2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC

Host Sponsor UNC Charlotte

Paper ID #51105

Promoting Equitable Accountability in STEM COIL through Multi-Source Peer Assessment

Dr. Dita Puspita Sari, Innovative Global Program, Shibaura Institute of Technology

Dita Puspita Sari is an Associate Professor in the Innovative Global Program at Shibaura Institute of Technology, Japan, and a Visiting Scientist at the RIKEN Nishina Center. She received her Ph.D. in Physics from Osaka University in 2018. Her research focuses on strongly correlated electron systems, particularly organic superconductors, using muon spin rotation (SR) and complementary experimental and computational approaches. In addition to condensed matter physics, she is actively engaged in interdisciplinary STEM education research, including Collaborative Online International Learning (COIL) and data-driven assessment methods. She has experience collaborating with international research facilities such as RIKEN in Japan, ISIS in the United Kingdom, Paul Scherrer Institut (PSI) in Switzerland, and TRIUMF in Canada, and her work spans both fundamental physics and global engineering education.

Prof. Darminto, Institut Teknologi Sepuluh Nopember

Darminto graduated from the doctoral program in Physics at the Bandung Institute of Technology, Indonesia, in 2001, and continued with a post-doctoral research program at Twente University, The Netherlands, from 2001 to 2003. Upon returning, he became a lecturer in the Physics Department at ITS and was appointed as a professor of condensed matter physics in 2011. He served as the university's vice president for research and partnership (2011-2015), the director of the Institute of Research and Community Engagement (2011-2015), and the executive secretary of the Board of Trustees (2016-2021). Previously, he participated in research internship programs at the Vienna University of Technology (1997), the University of Amsterdam (1998), and the University of Tokyo (1999). He has collaborated with RIKEN Nishina Center since 2007, the Federal Institute for Materials Research and Testing (BAM) in Berlin, Germany, since 2013, and the Synchrotron Light Research Institute in Thailand since 2016. Additionally, he has networked with numerous universities in Japan, Taiwan, Southeast Asian countries, and Germany. These collaborations have been financially supported by national and international research grants. He has supervised hundreds of undergraduate, master's, and doctoral students, resulting in approximately 500 publications, including articles, posters, books, chapters, and scientific and academic reports. His research interests lie in condensed matter physics, encompassing superconductivity, magneto-electronics, nanomaterials, and 2D materials. He is a member of the Indonesian Physical Society, the Materials Research Society of Indonesia, the American Physical Society, the American Association for the Advancement of Science, the International Association of Advanced Materials, and the International Solar Energy Society. He also served as the President of the Indonesian Magnetics Society (2018-2020).

Gabriele Trovato

Gabriele Trovato is currently Associate Professor in Shibaura Institute of Technology, where he leads LAB 22, as well as Senior Visiting Fellow at the University of Palermo, Italy, International Co-Investigator at the National Research Foundation of Korea (NRF), and Visiting Researcher in Waseda University, Tokyo, Japan.

He received his M.S. degree in Computer Engineering from the University of Pisa, Italy, and Ph.D. degree in Biorobotics in Waseda University. Within the relations between the two countries, Gabriele Trovato has been in the organising committee of Italy-Japan Workshops since 2011, and has been appointed "Ambassador of Livorno in the world" by the Municipality of Livorno, Italy.

He has been Visiting Researcher in Karlsruhe Institute of Technology (Germany), PUCP (Peru) and Imperial College London (UK) among others.

Gabriele Trovato has worked in the video game industry, being involved in the development of the world-wide known game series "Sid Meier's Civilization" and having created popular innovative mods for the game.

His main research interests are interdisciplinary and include Human-Robot Interaction, with focus on culture and religion related aspects, artificial emotions in humanoids, and robot aesthetics.

Gabriele Trovato's creations, such as SanTO robots, are a combination of engineering, AI, art and humanities, and raised interest among the worldwide press, including the Wall Street Journal and the BBC.

Sumito Nagasawa

Dr. Sumito Nagasawa received Ph.D. in Engineering from the University of Tokyo in 2001. He is a Professor in Adv. Mech. Eng. Course, Mech. Eng. Program, College of Engineering in Shibaura Institute of Technology. His main research topic is the development and application of small robots using Micro-Electro-Mechanical Systems technologies (MEMS) technologies. In the field of engineering education, he focuses on experiential learning using robotics technology, including learning support based on embedded systems and sensor measurement, as well as the development of practical educational methods for PBL-based courses and international collaborative programs. He also conducts research on methods for evaluating learning processes and educational outcomes in group work.

