



Quantifying individual contribution in team-based learning: Integrating peer evaluation with learning analytics

Ioannis Katsenos ¹

 0000-0003-3311-4170

Spyridon Papadakis ²

 0000-0001-8592-4066

Christos Pierrakeas ^{1*}

 0000-0003-1068-0444

¹ Department of Management Science and Technology, University of Patras, Patras, GREECE

² Hellenic Open University, Patras, GREECE

* Corresponding author: pierrakeas@upatras.gr

Citation: Katsenos, I., Papadakis, S., & Pierrakeas, C. (2026). Quantifying individual contribution in team-based learning: Integrating peer evaluation with learning analytics. *Contemporary Educational Technology*, 18(4), Article ep696. <https://doi.org/10.30935/cedtech/19553>

ARTICLE INFO

Received: 8 Apr 2026

Accepted: 14 Sep 2026

ABSTRACT

We used data from a training program involving a large group of informatics schoolteachers who were trained in the learning activity management system (LAMS) using team-based learning methodology. Extending previous work, we introduced a numerical indicator for the peer evaluations that everyone performed on their peers in their group by introducing the improved project contribution score (IPCS). This way we quantified and summarized the feedback received by individuals within the teams by their colleagues. In parallel, time spent on activities' sequences in LAMS was used as an objective indicator of learner engagement and its relation to IPCS was examined at different levels. While there was observed only a negligible negative correlation at the total sample level, using generalized linear mixed models and generalized additive mixed models, random forest, and clustering, we identified that team membership and team size were the dominant explanatory factors and showed that cluster-based profiles (high-time/average-IPCS, low-time/high-IPCS, low-time/low-IPCS) exist at deeper level. This could be a useful insight for the trainer as it can serve to identify differentiated support strategies to be applied to team members, especially if applied in a formative way. Finally, the application of peer evaluation and engagement within LAMS-supported group work fills in a sectoral gap, and our article applies patterns reported mainly in university coursework to competency-based, practice-oriented training programs.

Keywords: individual contribution, teamwork, objective evaluation, trainee engagement, lifelong learning

INTRODUCTION

Working in teams nowadays, whether in physical or digital environments, is highly desirable in many areas of human activity, however understanding and measuring individual contributions in the context of team activities remains a complex challenge. The main problem is that the final deliverables/artefacts (e.g., document, code, etc.) reflect the collective effort, without being able to distinguish the contribution of each member. This difficulty has significant implications for both the evaluation and management of human resources and incentives in business, education and science environments. In the field of education, which is our focus here, fair evaluation of trainees working on team projects is particularly important, both for their individual grading fairness and for the development of the desired collaboration skills. Peer evaluation in conjunction with objective data gathered from the logs of a digital collaboration platform seem an attractive means of providing an insight into the support needed by trainees.

In this article, we use data collected from the learning activity management system (LAMS) after the completion of an adult education (lifelong learning) program following the team-based learning (TBL) methodology, to attempt to provide educators with means to have an insight into the performance of the learners, combining peer evaluations and the total time spent on learning activities.

The research questions (RQs) attempted to be answered in this article are:

RQ1. Is there a correlation between the time learners spend on LAMS activities and the peer evaluations they receive in TBL?

RQ2. Can the teachers/trainers use any of these two measures or their combination to gain insight into the support needed by teams or individuals?

We addressed the above RQs using a dataset comprising 190 adult trainees in 50 teams of a training program for primary and secondary education informatics teachers. Their time spent ('TimeSpent') on different activities of the training program was measured by the LAMS platform, which was the collaboration environment used. At the conclusion of the program, the trainees provided peer evaluation for their teammates by distributing marks to those who mostly contributed to the teams' deliverables and effort, using again a LAMS activity.

For addressing the above RQs, initially we defined the improved project contribution score (IPCS) and examined its correlation and linear regression to the time spent on LAMS sequences of activities ('TimeSpent'). We then examine whether classification of trainees according to their project contribution (IPCS) and/or 'TimeSpent' on LAMS sequences might yield meaningful insights for the trainer, e.g., for quickly identifying special trainees' cases needing further attention.

BACKGROUND

Groups, Work Groups, and Teams

In general, the term group is used to refer to any collection of individuals who regard themselves as belonging together without necessarily sharing tightly interwoven tasks or collective goals. Narrowing the concept, a workgroup typically refers to a formally established set of employees who carry out related but largely independent activities under proper supervision. Each member's output contributes to a common institutional objective, but accountability remains mainly individual, even though members draw on a shared resource pool and follow coordinated schedules (Cohen, 1997; Sundstrom et al., 1990).

