Research Paper

Recommended citation: Izadkhast, S., Ercan, I., & Abdikivanani, B. (2025). Peer Evaluation and Academic Performance in Bachelor of Science Electrical Engineering: A BuddyCheck Case Study with Statistical Analysis. In Kangaslampi, R., Langie, G., Järvinen, H.-M., & Nagy, B. (Eds.), SEFI 53rd Annual Conference. European Society for Engineering Education (SEFI), Tampere, Finland. DOI: 10.5281/zenodo.17631636.

Peer Evaluation and Academic Performance in Bachelor of Science Electrical Engineering: A BuddyCheck Case Study with Statistical Analysis

S. Izadkhast [a,1], **I. Ercan** [b], **B. Abdikivanani** [c]

[a] Microelectronics Dept., TU Delft, Delft, Netherlands, ORCID: 0000-0002-7594-6322
[b] Microelectronics Dept., TU Delft, Delft, Netherlands, ORCID: 0000-0003-1339-9703
[c] Microelectronics Dept., TU Delft, Delft, Netherlands, ORCID: 0000-0001-5466-1638

***Conference Key Areas**: Engineering skills, professional skills, and transversal skills, Improving higher engineering education through researching engineering education*
***Keywords**: Academic Performance, BuddyCheck, Correlation, Peer Evaluation.*

ABSTRACT

This paper evaluates the effectiveness of the peer evaluation in a first-year Bachelor of Science (BSc) Electrical Engineering project involving 17 groups of 6–9 students. Students evaluated each other and themselves on five peer evaluation dimensions, namely job performance, attitude, leadership/initiative, communication, and teamwork, using a 1–5 scale (5 being the highest). The academic performance metrics (first-year BSc overall grade point average (GPA) and project final grade) were collected for our analysis. A dedicated measure, the “Factor” (a student’s average peer rating divided by the overall group average), was used to measure the peer evaluation results. Overall, though the correlation between peer evaluations and academic performance was low (r = 0.04), we found a strong correlation (r = 0.71) among students with lower peer evaluation scores. In general, in groups, self-assessments and peer evaluations were highly correlated (r = 0.82). We performed further statistical analyses such as multiple linear regression, clustering, mediation analysis and random forest regression in this study. While peer evaluations capture important aspects of teamwork and interpersonal skills, for most of the students, they seem likely more reflective of project-related competencies than necessarily only the overall GPA. Our findings suggest that the insights from BuddyCheck data can serve as an early indicator for targeted future interventions, enhancing collaborative learning outcomes in our projects. Note that, to further preserve anonymity, neither the project name nor the academic year / cohort is disclosed in this paper.

[1] *Corresponding Author*
S. Izadkhast
s.izadkhast@tudelft.nl

1 INTRODUCTION

Peer evaluation (PE) is a powerful tool to tackling the crucial challenges of project group work by enforcing accountability and fostering essential professional and academic skills such as communication, performance, teamwork, leadership, and critical thinking, etc. (Cheetham & Varga-Atkins, 2021) demonstrate that well-designed peer learning groups and activities can significantly improve the quality of group dynamics as well as the academic success and well-being of the students.

1.1 Literature on Challenges and Limitations in Peer Evaluations

(Gueldenzoph & May, 2002) emphasize the practices in collaborative peer evaluation, noting that subjective biases, which is present in the group working process in peer evaluations, may lead to inconsistent grading and it proposes ways to alleviate these subjective biases. (Van den Bogaard & Saunders-Smits, 2007) further illustrates that incorporating both peer and self-evaluations in project-based learning can further improve the accuracy of assessments. However, their work also reveals that the way students grade each other can remain perhaps inconsistent, creating challenges in obtaining fair and true contributions of each individual student. This inconsistency is emphasized by (Van Helden et al., 2023), whose very comprehensive umbrella review of digital peer assessment in higher education highlights critical issues and problems regarding unfair and inconsistent grading standards. (Van Helden et al., 2023) discuss how digital peer assessment tools have been employed across various educational programmes and emphasize that some research papers report a weak to moderate correlation between peer evaluation ratings/scores and students' academic success outcomes. This is perhaps likely affected by factors such as evaluation (in-)consistency and implementation practices. (Gueldenzoph & May, 2002) imply that when the PE is implemented properly, lower peer evaluation scores probably tend to be associated further with lower performance in group work. In addition, (Johnson, et al, 2022) report that digital tools such as the BuddyCheck software have been utilized to facilitate the early detection of dropouts and provide valuable insights into individual project contributions. Despite all these benefits and advantages, challenges such as unequal and unfair workload distribution and free riding can still remain in the project work. (Musgrove, 2023) and (Topping, 2023) mention that discrepancies between self-assigned and peer-evaluated grades might affect the reliability of PE results. (Topping, 2023) highlights the need for more rigorous research designs to better understand how digital peer assessment may relate to future academic performance and success.