Eiji Kamioka

Eiji Kamioka (Member, IEEE) is currently a Professor and a Deputy President at the Shibaura Institute of Technology (SIT). He also leads the initiative of implementing COILs at SIT. Before joining SIT, he was with the Sharp Communication Laboratory, the Institute of Space and Astronautical Science (ISAS) as a Research Fellow of the Japan Society for the Promotion of Science (JSPS), and the National Institute of Informatics (NII) as an Assistant Professor. His research interests encompass user-friendly information and communication systems, such as mobile multimedia communication and intelligent information systems. He received his B.S., M.S., and Ph.D. degrees in physics from Aoyama Gakuin University, Japan.

Prof. Hatsuko Yoshikubo Ph.D., Shibaura Institute of Technology

Dr. Hatsuko Yoshikubo is a Professor of Humanities in the Innovative Global Program at the College of Engineering, Shibaura Institute of Technology, a leading engineering university in Tokyo, Japan, where she leads the Smart Education for Creative Engineers Laboratory (SECE Lab). She earned her Ph.D. in English Literature from Chiba University in 2002. Her research explores the value of incorporating humanities into science and engineering curricula, innovative assessment strategies in hybrid and international learning environments, and collaborative learning and assessment in global and COIL-based engineering education contexts, with a recent focus on AI-enhanced learning, data-driven assessment, and interdisciplinary engineering education design. She currently leads a research project funded by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (No. 24K06133, 2024–2027). She has presented her work at major international conferences including ASEE, WERA, BERA, IIAI-DSIR, and APAIE. Her work has been recognized with multiple awards, including the ASEE Best Diversity Paper Award (2023), the DSIR Outstanding Paper Award (2021), SIT's Presidential Award (2023), and recent recognitions including the 20th Kanto Society for Engineering Education Award (Best Paper/Essay Award) and the JSEE Annual Conference International Session Award (2025).

Promoting Equitable Accountability in STEM COIL through Multi-Source Peer Assessment

Abstract

Group-based learning in engineering education fosters technical competence and teamwork skills; however, assessment practices often struggle to ensure individual accountability, leading to grading bias and uneven participation. This study implements a multi-source peer assessment model in a Collaborative Online International Learning (COIL) program jointly conducted by Shibaura Institute of Technology (Japan) and Institut Teknologi Sepuluh Nopember (Indonesia). The whole 14-week hybrid online–onsite physics course (Magnetism and Magnetic Materials) was embedded within an interdisciplinary COIL module that integrated STEM content. Additionally, a two-week collaboration with a humanities module (History of Japan) broadened the notion of "contribution" by highlighting sociocultural and ethical dimensions of teamwork. To promote fairness, a weighted scoring model was developed, integrating within-group peer evaluations (20%), cross-group peer evaluations (10%), Indonesia-based instructor evaluations (30%), and Japan-based physics lead instructor evaluations (40%). Fairness is conceptualized as both procedural and representational, capturing transparency in evaluation and the visibility of individual contributions. The model was compared with a conventional instructor-only assessment approach using data from two cohorts (FY2024 and 2025). Instructor-only averages exceeded peer-inclusive adjusted scores by approximately 2–5 points across cohorts, suggesting grade inflation under single-source evaluation. Statistical analysis showed significant differences ($p < 0.01$) across both academic years, while maintaining strong correlation ($r > 0.87$) and rank-order stability ($\rho > 0.82$). These findings indicate meaningful score correction without substantially altering performance ranking. The proposed multi-source model reduces reliance on a single evaluator, enhances transparency of individual contributions in collaborative contexts, and offers an empirically validated assessment framework for interdisciplinary, cross-institutional STEM COIL environments—an area that remains underexplored in engineering education research.

Keywords: Peer assessment, COIL (Collaborative Online International Learning), multi-source evaluation, engineering education, grading bias, individual accountability, international collaboration