By contrast, a "team" corresponds to the most integrated form of collective work. Team members bring complementary expertise to common objectives while depending on one another in real time and hold themselves mutually as well as individually responsible for outcomes (Kozlowski & Ilgen, 2006; Mathieu et al., 2017). Decision-making within teams can be shared or distributed, precisely because the high interdependence usually demands rapid collective response. Consequently, while groups and workgroups operate adequately with individual evaluation schemes, teams require performance measures that blend personal achievement with joint results (Salas et al., 2005). The progressive shift from group to workgroup to team is based on increasing task interdependence, synergy expectation, and collective accountability and, in parallel, each step requires the necessary changes in leadership style, reward structures, and evaluative practice.

Individuals' Peer-Assessment in Teamworking

Reliability and ethics in the evaluation of team members concerns both the proper distribution of recognition for individuals' contributions, but also the maintenance of trust, transparency, and fairness among them. The adoption of an evaluation system that combines objective data (e.g., system logs) with subjective assessments (peer scoring) has been confirmed as an effective, safe, and fair strategy in the literature (Hundhausen et al., 2023; Shi et al., 2021).

The need for reliable and fair evaluation methods arises from the imperative requirement of transparency, accountability and responsibility, especially when success depends on collective efforts, as the lack of clear role allocation and evaluation of individual performance leads to a decrease in responsibility and an increase in "free riding" phenomena (Tasi et al., 2025). Johnson et al. (1991) states successful student teams require

positive interdependence, where members' successes are linked, and individual accountability, where each person is assessed for their contribution to the shared goal.

Peer assessment, peer evaluation, peer review, and peer rating/ranking are closely related methodologies used to identify individual performance within a team. The terms are often used interchangeably, but they do not reflect identical concepts. Peer assessment refers primarily to the more general process, emphasizing skill development and learning through feedback. Peer review focuses primarily on the critical evaluation of the quality of work, based on predetermined criteria. Peer rating/ranking refers primarily to the assignment of a numerical score, usually for final/summary evaluation purposes.

There are different methodological models of peer assessment. In point allocation, students/evaluators evaluate others by distributing a fixed number of points to fellow students/collaborators. In questionnaire distribution/collection, students/evaluators complete questionnaires for their peers, usually answering on Likert scales and evaluating concepts such as collaboration, professionalism, quality of contribution, communication, and timely participation, but also the degree of contribution to specific tasks or deliverables. For example, Mathieu et al. (2015) provides a questionnaire to determine the role assumed by each member of the group. In rank-ordering or forced distribution (Baker, 2008; Davison et al., 2014; Fete et al., 2017) group members are asked to rank the other members according to some criterion/criteria, e.g., from the most active to the least. Ranking prevents giving the same ratings to everyone (leniency bias). Finally, in Open-Ended feedback, raters record comments for each member and therefore the assessment is qualitative but can help triangulate the results.

Peer-Assessment, Reliability, and Bias

The validity of peer assessment depends on factors related to the assessors, the tasks and systems used, as well as the broader educational context.

Regarding raters, the literature suggests that a larger number of raters improves validity and reliability (3–6 raters are reported for individual work and 14–17 for group work to be sufficient), but this is also influenced by other factors such as proper training of the raters and anonymity (Cho et al., 2006; Sung et al., 2010; Tong et al., 2023). Also, the education and experience of raters reduce the number of ratings required for accuracy, and that higher academic ability similarly reduces this need. On the contrary, rater bias – such as leniency, harshness, or a tendency toward central ratings – can reduce validity, while one studies (Dunning et al., 2004; Kruger & Dunning, 1999) report that low or high achievers may overestimate or underestimate themselves when self-rating.

Bias, on the other hand, is defined as the systematic deviation of evaluators' judgment from objective scientific criteria, due to characteristics unrelated to the scientific content (e.g., gender, country, and institutional prestige) and can affect the quality and fairness of decisions. Some sources of bias can be: Friendly/in-group favoritism, in which members with already close ties or shared characteristics rate each other favorably. Bias can be limited when ratings are made after external confirmation of success, otherwise it persists or intensifies (Speckbacher & Wiernsperger, 2025). Indulgence and central tendency bias, in which everyone gets "good" grades, so the scale loses discrimination ability. Reciprocity/strategic cooperation, in which team members exchange high ratings, expecting mutual benefit. Demographic biases, related to gender, age, country of origin, race. Stonewall et al. (2018) documented lower scores for women and students of color compared to white men, even with equal objective performance.

Student engagement, crucial for persistence and success, is a complex, multidimensional construct. Learning analytics (LA) captures digital footprints from systems like LMS, offering measurement potential. However, studies often over-rely on simple behavioral metrics like time-on-task, insufficiently addressing cognitive and affective dimensions (Bergdahl et al., 2024; Winter et al., 2024). Recent work shows selected LMS traces (e.g., forum and quiz data) can model these broader engagement strands, moving beyond mere participation (Ahmadi et al., 2023; Winter et al., 2024), though context remains vital.