1.2 Effective Peer Assessment Procedures

Note that we are aware of the importance of peer evaluation as well as peer assessment where the students have a chance to grade the work of other students. (Van Gennip, 2012) discusses that the interpersonal dynamics within peer assessment like trust and mental safety play a key role in improving students' reflective skills. (Fellenz, 2006) made a development in peer assessment protocol that focusses on fair evaluation and grading practices in group work, ensuring that individual contributions are properly incorporated into the final evaluations. (Panadero & Alqassab, 2019) highlight that the impact of peer assessment on academic outcomes can be in general context dependent influenced by factors like

educational level. (Topping, 2009) provides some proof that systematically implemented peer assessment is reliable and valid leading to enhance the student's learning and more involvement with the learning objectives.

1.3 Our Research Goal

In our peer evaluation practices, we aim to determine whether students who underperform academically (as measured by overall GPA) also tend to receive lower scores in project-based peer evaluations. Additionally, we aim to evaluate the extent to which self-assessment scores highlighting self-efficacy correlate with peer evaluation scores, and we explore the relationship between these peer evaluation scores and overall academic performance.

2 METHODOLOGY

2.1 Research Question

Based on the challenges and gaps identified in our literature review, we have revised our main research question (RQ) as follows: How does BuddyCheck grade normalization improve peer-evaluation consistency and correlation with self-evaluation, and how well do normalized peer-evaluation dimensions (performance, attitude, leadership, communication, teamwork) predict first-year GPA?

2.2 Methods and Statistical Analyses

Study Context and Data Collection: This study was conducted within a first-year BSc Electrical Engineering project involving 17 groups of 6–9 students. The BuddyCheck system was used for both peer and self-evaluation. The students rated each other and themselves on five peer review questions: 1) job performance, 2) attitude, 3) leadership/initiative, 4) communication, and 5) teamwork, where they give a grade using a 1–5 scale. The details of these five peer questions are provided in Appendix A. Academic performance was assessed using first-year GPA and the project grade. Note that all GPAs and course grades in this study are reported on a 10-point scale (where 10.0 corresponds to the highest performance, 9.0–10.0 to an A range, 8.0–8.9 to a B range, etc). A derived measure, the "Factor," was calculated by dividing a student's average peer rating by the overall group average, providing insight into relative individual contributions. We analyze our extensive data using statistical software (e.g., Python's pandas, statsmodels, and sklearn).
Ethical / Consent Considerations: Note that all our analyses were performed on anonymized and aggregated data. Data collection received prior approval from the Ethics Committee, as well as informed consent forms obtained from all students involved in this research for publication purposes.
Analytical Procedures: Considering our existing data, we perform our research on several relevant questions using a wide range of statistical analyses as follows:

- **Descriptive Analyses :** These analyses evaluated the statistical distribution and variability of peer ratings across the project groups. By summarizing and categorizing the data, we can better understand overall trends.
- **Correlation Analyses :** This analysis explored the relationships and associations between raw peer ratings on 1-5 scale, the derived adjusted grade/ "Factor" and academic outcomes (GPA and project grade). Correlation analyses help determine whether there is any linear relation between our PE metrics and academic performance (GPA).