1. Introduction

Group-based learning has become a cornerstone of engineering and physics education because it fosters essential competencies such as collaboration and communication [1]–[3]. However, assessing individual contributions within collaborative environments remains difficult. Instructor-only evaluation can result in inflated grades for low contributors or insufficient recognition for highly active team members [2], [4]. Peer assessment has been adopted to increase fairness by capturing teamwork processes not visible to instructors [5]–[7], and it can improve learning outcomes for students with varying achievement levels [8]. Still, students often trust instructor feedback more than peer ratings due to concerns about credibility and bias [9]. Peer scores may be affected by social relationships, friendship norms, or reluctance to provide negative feedback [2], [4]. These challenges become more complex in cross-institutional and culturally diverse COIL environments, where communication norms differ and accountability is difficult to maintain across distributed work [1], [2], [10]. Thus, there is a growing need for assessment models that combine multiple viewpoints to strengthen validity, transparency, and fairness in collaborative STEM learning [6], [11]. To situate this study within a broader interdisciplinary context, the COIL course examined here integrates both STEM and humanities perspectives, including a two-week collaboration with a History of Japan module, expanding the notion of contribution beyond technical tasks to include sociocultural, collaborative, and ethical dimensions. In this context, peer assessment becomes particularly valuable, as it provides insight into internal team roles, communication, and effort distribution [5]–[7]. Evidence from online peer assessment (OPA) indicates significant positive effects on academic performance [8], [12] and higher-order thinking [13], and OPA increases scalability and feedback immediacy in hybrid settings [7], [12], [13]. Despite these benefits, peer assessments can be unreliable due to interpersonal dynamics and bias [2], [4], and their trustworthiness may be questioned by students [9]. These issues are intensified in culturally mixed teams, where norms of critique and communication differ [2], [4], [10]. Physics education research emphasizes composite frameworks that triangulate evaluation sources to maximize validity [11], [12], [14], [15]. Similar triangulation principles have been discussed in broader assessment and fairness literature, particularly in relation to validity, transparency, and accountability in group-based and cross-cultural learning environments. Thus, structured multi-source assessment is an appropriate method to address both fairness and accountability in global teamwork environments.

The present study introduces a multi-source peer-assessment model integrating instructor scoring, within-group peer ratings, and cross-group peer ratings into a single composite score. This model is evaluated in a hybrid, cross-cultural COIL engineering-physics course to determine whether it reduces grade inflation and better reflects individual contribution.
Research questions:

1. Does the proposed multi-source assessment model significantly reduce score inflation compared to instructor-only scoring?
2. Does the model preserve alignment with instructor judgment, as reflected in linear correlation and rank-order stability?
3. How do peer and instructor evaluations differ across academic cohorts, and what do these differences imply for the feasibility of implementing multi-source assessment in COIL environments?

Beyond its technical objectives, the course design was informed by humanities perspectives on fairness, ethics, and social accountability in collaborative work. Students were encouraged to reflect on invisible labour, contribution asymmetry, and responsibility within multinational teams—issues commonly examined in humanities and social science scholarship. These perspectives directly shaped the assessment design by motivating the use of peer-based evaluation as a mechanism to surface contribution patterns that are not accessible through product-based evaluation alone. In this study, "fairness" is conceptualized not only as statistical adjustment of scores, but also as procedural fairness and representational fairness. Procedural fairness refers to the transparency and consistency of the evaluation process, while representational fairness refers to the extent to which individual contributions—particularly those not directly observable by instructors—are appropriately reflected in assessment outcomes.

2. Methodology

2.1 Course and intervention design

The study was embedded in an undergraduate "Magnetism and Magnetic Materials" COIL course jointly delivered by Shibaura Institute of Technology (SIT, Japan) and Institut Teknologi Sepuluh Nopember (ITS, Indonesia). The hybrid learning environment combined physical lab work with synchronous Zoom discussions and asynchronous collaboration through Slack. A total of 51 students participated from SIT and ITS, organized into mixed-nationality teams of 4–6 members. There were 4 Japanese, 38 Indonesian (online), 2 Brazilian, 1 Korean, 1 Chinese, 1 Finnish, 1 Russian, and 3 Dutch. Teams were formed to ensure a mix of countries, gender, and research interests. Each team completed the discussion and oral presentation based on reviewing a research paper addressing a magnetic materials topic, with distributed responsibilities including reviewing the background study, experimental and computational methods, analysis results, and the conclusion part. The research papers were chosen from the part of the research done by the Japan-based instructor and collaborators and can represent the whole syllabus of the course. Three structured peer-assessment events were embedded in the project timeline to measure evolving collaborative contributions and social presence. Because teamwork was partially remote, direct instructor visibility into internal task distribution was limited — making this setting appropriate to test multi-source assessment. The 14-week module was conducted as a COIL course between SIT and ITS. A separate two-week humanities collaboration, involving approximately 40 students, was not included in the peer assessment model due to its limited proportion within the overall course. Nevertheless, this period contributed to building interaction and collaborative dynamics among students through discussions of scientific developments in their respective countries.