Time spent on activities and time-based traces (durations AND session counts) also show predictive behavior for engagement and achievement, but only when they are interpreted cautiously and paired with other behavioral/relational indicators. Bergdahl et al. (2024) concluded that learning-analytics studies continue to use extensively behavioral counts – especially clicks and time-on-task – and under-represent

Table 1. Characteristics of the initially enrolled group (n = 400)

Participants' characteristic	Category	Percentage (%)
Gender	Female	68.0
	Male	32.0
Employment status	Permanent	87.3
	Substitute/contract	12.7
Teaching experience	≥ 16 years	68.8
	1–7 years	22.0
	8–15 years	9.2
Employment position/school type	Primary school	30.9
	Junior high school	28.9
	General senior high school	17.3
	Vocational senior high school	13.0
	Other (e.g., special education)	9.9
Educational attainment & training (multiple selections allowed)	MSc	68.3
	Level B advanced ICT certification	47.6
	MEd	14.7
	PhD	7.4

cognitive, affective, and social facets of engagement, which limit interpretability if time is used alone, i.e., time should be used in conjunction to other indicators.

Lately it is not raw duration that is focused on, but temporal profiles of participation. Goh (2025) modelled persistence (how often students engage across weeks) and consistency (how evenly they distribute that engagement), showing that both measures were correlated positively with performance, with effects strengthening late in the semester. This was interpreted as evidence that time distribution matters more than a single “hours online” total. Adding to this, Hernández-García et al. (2024) argue that simple time online measurements can be misleading, unless they are tied to the learner’s work context (e.g., preparatory vs. production vs. consolidation work).

But does higher engagement also mean higher individual performance? There are studies claiming so. For example, Hundhausen et al. (2023) report significant positive correlations between individuals’ GitHub contributions and their peer-evaluation ratings. Also, Roberts et al. (2024) report that visible, production-oriented engagement (e.g., edits/commits, issue activity, assessed forum work) shows small-to-moderate positive associations with peer-evaluated individual contribution.

Extending to teams, recent education/higher education evaluations highlight that assessment and team engagement are intertwined: assessment practices influence patterns of participation and vice versa, e.g., peer assessment frameworks used in groups tend to increase accountability and ongoing participation (Alizadeh et al., 2024; Gunning et al., 2022). These evaluations do not limit “engagement” to time alone, but reinforce that visible, trackable effort is part of how colleagues and instructors assess contribution.

METHODOLOGY

Source of the Data

LAMS is an open-source platform, allowing for the graphic creation of sequences of learning activities integrating all kinds of multimedia content and various innovative types of interactions both among the learners and the teachers and among the learners themselves. LAMS then allows for execution of the sequences of activities, collaboration, communication for trainees and control of the flow per learner, marking and getting LA reports for instructors. LAMS has also built-in templates for sequences following pedagogical concepts such as TBL.

The data used for this study originates from a training program for informatics schoolteachers, held using the LAMS platform under the title “LAMS in action: Innovation in practice”. The training program was organized by secondary education teachers’ counselors of different regions all over the country following the distant learning, TBL, and flipped classrooms principles.

At the beginning of the program, 400 trainees enrolled. **Table 1** shows that most of this first group were women (68.0%) and permanent staff (87.3%). Also, most trainees had a lot of experience, with 68.8% teaching

for 16 years or more, and a high level of education, with 68.3% having a Master of Science degree. Please note that these percentages are only for the whole starting group. In the end, we only analyzed 190 participants. We do not have the exact demographic data linked just to these 190 people. As we explain later in the Results section, the individuals we removed from the study spent less time on the platform. Because this removal was not random, the final 190 people might have somewhat different percentages than the starting group. For this reason, we provide [Table 1](#) just to give a general picture of the trainees, and we did not use these personal details as variables in our statistical models.

Objectives of the training program were as follows:

- a. To learn how to use the LAMS to improve the quality of teaching by integrating innovative instructional practices, through the design and implementation of lessons in LAMS, their pilot application in the classroom, the revision and enhancement of the lessons through appropriate support and feedback, and the exchange of experiences.
- b. To share the developed lessons under creative commons licenses in subject-specific repositories (per educational level) of computer science educational resources.

The main points of the educational methodology were as follows:

- a. The training program was organized voluntarily by both experienced and less experienced secondary education teachers' counselors and computer science teachers, employing the flipped classroom model, a blended learning (combined asynchronous and synchronous distance learning modes) and TBL methodologies.
- b. It offered, through an open invitation, the opportunity for a large group within the computer science teaching community to participate actively, drawing on their prior experiences, needs, and expectations as stimuli for engagement. The training program employed crowdsourcing and a bottom-up approach for the improvement of computer science teaching and the facilitation of classroom practices.