- **Regression Analysis :** We use a multiple linear regression model to predict first-year GPA from the Adjusted Average Without Self (Adj_Avg_without_self) peer review grades. This method helps better quantify the extent to which peer evaluations explain variations in academic performance, testing whether these evaluations can serve as a reliable predictor for the GPA.
- **Clustering :** Hierarchical and K-Means clustering techniques help identify natural groupings within our data related to peer evaluation scores and academic performance. Clustering helps reveal whether distinct subgroups exist (for instance, at-risk low performing subgroups versus high-performing students).
- **Comparative Analyses :** To further confirm the results from previous analyses, we performed T-tests, Mann–Whitney U tests, and MANOVA where we compare the clusters and examine differences in individual peer evaluation questions (Q1–Q5). These tests help us evaluate whether the identified clusters differ significantly in terms of academic outcomes / GPA.
- **Additional Analyses : 1) Mediation Analysis:** We perform it to test whether self-evaluation mediates the relationship between peer evaluation and GPA. **2) Random Forest Regression:** To capture potential non-linear relationships between evaluation metrics and GPA, we conduct this modeling approach.

Each of the analyses above contributes to a better understanding of how our peer evaluation data relates to academic performance.

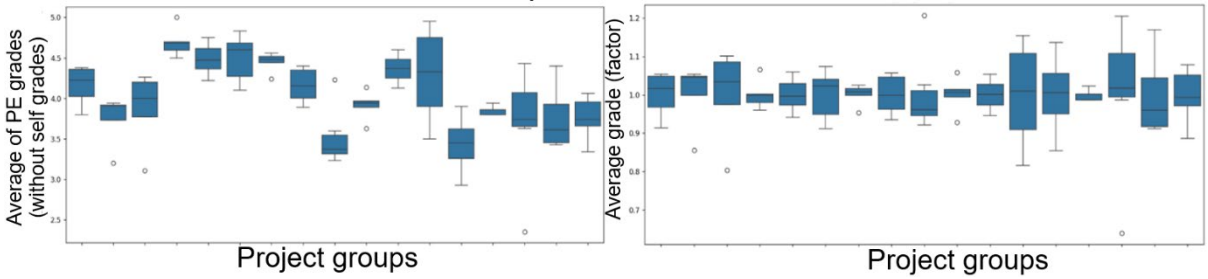


Fig. 1. (a) Average of peer grades peer group (without self) for each student on the scale of 1-5. (b) Adjusted grade (factor) (without self) of peer grades for each student.

3 RESULTS

3.1 Comparison of Raw PE Data (1–5 Scale) & BuddyCheck Factor Using Descriptive and Correlation Analyses

We start our study with descriptive analysis where we first evaluate the distribution and mean value of peer evaluation grades per group. Figure 1(a) illustrates the average peer ratings received by each student within their respective groups. It is evident that both the mean and the variation of these ratings differ substantially from one group to another, raising questions about why students interpret the same 1-to-5 grading rubric in different ways (for further details, see Appendix A). In addition, Figure 1(b) presents a metric called the "Factor" for all 17 groups of the project. This Factor, derived from the peer evaluation responses, reflects each student's contribution relative to their team average PE grades. In practice, it is calculated by dividing a student's average rating by the overall average rating of the team. If the factor is equal to 1, then it implies that the student performs at the average of the group. If it is above 1, then the student performs above the average of the group (and vice versa). The term “With Self” in this paper refers to the student's self-assessment as reflection for self-efficacy, the grades each student gives to themselves. We calculate the overall variance for both sets of raw and factor data.

Our results show that the overall variability in the peer ratings on scale 1-5 (5 being the highest) is notably higher than in the factor scores. In particular, the overall variance for the 1-5 scale ratings is about 0.245, whereas the variance for the factors is only 0.007. This notable reduction in variance suggests that perhaps normalizing the peer ratings by the group average (i.e., using the Factor) may remarkably reduce the dispersion in scores, resulting in a more consistent measure across students. As a result, for our further analyses, we mainly use the factor to compare the groups.

3.2 Group-Level Associations Between Peer/Self Evaluations and Academic Performance Using Correlation Analyses

Figure 2 displays the average self-grades per group versus the average peer evaluation scores received from fellow group members. The correlation analysis reveals a strong positive correlation (r= 0.82) between self-grades and peer grades. This indicates that within each group, students who give higher scores to their peers also tend to rate themselves generally at a higher level as well.
Moreover, we perform a correlation analysis to check whether the groups with higher peer evaluation grades, on average in an aggregated way, also have a higher GPA or not. The correlation between the group average peer rating and GPA is about 0.22, while the correlation for the group average self-rating and GPA is around 0.24. Thus, it is likely that groups with higher average peer and self-evaluation scores moderately tend to have higher average GPAs as well. Since the correlation is modest and not high value, these results should be treated as only complementary to other indicators.