2.2 Interdisciplinary course design: Two-week Physics–Humanities collaboration

COIL programs provide opportunities for students to collaborate across cultural and disciplinary contexts. Prior research in engineering education has shown that integrating humanities perspectives into STEM courses supports the development of sociotechnical thinking, contextual awareness, and interdisciplinary communication essential for engineering practice [17],[18]. Consistent with these findings, prior COIL implementations also indicated that interdisciplinary collaboration encouraged diverse forms of student contribution beyond technical task execution [19]. These considerations informed the decision to incorporate a two-week collaboration with a humanities module, History of Japan, into the hybrid COIL physics module (Magnetism and Magnetic Materials). This interdisciplinary integration influenced how students contributed to group work. In addition to solving physics problems, students contributed by interpreting historical context, facilitating discussion, connecting ideas across disciplines, and supporting team coordination. Many of these contributions emerged during group interaction and were not always visible in final presentations or other instructor-observed outputs. As a result, an instructor-only assessment may not fully reflect individual participation in collaborative processes. To implement this collaboration, the humanities module was integrated within the same course structure and group organization as the physics module. Student teams engaged in both onsite and online interactions, enabling interdisciplinary contributions to emerge through multiple forms of participation. Because peer evaluations capture interaction within groups, they provide information about contribution patterns that may not be directly observable to instructors. This course design supported the use of a multi-source assessment model that incorporated both peer and instructor perspectives. During the two-week collaboration, student groups were encouraged to relate physics concepts, such as the properties and applications of magnetic materials, to historical developments and socio-cultural contexts introduced in the humanities module. This interdisciplinary framing expanded the scope of contribution in student team projects and formed the basis for the assessment model described in Section 2.3. Although the humanities module was not included as a separate scoring component in the multi-source assessment model, it broadened the forms of contribution that emerged during teamwork, including sociocultural interpretation, facilitation, and coordination. These expanded contribution patterns provided additional justification for incorporating peer assessments, which capture internal group processes not directly observable by instructors.

2.3 Multi-source peer-assessment model

Instructor scoring has strong validity for evaluating technical outcomes and presentation quality, while peer scoring captures internal work dynamics and behavioural contributions not visible externally [4]–[6]. Within-group peers observe collaboration processes continuously and can therefore provide direct evidence of individual contribution. Cross-group peers provide a limited external perspective, primarily based on presentations, on clarity, communication, and scientific reasoning. Such triangulation improves both fairness and credibility in collaborative assessment [6], [8], [11]. This need for triangulation becomes more salient in interdisciplinary COIL settings, where meaningful contributions may include facilitation and contextual interpretation in addition to technical problem-solving. The final individual score for student *i* is defined as having an adjusted score S_i:

$$S_i = 0.2 \times W_i + 0.1 \times C_i + 0.3 \times I_{INA,i} + 0.4 \times I_{JPN,i}, \tag{1}$$

where:

- W_i : within-group peer assessment
- C_i : cross-group peer assessment
- $I_{INA,i}$: Indonesia-based instructor evaluation
- $I_{JPN,i}$: Japan-based physics lead instructor evaluation

In this study, the Japan-based instructor refers to the physics lead instructor (Indonesian) who continuously observed onsite teamwork processes in Japan, whereas the Indonesia-based instructors participated remotely. The weighting scheme was designed to reflect both epistemic authority and visibility of student contributions. Instructor evaluations were assigned the largest weight because instructors possess disciplinary expertise and responsibility for summative judgment. However, instructor observation is inherently limited in collaborative project settings, particularly in COIL environments where internal group dynamics and task allocation are not fully visible. Peer assessments were therefore incorporated as corrective inputs, with within-group peers providing the most direct insight into individual contribution, and cross-group peers offering an external perspective on collaborative performance. The weighting coefficients were defined *a priori* and were not directly adopted from any single prior study. Instead, they were determined based on pedagogical considerations of epistemic authority and visibility of student contributions. The weighting factors were fixed prior to analysis and documented in course implementation records to avoid *post hoc* optimization, as summarized below.

The 70% weighting assigned to instructor evaluations is analytically justified by their role as the primary source of domain-specific validity and summative judgment, consistent with prior literature emphasizing expert evaluation as the anchor in composite assessment systems. Peer assessments function as corrective signals rather than primary evaluative sources, complementing instructor judgments by capturing aspects of contribution that are not directly observable. Specifically:

- Instructor ratings receive 70% total weight, serving as the primary validity anchor based on disciplinary expertise and summative judgment
- Within-group peer ratings (20%) capture direct observation of effort and teamwork
- Cross-group ratings (10%) provide external perspectives on communication and performance
- Within-peer weight > cross-peer weight reflects stronger access to internal processes [4]–[6], [11]

This weighting strategy is consistent with prior research recommending composite scoring to enhance fairness in collaborative STEM settings [9]–[14]. This composite model tends to reduce inflated scores for low-effort contributors relative to instructor-only evaluation, raises suppressed scores for high-effort contributors, and increases accountability and alignment between contribution and grade. The weighting scheme was determined prior to implementation and was communicated to students at the beginning of the course to ensure procedural fairness. The selected weights reflect a balance between disciplinary authority and visibility of contribution. Instructor evaluations were assigned a combined weight of 70% to preserve the validity of expert judgment, while peer evaluations were incorporated as corrective signals that capture internal group processes. Alternative weighting schemes (e.g., equal weighting or peer-dominant weighting) were considered during course design but were not adopted due to concerns regarding reliability and student trust in peer scoring. The adopted scheme therefore represents a theoretically grounded and pedagogically practical balance between fairness, transparency, and institutional expectations for summative assessment.