Thematic units were as follows:

- a. Introduction to LAMS – Presentation of the LAMS environment, repository, and indexes.
- b. LAMS – Practice as an author. Supervisor's environment, lesson design in a .docx template with the prospect of implementation in LAMS. Other tools in LAMS (optional, supportive, branches, grouping, gateways, etc.).
- c. Classroom application – Creating a class and enrolling students. Adding a lesson into the LAMS platform of the Greek school network (sch.gr). Depending on LAMS, differentiated instruction, and collaborative learning.
- d. International community – Uploading a lesson to the international repository. Sharing in the OER index and computer science courses OER index.

The planned duration of the seminar was 12 weeks (January to March 2025), while the trainees were given a requirement to spend no less than 20 hours. During the training program period, the learners were divided into teams of 3 to 7 people, supervised by 1 to 2 instructors. There were in total 13 very competent informatics teachers who assumed the role of team instructor. Both trainees and instructors were all volunteers, and all teaching and learning process was held outside of working hours.

Ethical review and formal committee approval were not required for this study in accordance with national legislation and institutional guidelines (see Ethics Declaration) as this work is carried out within the externally unfunded doctoral research of the first author. All participants in this study were adult professionals (informatics teachers). Prior to the training, participants were verbally informed by their trainers that if they chose to sign up, their data would be used anonymously for scientific research purposes. Consequently, voluntary registration and participation in the training constituted informed consent. To ensure data privacy and confidentiality, access to the collected data within the LAMS platform was strictly restricted to the primary investigators.

The Improved Project Contribution Score

Katsenos and Pierrakeas (2025) introduced the concept of a project contribution score (PCS) for summing up the peer evaluations of contributions of everyone in the team, as perceived by his/her team-mates. The initial PCS was based on a forced-ranking methodology, calculating a student's contribution through an algorithm that considered ordinal rank positions, number of evaluated tasks, and team size. However, the original PCS presented two main limitations:

- (a) it did not allow for equal scores, forcing students to rank teammates with identical contributions and
- (b) it assumed equal mathematical distances between all ranking positions.

To overcome these constraints, the IPCS uses the point-distribution method. This allows evaluators to assign equal scores when contributions are indistinguishable and more accurately captures the proportional differences between team members' efforts. Furthermore, compared to other efforts (Fellenz, 2006), IPCS uses a cumulative ranking approach to allow learners to rank the total contribution of their colleagues according to criteria and weights that they themselves give, which ultimately are what give the final sense of contribution to the group result.

The IPCS is calculated following the steps below: Each person "i" (rater) distributes marks to his/her "j" team-mates (ratees), rating each one's contribution to team tasks. The sum of points given is fixed (e.g., 20) and then transformation into the scale 0–1 is done, by dividing with the sum of points each person provided for his/her colleagues in the team. Each Rater can rate up to $n-1$ ratees where n is the team size. Therefore, if adding all individual normalized ratings r_{ij} provided from the same rater for all different ratees in his/her team should equal to one, i.e., $\sum_{j=1}^{n-1} r_{ij} = 1$.

For each rated person (ratee) j , his/her IPCS is calculated by summing all ratings he/she has received and divided by the maximum possible number of ratings he/she could receive in his/her team ($(n-1)$ if n is the team size), i.e., using the formula: $IPCS_j = \frac{\sum_{i=1}^a r_{ij}}{n-1}$, where a is the number of ratings a person received and $1 \leq a \leq n-1$.

It should be noted that normalization to $(n-1)$ makes the $IPCS_j$ comparable across all teams and accounts for team size. At the same time, of course, this division by $(n-1)$ implies that if someone should have received $(n-1)$ ratings but received fewer, then the remaining ones are effectively considered as zero. After this step the $IPCS_j$ of every rated person (ratee) is produced.

Note that

- a. it is not obligatory for someone to provide a peer evaluation for all others, he/she can allocate his/her marks to any or all the team members of his/her team,
- b. one can allocate same number of marks to different persons, equally classifying their contribution to a position,
- c. differences in allocated marks reflect the differences in contribution,
- d. LAMS provides a relevant activity for allowing each trainee to distribute marks to his/her colleagues, and
- e. confidentiality was preserved, as no-one could see the others' marks through LAMS and researchers did not make this information available.

Time Spent in Sequences of Activities ('TimeSpent')

The LAMS platform records detailed log data of user interactions, including timestamps for every activity. These include start and end times, as well as time duration per trainee per activity ('TimeSpent'), in milliseconds. This duration refers to the time taken for an activity to be completed and includes idle time within each activity. Educators can view reports on the duration students spend on specific learning materials, such as content pages, videos, assignments, or forums. This helps identify modules that might be too difficult or not engaging enough.

As LAMS provides time data per activity, summing up the activity durations, results in removal of the idle time between activities. Moreover, this is the easiest measurement to be taken, without the need to write

Table 2. Indicative part of the post-processing dataset used in this study

Trainee	Team ID	Team size	'IPCS'	'TimeSpent' in hours
T11U1	11	7	0.1835	100.5
T11U5	11	7	0.1573	99.7
T11U6	11	7	0.2418	398.2
T11U7	11	7	0.2432	242.6
T14U2	14	5	0.2937	969.3
T14U3	14	5	0.3532	83.3
T14U4	14	5	0.3532	223.4
...	...	...	...	...