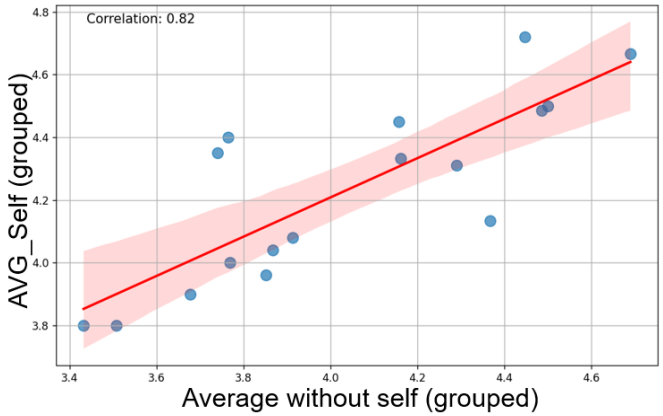
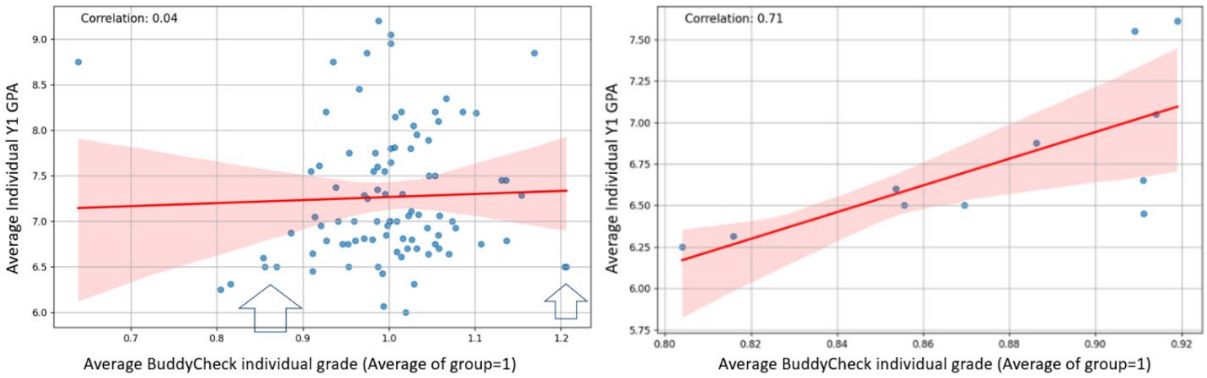


Fig. 2. Average grades per group for self-grades versus peer grades.

3.3 Evaluation of Student's PE Performance & Academic Performance (GPA)

Our third research question in this paper is whether there is a correlation between a student's performance in peer evaluations and their overall academic performance in the programme. To assess this, we compared each student's first-year average GPA with the average BuddyCheck grade received from other peers. Figure 3(a) shows a very low correlation (r= 0.04), suggesting that, in general, there is little relationship between academic performance and peer evaluation outcomes considering the entire set of students. We also performed the multiple linear regression model aiming to predict first-year GPA (AVG Year1 GPA) using predictor Adjusted Average (factor) Without Self. The overall model resulted in only 3.8% of the variance in GPA with a p-value of 0.1769, indicating that the regression model is not statistically significant. Note that in Figure 3(a), there is a student with GPA of 8.7 with the PE factor below 0.65. This was an exception case where the student did stop with this project in first year and joined it in another year.

Figure 3(b) shows that among students with an average peer evaluation grade below 0.92, the correlation with academic performance is considerably higher (r= 0.71). This finding indicates that while no general conclusion can be drawn for all students, those who underperform in peer evaluations are likely to also show lower academic performance. Conversely, high performance in peer evaluations does not necessarily guarantee a high overall GPA. This may suggest that students with strong communication, interpersonal and leadership skills may tend to receive better peer evaluation scores, yet these skills do not always translate into a high grade in other courses of the program. In short, low peer evaluation scores likely raise a red flag, but high peer evaluation scores alone do not per se fully predict a high GPA.



(a) (b)

Fig. 3. (a) Average individual grade Y1 GPA versus the average BuddyCheck individual factor [r=0.04]. (b) Average individual grade Y1 GPA versus the average BuddyCheck individual factor of the students with the factor below 0.92 [r=0.71].