3. Results and discussion

Building on the interdisciplinary COIL course design described in Section 2.2, we evaluated whether the proposed multi-source assessment model improves transparency and procedural fairness by reducing score inflation while maintaining ranking stability. Figure 1 compares the distribution of assessment outcomes across two academic cohorts. In FY2024 (Fig. 1(a)), instructor assessments were tightly clustered and higher than peer evaluations, resulting in substantial downward adjustments by the model. In FY2025 (Fig. 1(b)), instructor and peer scores were more aligned (both higher overall), with the adjusted score distribution showing less dramatic correction.

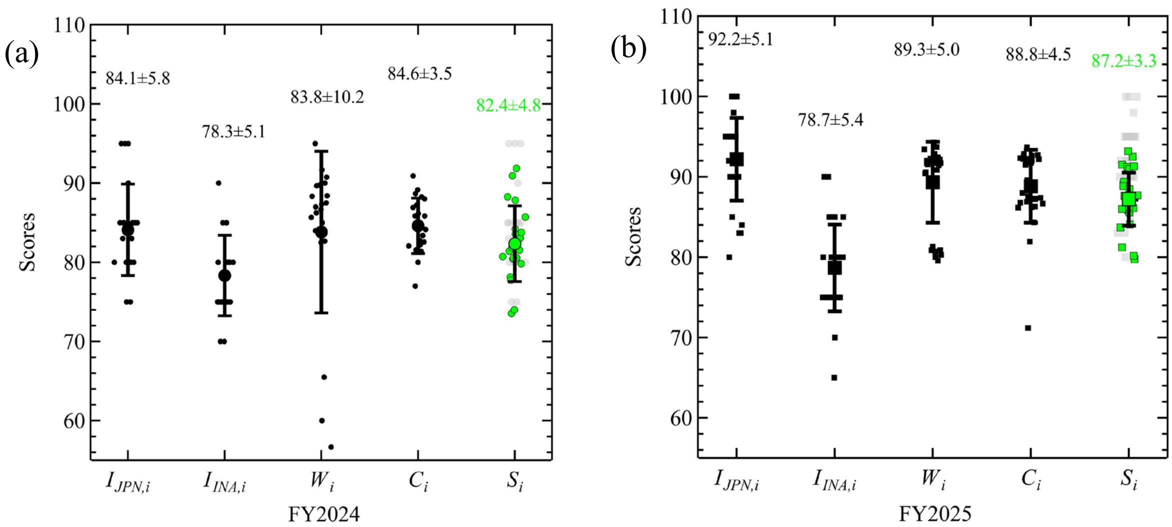

Figure 1. Comparison of score distributions across two academic years: (a) FY2024 and (b) FY2025.

For each year, scores from the Japan-based physics lead instructor ($I_{JPN,i}$), Indonesia-based instructors ($I_{INA,i}$), within-group peers (W_i), cross-group peers (C_i), and the final adjusted composite score (S_i) are shown. $I_{JPN,i}$ refers to the physics lead instructor based in Japan, who had continuous onsite access to teamwork processes. Grey dots in the S_i column indicate original instructor scores ("before"), while green dots indicate adjusted scores ("after"). Large black markers denote means of the data, with error bars representing ± SD.