Table 3. Descriptive statistics for IPCS score and 'TimeSpent'

	n	Mean	Standard deviation	Minimum	Q1	Median	Q3	Maximum	IQR
IPCS score	190	0.297	0.115	0.089	0.214	0.270	0.353	0.812	0.139
'TimeSpent' (h)	190	262	207	0.500	151	207	306	1,180	155

Note. IQR: Interquartile range

code or use special tools, so we aimed to provide the trainers using LAMS the easiest and fastest way to collect objective data, being aware of its limitations as a trainee engagement index. Therefore, in this study we computed the overall duration for each LAMS activity sequence, which is different from the login time to the LAMS platform, as the sum of all partial time durations for all core activities of the training program and a single indicator 'TimeSpent' was formed. 'TimeSpent' on LAMS activities sequences resembles more of a time-to-completion indicator found in literature and was used as a fast behavioural engagement indicator and an inferred measure of the engagement of each student.

Export of time duration data is possible through the "gradebook" functionality in LAMS and is accessible without direct accessing the LMS database, which has the advantage to be easy to be extracted without coding and the need of specialized tools. Because the durations are in the range of milliseconds, a conversion of the durations in hours was done to improve understanding.

RESULTS

Initially, 400 trainees enrolled in the training program, forming 72 teams. During follow-up, 73 trainees discontinued participation at various stages and were excluded from the analysis. Among the 327 participants who completed the program, 137 (42%) did not receive any peer reviews due to the voluntary nature of the evaluation process, resulting in default zero values. To avoid artificially distorting the statistical model, these 137 participants were excluded from our sample. A Wilcoxon rank-sum test revealed that these excluded trainees exhibited significantly lower baseline engagement, spending less time on the platform (median $\approx$ 170 hours) compared to the included trainees (median $\approx$ 207 hours), $p < 0.001$.

The final analytical sample consisted of 190 trainees who completed the program and for whom peer rankings were available, allowing computation of the IPCS.

Pseudo-anonymization of the data was done by assigning to each trainee a code name of the form T<team ID>U<user ID>, where team ID was retrieved by LAMS log and user was assigned randomly among the team members of each team. Thus, e.g., T36U6 is the sixth user of team 36.

In total, 683 peer rankings were collected. After aggregation to compute the individual peer contribution score (IPCS), this yielded IPCS values for 190 distinct trainees, who were distributed across 50 teams.

Time spent on different LAMS sequences of activities was retrieved from LAMS logs for these persons, converted to the hours scale and summed to calculate "TimeSpent (h)" in various LAMS sequences.

Ultimately a data table was constructed, which had the format shown in [Table 2](#).

Results Associated With RQ1

Table 3 summarizes statistics for IPCS scores received by ratees and for 'TimeSpent' (time spent on LAMS sequences).

Table 4. Statistical tests for normality, skewness, Shapiro-Wilk test (W and p-values), and Anderson-Darling test results

Variable	n	Skewness	Excess kurtosis	Shapiro W	Shapiro p	AD_A	AD_p
'IPCS'	190	1.40	2.66	0.89	$< 2 \times 10^{-10}$	5.18	$< 8 \times 10^{-13}$
'TimeSpent'	190	2.19	5.85	0.78	$< 2 \times 10^{-15}$	11.20	$< 4 \times 10^{-24}$

Further testing to the data to examine normality (Shapiro-Wilk and Anderson-Darling tests), skewness and kurtosis revealed that both variables (IPCS and 'TimeSpent'), deviate from normality (IPCS: $W = 0.948$, $p < 10^{-6}$; 'TimeSpent': $W = 0.786$, $p < 10^{-15}$) with positive skewness and heavy tails ('TimeSpent' skewness = 2.15, excess kurtosis = 5.6). The results of the above tests are summarized in [Table 4](#).

Since the normality tests suggest significant deviation from normality, we used Spearman's rank correlation as the primary measure of association between IPCS and 'TimeSpent'. Spearman's ρ was calculated $\rho = -0.148$ ($p = 0.04$ asymptotic), indicating a weak negative monotonic association, i.e., as IPCS increases 'TimeSpent' tends to be lower in rank (and vice-versa). p -value 0.04 (< 0.05) implies modest evidence against the null hypothesis of non-monotonic relationship. Clearly, there is a real but small inverse trend that higher IPCS is associated with slightly lower 'TimeSpent' and because Spearman correlation method is rank based, this finding is robust to skew/outliers and does not assume linearity or normality. Ordinary Linear Regression tested initially to examine whether 'TimeSpent' predicts IPCS showed tiny predictive power of 'TimeSpent' for IPCS, as the slope was found negative at -1.03×10^{-4} , ($p = 0.01$) but $R^2 = 0.035$ (adjusted $R^2 = 0.030$) indicates poor model fit, thus analysis was continued using more robust methods.