To look more closely into the associations between the peer review questions, overall GPA and project grades, we apply the correlation analysis. The resulting correlation matrix (Figure 4) reveals that the peer evaluation items, namely Q1 (Job Performance), Q2 (Attitude), Q3 (Leadership/Initiative), Q4 (Communication), and Q5 (Interacting with Teammates), are highly correlated and interrelated, with most correlations above 0.70 (for instance, Q2 and Q5 correlate at 0.91, and Q1 with Q3 at 0.87 that is very high). This reveals that students tend to rate peers consistently across different questions and dimensions. In contrast, in Figure 4, the correlations between these peer evaluation items and AVG Year1 GPA are generally low (ranging from 0.05 to 0.24, which is for Q1 performance), indicating that academic performance, as measured by GPA, does not strongly align with these peer evaluations. However, the project grade shows moderate correlations with the peer evaluations (between 0.43 and 0.64), showing that peer evaluations are more reflective of performance in our course project. This result is consistent with our expectations, since a large part of project final grade is a group grade component, and the final grade of individual students is modified by the peer review results. Hence, we evidently expect to obtain a moderate correlation between the PE questions and final project grade. Moreover, we performed a GPA threshold analysis where it resulted in a threshold/cut-off of 7.00. The average probability that the students with GPA below 7.00 perform well in the peer evaluations is 0.57, while it is 0.84 for those at or above 7.00. Our t-test result, that aims to measure the amount of difference between these two cluster groups, provides a high negative t (t = -8.09) and very low p value (p = 0.0002), indicating a large difference between these two cluster groups. A point-biserial correlation, aiming to measure the relationship between a binary variable (i.e., being above or below the GPA threshold of 7.00) and a continuous variable (i.e., peer evaluation score), gives a correlation of 0.38 with very low p value (p

= 0.0003) further supporting the link between academic performance and these outcomes. These findings again affirm that while peer evaluations consistently capture important aspects of teamwork, leadership and interpersonal skills, they may not translate directly into overall academic success as it is measured by course grades and GPA. Instead, the PE results perhaps seem affected largely also by the practical, project-specific competencies. The valuable insights from the peer evaluation data analysis could maybe serve for early identification of students who might benefit from targeted supportive interventions, ultimately enhancing both collaborative processes and project-based performance in project-based learning environments.

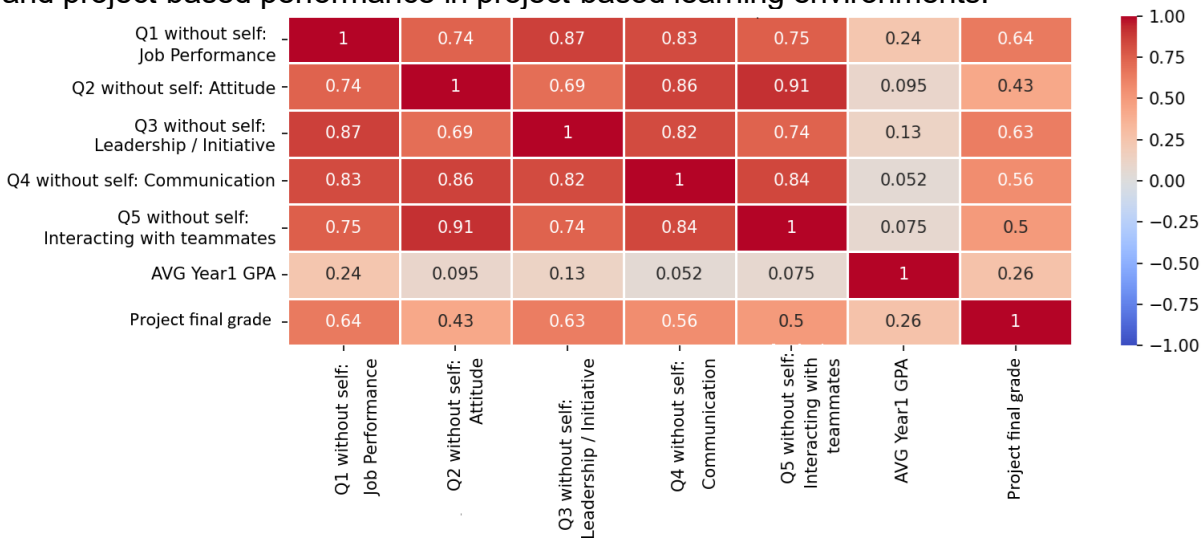


Fig. 4. Correlation matrix of give peer evaluation questions (Q1-Q5), average GPA of individual students, and the final project grade.