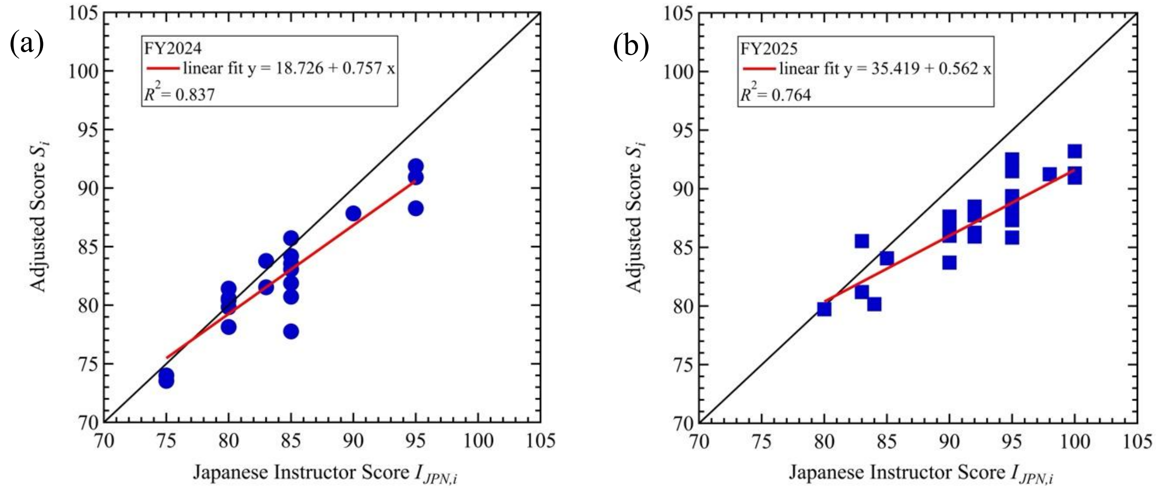

Figure 2. Relationship between Japan-based physics lead instructor scores ($I_{JPN,i}$) and adjusted composite scores (S_i) for (a) FY2024 and (b) FY2025. Each point represents one student. The black diagonal line indicates $S_i = I_{JPN,i}$ while the red line shows the least-squares fit to the data. For FY2024, the regression is S_i = 18.73+0.7 $I_{JPN,i}$ (R^2 =0.837); for FY2025, S_i =35.42+0.56$I_{JPN,i}$ (R^2=0.764). Slopes less than 1 and intercepts greater than 0 indicate that the adjustment compresses extreme instructor scores toward the center while preserving a strong linear relationship, supporting both fairness and validity.

Figure 2 examines the degree of alignment between adjusted scores S_i and instructor evaluations $I_{JPN,i}$ across the two cohorts. In both cohorts, the correlation is strong (R^2=0.837 in FY2024 and R^2=0.764 in FY2025), indicating that the adjustment model remains closely aligned with expert judgment. At the same time, the regression slopes are noticeably less than 1 and the intercepts are positive, showing that the transformation compresses very high and very low instructor scores toward the middle of the scale. In other words, the adjustment reduces extremes and potential inflation while still preserving the overall ordering of students, which is a desirable property for a fairness-oriented grading scheme. To evaluate whether the adjustment procedure altered relative student evaluation, Spearman's rank correlation coefficients were calculated between instructor-only scores and adjusted scores. High rank correlations were observed in both FY2024 (ρ=0.846) and FY2025 (ρ=0.826), indicating that student ranking was largely preserved after adjustment.

Taken together, these results show that the proposed multi-source adjustment model introduces statistically meaningful corrections while maintaining both linear agreement and rank-order stability across cohorts. These year-to-year differences likely reflect increased familiarity with peer-based evaluation and collaborative assessment practices over time, potentially contributing to reduced inflation. This pattern is particularly relevant in this interdisciplinary COIL setting, where contributions may include facilitation and contextual interpretation in addition to technical outputs.

Table 1. Summary of statistical analysis for data in FY2024 and FY2025.

Metric	FY2024	FY2025
Mean (Japan-based physics lead instructor)	84.1	92.2
Mean (Adjusted Score)	82.36	87.23
Paired t-test (p, two-tailed)	0.003	9.8×10^{-11}
Pearson r	0.915	0.873
Spearman ρ	0.846	0.826
Cohen's d	0.33	1.15

Table 1 summarizes the statistical comparison between J Japan-based physics instructor scores and the adjusted composite scores for FY2024 and FY2025. In both cohorts, the mean adjusted score is lower than the instructor-only mean, indicating systematic correction of grade inflation. In FY2024, the instructor mean was 84.10, while the adjusted mean decreased to 82.36. A paired-sample t-test revealed a statistically significant difference between the two score sets (two-tailed p = 0.003). In FY2025, the difference was more pronounced: the instructor mean reached 92.20, whereas the adjusted mean decreased to 87.23, with a highly significant paired difference ($p = 9.8 \times 10^{-11}$). Despite these mean shifts, strong linear agreement between instructor-only and adjusted scores was preserved. Pearson correlation coefficients remained high in both cohorts (r=0.915 for FY2024 and r=0.873 for FY2025), demonstrating that the adjustment did not disrupt overall alignment with instructor judgment. The results indicate that instructor-only evaluation in collaborative physics courses tends to produce inflated and tightly clustered scores, particularly evident in FY2025, where the instructor mean exceeded 92. The proposed adjustment model systematically moderates this inflation by incorporating peer-based assessments, leading to lower and more distributed adjusted scores. Importantly, the adjustment does not undermine evaluation validity. High Pearson correlations r indicates strong agreement with instructor judgment, while Spearman rank correlations ρ above 0.8 in both cohorts confirm that relative student ranking is preserved. This balance between correction and stability is critical for acceptance of alternative assessment schemes in formal engineering curricula, where concerns about fairness must be weighed against concerns about arbitrariness. The effect size was small-to-moderate in FY2024 (d = 0.33) and large in FY2025 (d = 1.15), indicating that the magnitude of correction increased substantially as instructor scores became more inflated. Importantly, the adjustment does not substantially alter student rank ordering, ensuring that grading decisions such as pass/fail thresholds or scholarship eligibility would remain stable. The stronger adjustment observed in FY2025 can be attributed to higher baseline instructor scores, rather than a failure of the model. In fact, the ability of the model to apply larger corrections while still maintaining rank-order consistency suggests robustness under varying grading cultures and cohort characteristics. This finding is particularly relevant in COIL contexts, where differences in instructional norms and cultural expectations may exacerbate grading inconsistencies. From a pedagogical perspective, the results support the use of peer assessment not as a replacement for instructor judgment, but as a corrective lens that reveals contribution disparities invisible to external observers. Particularly, the implementation of Humanities encouraged students to relate physics concepts to historical developments and socio-cultural contexts introduced in the humanities module. This interdisciplinary framing expanded the scope of contribution in student team projects and may have contributed to students' understanding of contribution and collaboration, which in turn may support acceptance of the proposed assessment model. By weighting peer evaluations appropriately, the model enhances accountability within collaborative teams while preserving instructor authority over final grading outcomes.

