education,* vol. 66, no. 2, pp. 174-187, 2022.
[30] F. Mattioli and S. D. Ferraris, "Self-evaluation and peer evaluation tool for design and engineering teams: experiment conducted in a design studio," in *DS 95: Proceedings of the 21st International Conference on Engineering and Product Design Education (E&PDE 2019), University of Strathclyde, Glasgow. 12th-13th September 2019*, 2019.
[31] A. Pagano, T. T. Parks, and S. Shehab, "Developing a Human-Centered Engineering Design Self-Assessment Survey," in *2024 ASEE Annual Conference & Exposition*, 2024.
[32] K. A. Douglas, R. E. Wertz, M. Fosmire, S. Purzer, and A. S. Van Epps, "First-year and junior engineering students' self-assessment of information literacy skills," in *2014 ASEE Annual Conference & Exposition*, 2014, pp. 24.607. 1-24.607. 10.
[33] B. Jaeger, S. Freeman, and R. Whalen, "Do they Like What They Learn, Do They Learn What They Like–And What Do We Do About It?," in *2007 Annual Conference & Exposition*, 2007, pp. 12.560. 1-12.560. 23.
[34] P. R. Pintrich, "The role of goal orientation in self-regulated learning," in *Handbook of self-regulation*: Elsevier, 2000, pp. 451-502.
[35] B. J. Zimmerman, "Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects," *American educational research journal,* vol. 45, no. 1, pp. 166-183, 2008.
[36] E. Panadero and J. Alonso-Tapia, "Self-assessment: theoretical and practical connotations, when it happens, how is it acquired and what to do to develop it in our students," 2013.

[37] E. Panadero, G. T. Brown, and J.-W. Strijbos, "The future of student self-assessment: A review of known unknowns and potential directions," *Educational psychology review,* vol. 28, no. 4, pp. 803-830, 2016.
[38] H. W. Goodrich, "Student self-assessment: At the intersection of metacognition and authentic assessment," Harvard University, 1996.
[39] J. Hattie and H. Timperley, "The Power of Feedback," *Review of Educational Research,* vol. 77, no. 1, pp. 81-112, 2007, doi: 10.3102/003465430298487.
[40] T. Sitzmann, K. Ely, K. G. Brown, and K. N. Bauer, "Self-assessment of knowledge: A cognitive learning or affective measure?," *Academy of Management Learning & Education,* vol. 9, no. 2, pp. 169-191, 2010.
[41] P. H. Winne, "Paradigmatic dimensions of instrumentation and analytic methods in research on self-regulated learning," *Computers in Human Behavior,* vol. 96, pp. 285-289, 2019.
[42] B. J. Zimmerman, "Attaining self-regulation: A social cognitive perspective," in *Handbook of self-regulation*: Elsevier, 2000, pp. 13-39.
[43] S. Fuchs, V. Vasudevan, and J. Butcher, "Embedding Studio-Based Learning in the Biomedical Engineering Curriculum to Improve Quantitative Problem-Solving Skills," *Biomedical Engineering Education,* 2025/09/15 2025, doi: 10.1007/s43683-025-00195-5.
[44] S. Fuchs, A. Werth, and J. Butcher, "Work in Progress: Evaluating the impact of student cognitive and emotional responses to real-time feedback on student engagement in engineering design studios," presented at the American Society for Engineering Education Annual Conference & Exposition, Portland, Oregon, 2024. [Online]. Available: https://nemo.asee.org/public/conferences/344/papers/42383/view.
[45] D. P. Crismond and R. S. Adams, "The informed design teaching & learning matrix," *Journal of Engineering Education-Washington,* vol. 101, no. 4, p. 738, 2012.
[46] S. Fuchs, M. Saikia, and J. Butcher, "Work In Progress: A framework for evaluating student cognitive and affective reflections in BME studio learning," presented at the 2025 ASEE Annual Conference and Exposition Montreal, Canada, 2025.
[47] M. Saikia and S. Fuchs, "WIP: Can studio-style instruction promote the application of engineering principles in biomedical problem solving. Analysis of type 1 diabetes treatment designs submitted by biomedical engineering students in their sophomore and junior year studio," in *2025 Rocky Mountain Section Conference*, 2025.
[48] V. S. Vasudevan, S. M. Fuchs, and J. Butcher, "Engineering Studio Pedagogy–First Experience in Integrating Novel Studio Sessions in Biomedical Engineering Courses," in *2024 IEEE Frontiers in Education Conference (FIE)*, 2024: IEEE, pp. 1-9.
[49] K. L. Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," *British Journal of Mathematical and Statistical Psychology,* vol. 61, no. 1, pp. 29-48, 2008.
[50] N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet, "A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples," *BMC Medical Research Methodology,* vol. 13, no. 1, p. 61, 2013/04/29 2013, doi: 10.1186/1471-2288-13-61.