We deepened our analysis to examine how time spent on tasks ('TimeSpent') relates to peer-evaluated collaboration quality (IPCS) using generalized linear mixed models (GLMMs) to test linearity assumptions while accounting for team clustering. Results showed no significant linear relationship ($\beta = -0.015$, $p = 0.341$), with the smooth term confirming linearity was appropriate (edf = 1.0). Substantial between-team heterogeneity existed (random intercept $\sigma = 0.295$), indicating team-level factors strongly influence IPCS. The later applied GAMM yielded consistent conclusions ($\beta = -0.027$, $p = 0.221$), validating the linear specification and robustness of null findings across modelling approaches.

The k-means cluster analysis ($k = 3$, as inspection of the within cluster sum of squares "elbow" plot showed a clear inflection at this value, beyond which additional clusters yielded only marginal improvement) of scaled IPCS and 'TimeSpent', which followed, revealed three distinct groupings with moderate overlap, explaining meaningful variance in trainee peer score-engagement patterns. [Figure 1](#) visualizes the three identified clusters. Cluster 3 ($n \approx 113$) represents a 'low-low' profile with mostly below average scores on both dimensions. Cluster 2 ($n \approx 53$) shows mostly above-average IPCS scores coupled with slightly below-average 'TimeSpent', suggesting efficient perceived contribution where less time investment yields superior peer evaluations. Cluster 1 ($n \approx 24$) demonstrates the opposite pattern: substantially above-average 'TimeSpent' paired with moderately below-average IPCS.

Results Associated With RQ2

Previous generalized additive mixed model (GAMM) focused on inference and found no significant effect of 'TimeSpent', though it did reveal significant random effects for teams. To explore potential complex, non-linear interactions and to simultaneously examine the effects of categorical team membership ('teamID') and team size ('teamSize'), we employed random forest (RF) regression, a machine learning approach well-suited to modest sample sizes like ours (though larger datasets would strengthen inference).

We trained an RF model to predict individual IPCS scores from three predictors: 'TimeSpent', 'teamID', and 'teamSize', evaluated via 5-fold cross-validation repeated 10 times (50 resampling iterations total). This resampling strategy provided robust out-of-sample performance estimates. The model's hyperparameters were optimized by minimizing root mean squared error (RMSE). The optimized RF model achieved an RMSE of 0.0792, $R^2 = 0.5461$, and MAE = 0.0519, indicating that approximately 55% of the variance in IPCS scores is explained by these three predictors, with average prediction errors of 0.05–0.08 IPCS units. The relatively stable performance across evaluated hyperparameter configurations suggests a robust balance between bias and variance. These results demonstrate that RF captures meaningful predictive structure in the data beyond what linear inference methods detected, with 'teamID' emerging as the dominant predictor, followed by

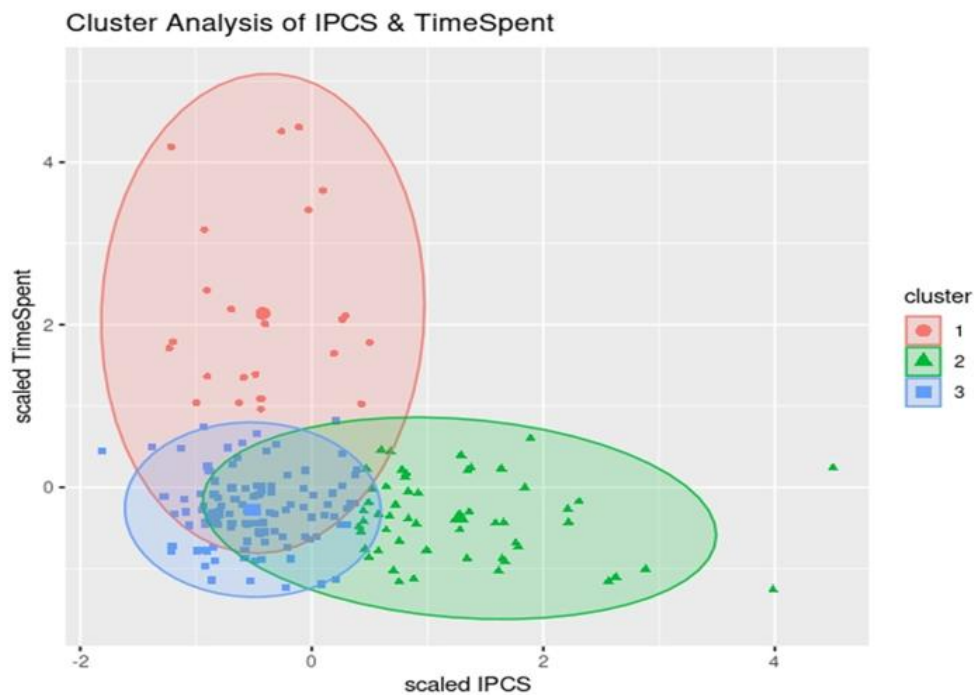


Figure 1. Scaled variables' cluster analysis visualization (the authors' own elaboration)

'teamSize' and 'TimeSpent', a finding that aligns with and extends the mixed-model inference by quantifying their relative importance in a non-linear, predictive framework.