4. Conclusion

This study proposed and empirically evaluated a multi-source peer assessment framework designed to improve fairness and accountability in collaborative engineering education, particularly within COIL-based learning environments. By integrating instructor evaluations with within-group and cross-group peer assessments through a fixed weighted model, the framework addresses a central limitation of instructor-only grading: limited visibility into individual contributions within team-based projects. Across two academic cohorts, instructor-only evaluation produced systematically inflated and tightly clustered scores, especially in highly collaborative contexts. The proposed adjustment model significantly moderated this inflation while preserving strong agreement with instructor judgment. High Pearson correlations and Spearman rank-order stability demonstrate that the framework maintains grading validity while introducing meaningful correction. Importantly, the model proved robust across cohorts with different grading baselines, applying stronger correction when inflation was greater without disrupting relative student ranking. From a pedagogical perspective, these findings support the role of peer assessment as a corrective complement rather than a replacement for instructor authority. Properly weighted peer input functions as a mechanism to surface invisible labour and contribution asymmetries that are otherwise inaccessible in COIL environments, while maintaining institutional expectations for summative evaluation.

Several directions for future work emerge from this study. First, systematic exploration of alternative weighting schemes, for example, using constrained optimization or Bayesian approaches, could further refine the balance between instructor authority and peer insight. Second, longitudinal analysis at the individual-student level may clarify how repeated exposure to structured peer assessment influences contribution behavior, feedback literacy, and collaborative norms. Third, broader validation across disciplines, institutions, and cultural contexts is needed to assess generalizability, particularly in COIL settings where communication norms vary. Future studies could explore whether incorporating evaluations from non-physics instructors (e.g., from the humanities module) into the weighting scheme would further improve assessment validity in interdisciplinary COIL contexts. Finally, integrating qualitative components such as reflective self-assessment or rubric-based peer justification may enhance transparency and further reduce bias, strengthening the fairness and pedagogical value of multi-source assessment frameworks. In addition, the interdisciplinary design of the course, which incorporated a humanities module alongside physics content, broadened the notion of contribution beyond technical tasks by encouraging engagement with sociocultural perspectives, reinforcing the need for assessment models that capture diverse forms of participation in COIL environments.

Acknowledgements

The authors would like to thank the students from Shibaura Institute of Technology and Institut Teknologi Sepuluh Nopember who participated in this study. We also thank all the instructors and staff who supported this COIL module. This research is supported by JSPS Research Grants 24K06133.

References

[1] L. Kimura, H. Yoshikubo, and S. Nagasawa, “Enhancing assessment of student engagement in face-to-face global project-based learnings (gPBLs): Adding peer assessment to improve Slack-based evaluation,” in *Proc. ASEE Annu. Conf. Expo.*, Montreal, QC, Canada, Jun. 2025. doi: 10.18260/1-2--56401.

[2] J. H. Stonewall, M. C. Dorneich, and J. Rongerude, “Evaluation of bias in peer assessment in higher education,” *Int. J. Eng. Educ.*, vol. 40, no. 3, pp. 543–556, 2024.

[3] R. D. Handayani, M. U. Genisa, and Triyanto, “Empowering physics students’ performance in a group discussion through two types of peer assessment,” *Int. J. Instr.*, vol. 12, no. 1, pp. 655–668, 2019. doi: 10.29333/iji.2019.12142a.

[4] L. Li, N. Zhao, and F. Gao, “The effect of peer assessment on translation performance among EFL students at various achievement levels,” *Ital. J. Educ. Technol.*, vol. 31, no. 3, 2023. doi: 10.17471/2499-4324/1311.