3.4 Cluster Analyses

We performed hierarchical clustering of the performance data that resulted in two distinct clusters. 1) Cluster 1 includes students with a mean Adjusted Average Self (Adj_Avg_self) of approximately 0.976 and a mean AVG Year1 GPA of about 6.73, revealing that these students may give themselves lower self-evaluation ratings and achieve lower academic grades. 2) In contrast, Cluster 2 consists of students with a higher mean Adj_Avg_self of approximately 1.031 and a higher AVG Year1 GPA of around 7.95, implying that these students not only rate themselves higher but also tend to perform better academically with higher GPA. We also perform a Mann–Whitney U test comparing the AVG Year1 GPA between these clusters. It confirms a statistically significant difference with very small p value ($p = 0.0001$), emphasizing that these groups differ notably in their academic outcomes. Additionally, we carry out a MANOVA on the five peer evaluation questions. This analysis confirms significant differences between these two clusters with very small p value ($p = 0.0033$). When we evaluate each evaluation questions/dimension separately, we notice that Q1 (Job Performance) differed more significantly between clusters. These findings suggest that differences in job performance ratings can be the most prominent element of the overall clustering, with Cluster 1 representing students who may be at higher risk for underperformance and Cluster 2 including those with stronger academic performance.

3.5 Mediation Analysis

Mediation analysis was conducted with the objective of investigating whether self-evaluation mediates the relationship between peer evaluation and academic

performance (GPA). The analysis led to an indirect effect (mediation effect) of 0.2146 and a large p-value of 0.3600. This indicates that the mediation effect is not statistically significant. This suggests that while self-evaluation may play a role in the relationship between peer evaluations and GPA, it does not significantly mediate this association within our current dataset.

3.6 Random Forest Regression Analysis

So far, our analysis has been carried out using linear analysis. A random forest regression model was applied to capture potential non-linear relationships between the evaluation measures and academic performance. Using 100 trees, the model achieved a mean squared error (MSE) of 0.0799 when predicting AVG Year1 GPA. This result implies that the non-linear model accounts for some variance in academic performance, revealing that relationships between peer evaluation metrics and GPA might be more complex than what is captured by linear models alone.

4 DISCUSSION AND CONCLUSIONS

Our analyses show several valuable insights into the relationship between peer evaluations and academic performance. First, normalizing the raw peer ratings (on a 1–5 scale) by the group averages (using the Factor / adjusted grade) notably reduces the dispersion in scores, leading to a more consistent measure across groups and students. Within each group, a strong positive correlation ($r = 0.82$) was observed between self-assessment scores highlighting self-efficacy and peer evaluation scores, indicating that students who give higher scores to their peers also likely tend to rate themselves higher. In general, the overall correlation across all students between PE results and GPA was weak ($r= 0.04$), as confirmed also by some references in this paper. Notably, when focusing on students with lower peer scores, the association with GPA was much stronger ($r= 0.71$), revealing that low peer evaluation scores may serve as early indicators of academic underperformance. Hierarchical clustering further distinguished two groups with a significant GPA difference. The random forest regression indicate that non-linear and group-related components might be present. These findings highlight the potential of digital peer evaluation systems, like BuddyCheck, to serve as early indicators and signals for academic underperformance in project-based learning environments. Although high peer evaluation scores do not necessarily guarantee high GPA, consistently low PE scores, especially in key questions such as job performance, can reliably filter students who may benefit from targeted interventions. The modest yet positive group-level correlations indicate that the peer evaluations have a good potential to offer valuable supplementary insights into student performance and success. Our future research should focus on replicating these insights and findings in a larger number of BSc projects and across more diverse cohorts to assess both the consistency and predictive potential of our existing peer evaluation data.

5 ACKNOWLEDGEMENTS

The authors developed this work and collected the required data with informed consent from participating project students, securely stored it on a project safe drive, and obtained necessary approvals from the Ethics Committee. The authors transparently acknowledge that during the manuscript's development, ChatGPT was partly used for language refinement, idea generation, interpretation, and Python coding assistance.