Analysis of the importance of the three used predictor variables was performed to determine the relevant importance of the three examined factors. Typically, variable importance in RF is computed by randomly permuting each predictor and measuring the resulting increase in prediction error (e.g., MSE/RMSE), averaging this effect across all trees and resamples.

Figure 2 shows a plot of the relative importance of 'teamID', 'teamSize', and 'TimeSpent' in IPCS prediction. Y-axis (Predictor): Lists the three predictors 'TimeSpent', 'teamSize', and 'teamID'. Each bar represents the total importance of that predictor. X-axis (relative importance %): Shows the percentage derived from the RF's permutation-based measure (increase in prediction error when that predictor is permuted). Larger values mean a larger increase in error and thus greater influence on IPCS predictions.

'teamID' appears much more important, indicating that variation between teams explains most of the predictable variation in IPCS as permuting 'teamID' causes the largest degradation in predictive accuracy. 'teamSize' has a clear but much smaller relative importance, implying a secondary role: it improves prediction beyond 'teamID', but its effect on RMSE is modest. Finally, in consistence to previous observations, 'TimeSpent' has a very small relative importance, suggesting that, conditional on 'teamID' and 'teamSize', it contributes little to reducing RMSE and dropping or permuting 'TimeSpent' barely changes prediction. Statistically, this implies strong between-team heterogeneity and relatively weak within-team prediction based solely on logged time, with IPCS behaving almost like a team-level effect plus minor adjustment for team size.

DISCUSSION

Our analysis of 190 trainees across 50 project teams revealed no consistent relationship between time spent on tasks and peer-evaluated collaboration quality. While trainees varied considerably in their baseline performance levels, time investment neither universally improved nor diminished peer evaluations. Both linear (GLMM) and flexible (GAMM) modeling approaches converged on this null finding, with no evidence of optimal time thresholds or diminishing returns. This suggests the perceived by others individual contribution might depend more on how team members use their time together rather than the quantity alone, emphasizing pedagogical focus should be directed to interaction quality over time monitoring.