[5] S. S. Douglas, D. T. Brookes, and V. Monty, “Peer assessment of student-produced mechanics lab reports,” *Phys. Rev. Phys. Educ. Res.*, vol. 13, no. 2, p. 020126, 2017. doi: 10.1103/PhysRevPhysEducRes.13.020126.

[6] A. E. Kay, J. Hardy, R. K. Galloway, and K. Devine, “Learning from peer feedback on student-generated multiple-choice questions,” *Phys. Rev. Phys. Educ. Res.*, vol. 14, no. 1, p. 010119, 2018. doi: 10.1103/PhysRevPhysEducRes.14.010119.

[7] A. Werth, L. B. Limeri, L. Zangori, and M. M. Cooper, “Assessing student engagement with teamwork in an online physics CURE,” *Phys. Rev. Phys. Educ. Res.*, vol. 18, no. 2, p. 020128, 2022. doi: 10.1103/PhysRevPhysEducRes.18.020128.

[8] L. Li and F. Gao, “The effect of peer assessment on project performance of students at different learning levels,” *Assess. Eval. High. Educ.*, vol. 41, no. 6, pp. 885–900, 2016. doi: 10.1080/02602938.2015.1048185.

[9] J. Heil and D. Ifenthaler, “Effects of teacher, peer and self-feedback on student improvement in online assessment: The role of individuals’ presumptions and feedback literacy,” *Assess. Eval. High. Educ.*, pp. 1–20, 2025, doi: 10.1080/02602938.2025.2530452.

[10] K. Ketonen, H. Lehtinen, and H. Koskinen, “Assessment designs of instructional labs: A literature review and a design model,” *Phys. Rev. Phys. Educ. Res.*, vol. 19, no. 2, p. 020601, 2023. doi: 10.1103/PhysRevPhysEducRes.19.020601.

[11] A. M. Madsen, S. B. McKagan, and E. C. Sayre, “Research-based assessment affordances and constraints in introductory physics courses,” *Phys. Rev. Phys. Educ. Res.*, vol. 12, no. 1, p. 010115, 2016. doi: 10.1103/PhysRevPhysEducRes.12.010115.

[12] Y. Zhan, Z. Yan, Z. H. Wan, X. Wang, Y. Zeng, and L. Yang, “Effects of online peer assessment on higher-order thinking: A meta-analysis,” *Br. J. Educ. Technol.*, vol. 54, no. 4, pp. 817–835, 2023. doi: 10.1111/bjet.13310.

[13] K. K. Özyer, “The impact of online peer assessment on academic performance in higher education: A meta-analytic review,” *J. Educ. Online*, vol. 22, no. 3, 2025. doi: 10.9743/JEO.2025.22.3.11.

[14] J. Lee, J. Kim, and M. Chae, “Effect of group type on group performance in peer-assessed exams,” *Phys. Rev. Phys. Educ. Res.*, vol. 18, no. 2, p. 020112, 2022. doi: 10.1103/PhysRevPhysEducRes.18.020112.

[15] G. Geschwind *et al*., “Development of a global landscape of undergraduate physics education research,” *Phys. Rev. Phys. Educ. Res*., vol. 20, no. 2, p. 020117, 2024. doi: 10.1103/PhysRevPhysEducRes.20.020117.

[16] Y. Cao, T. Väljataga, J. K. T. Tang, H. Leung, and M. Laanpere, Eds., *New Horizons in Web Based Learning*, Lect. Notes Comput. Sci., vol. 8699. Cham, Switzerland: Springer International Publishing, 2014. doi: 10.1007/978-3-319-13296-9.

[17] L. Czerwionka, S. S. Joshi, G. O. Rios-Rojas, and K. A. Davis, “Developing students' sociotechnical thinking in the humanities-engineering classroom: A treatment versus control study using a scenario-based assessment,” *Journal of Engineering Education*, vol. 114, no. 3, May 2025, doi: 10.1002/jee.70004.

[18] National Academies of Sciences, Engineering, and Medicine, *The Integration of the Humanities and Arts with Sciences, Engineering, and Medicine in Higher Education: Branches from the Same Tree.* Washington, DC: National Academies Press, 2018. doi: 10.17226/24988.

[19] H. Yoshikubo, D. P. Sari, G. Trovato, E. Kamioka, and D. Darminto, “Exploring the Impact of STEM COIL Modules on Global Competence: A Study of Indonesian Physics Undergraduates and International Engineering Students in Japan,” in *Proc. ASEE Annu. Conf. Expo.*, Montreal, QC, Canada, Jun. 2025. doi: 10.18260/1-2--56537.