REFERENCES

Cheetham, J., & Varga-Atkins, T. (2021). Designing peer learning groups and activities. URL: https://www.liverpool.ac.uk/media/livacuk/centre-for-innovation-in-education/diy-guides/designing-peer-learning-groups-and-activities/designing-peer-learning-groups-and-activities.pdf

Fellenz, M. R. (2006). Toward fairness in assessing student groupwork: A protocol for peer evaluation of individual contributions. *Journal of Management Education, 30*(4), 570–591. URL: https://doi.org/10.1177/1052562906286713

Gueldenzoph, L. E., & May, G. L. (2002). Collaborative peer evaluation: Best practices for group member assessments. *Business Communication Quarterly, 65*(1), 9–20. URL: https://doi.org/10.1177/108056990206500102

Johnson, C., Macleod, M., & Visscher, K. (2022). The utility of a peer review application in interdisciplinary teamwork arrangements. In *Towards a new future in engineering education, new scenarios that European alliances of tech universities open up* (pp. 1989–1993). Universitat Politècnica de Catalunya. URL: https://upcommons.upc.edu/handle/2117/387392

Musgrove, A. (2023). 'Buddy check' peer observation activity for hands-on learning in analytical chemistry laboratories. *Analytical and Bioanalytical Chemistry, 415*(17), 3299–3303. URL: https://doi.org/10.1007/s00216-023-04755-y

Panadero, E., & Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assessment & Evaluation in Higher Education, 44*(8), 1253–1278. URL: https://doi.org/10.1080/02602938.2019.1600186

Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*(1), 20–27. Retrieved from https://doi.org/10.1080/00405840802577569

Topping, K. J. (2023). Digital peer assessment in school teacher education and development: A systematic review. *Research Papers in Education, 38*(3), 472–498. URL: https://doi.org/10.1080/02671522.2021.1961301

Van den Bogaard, M., & Saunders-Smits, G. (2007). Peer & self-evaluations as means to improve the assessment of project based learning. In *2007 37th Annual Frontiers In Education Conference – Global Engineering: Knowledge Without Borders, Opportunities Without Passports* (pp. S1G-12). IEEE. URL: https://ieeexplore.ieee.org/abstract/document/4417988/

Van Gennip, A. E. (2012). Assessing together. Peer assessment from an interpersonal perspective (Doctoral dissertation). Leiden University. URL: https://openaccess.leidenuniv.nl/handle/1887/20012

Van Helden, G., Van der Werf, V., Saunders-Smits, G., & Specht, M. (2023). The use of digital peer assessment in higher education – an umbrella review of literature. *IEEE Access*. URL: https://ieeexplore.ieee.org/abstract/document/10058959/

APPENDIX A: PEER EVALUATION QUESTIONNAIRE

Q1. Job Performance
5 – Very good: Consistently does more than required; work is of very high quality.
4 – Good: Sometimes does more than required; work is of high quality.
3 – Satisfactory: Performs all assigned tasks; quality meets standards.
2 – Marginal: Performs tasks but work must be redone to meet standards.
1 – Poor: Performs some tasks; work must be redone by others.

Q2. Attitude
5 – Excellent: Positive, professional, and influences the team favorably.
4 – Good: Positive attitude toward the project and team.
3 – Satisfactory: Neutral attitude.
2 – Marginal: Negative, unprofessional attitude.
1 – Poor: Negative attitude that adversely affects the team.

Q3. Leadership / Initiative
5 – Excellent: Provides strong leadership and takes initiative.
4 – Good: Accepts tasks and sometimes seeks additional responsibilities.
3 – Satisfactory: Completes assigned tasks and maintains some oversight.
2 – Marginal: Limited involvement and initiative.
1 – Poor: Minimal participation; relies on others.

Q4. Communication
5 – Very good: Excellent oral and written skills; communicates effectively.
4 – Good: Adequate communication; clear and effective most of the time.
3 – Satisfactory: Generally, communicates effectively, with occasional lapses.
2 – Marginal: Communication skills are ineffective.
1 – Poor: Poor communication with little effort to improve.

Q5. Interacting with Teammates
5 – Excellent: Actively engages with teammates; ensures clear communication.
3 – Satisfactory: Listens and respects teammates but is inconsistent.
1 – Poor: Disruptive or uncooperative; does not engage effectively.