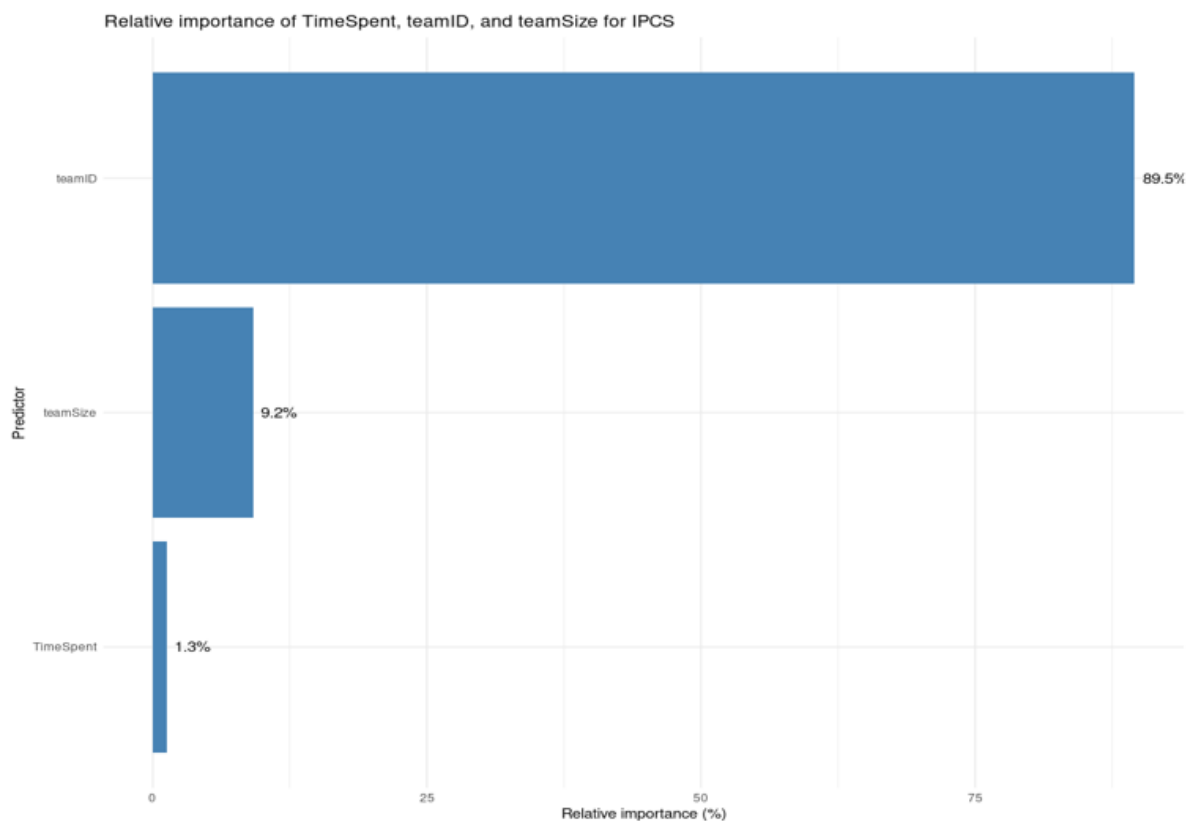


Figure 2. Relative importance of examined predictors in IPCS (the authors' own elaboration)

The analysis using a popular machine learning algorithm, which followed, confirmed the dominance of the factor to which team a trainee belongs (expressed by the 'teamID' variable) indicating which team a learner belongs to, and the team size are far more informative about perceived contribution (IPCS) than the recorded time the individuals spent on team activities. This implies that team size, culture, task structure, leadership, or peer norms likely drive much of the contribution pattern. The moderate importance of team size suggests that structural features of teams (e.g., smaller vs larger groups) have some systematic impact on how contributions are distributed and perceived. The minimal importance of 'TimeSpent' likely suggests that "time on the platform" alone is a poor proxy for meaningful contribution in this setting; learners who contribute more are not necessarily those who spend more logged time, highlighting the need to complement time-based analytics with richer indicators of engagement (quality of input, role distribution, interaction patterns) when designing assessment and feedback for collaborative learning.

The cluster analysis of IPCS and 'TimeSpent' (scaled) reveals three distinct engagement profiles among the 190 participants, highlighting heterogeneity in how time investment relates to perceived contribution. Cluster 1 (red and upper-left region) comprises individuals with moderate to high scaled 'TimeSpent' but not widely varying and close to average IPCS, suggesting a group where logged activity does not consistently translate into recognized contribution (it could reflect passive presence, lower-quality interactions, or roles with less visible output). Cluster 2 (green, right side) captures learners with low to moderate 'TimeSpent' yet consistently moderate to high IPCS, indicating efficient contributors whose impact exceeds their logged time (could be potentially strategic participants, experienced peers, or those in leadership roles). Cluster 3 (blue, lower-left) consists of low-time, low-contribution individuals, which may include disengaged learners or those whose contributions were limited by external constraints.

These patterns underscore that the relationship between time spent and contribution is not uniform across the cohort. While the RF analysis demonstrated that 'TimeSpent' has minimal predictive power when conditioned on team membership and size, the clustering reveals that unconditional bivariate segmentation exposes latent subpopulations with qualitatively different engagement dynamics. This divergence implies that team-level effects (as captured by 'teamID's dominance in the RF) may mask within-team or cross-team

heterogeneity in how individuals leverage their time. From an educational standpoint, recognizing these distinct profiles can inform differentiated support strategies: high-time/variable-contribution learners (cluster 1) may benefit from guidance on effective collaboration practices, while high-contribution/low-time learners (cluster 2) could serve as peer mentors or exemplary participation. The cluster structure also highlights the limitations of using time-on-platform as a proxy for meaningful engagement and reinforces the need for multidimensional analytics that incorporate interaction quality, peer feedback, and role distribution alongside quantitative time metrics.

As far as **RQ1** is concerned, we conclude that the observed correlation for the overall sample between peer score ('IPCS') and the time spent on team activities ('TimeSpent') is negligible and that team effects are far more important (which is in line with literature). However, we argue for **RQ2**, that the combination of time spent on activities and peer assessment can provide useful information to the educator. Based on the different profiles for the combination of these two measures as revealed by the cluster analysis, heterogeneities in the combination of engagement and individual performance are revealed and therefore the types of support that should be provided to individuals with different profiles can be adapted accordingly.

The sample size of 190 participants across 50 teams, while adequate for initial exploration, is modest for machine learning applications and may limit generalizability imposing a limitation on this study. Similarly, generalizability of the outcomes is limited by the facts that platform-logged time is a proxy for cognitive engagement and may not capturing off-platform collaboration or effort quality, peer-evaluated IPCS scores are subject to bias and team-specific norms not explicitly modeled, the dominance of 'teamID', while lacking mechanistic explanation, points to unmeasured here constructs like team culture, leadership, and communication patterns and cluster analysis is exploratory and requires external validation.

Another methodological limitation of this study relates to how the RF model handles team-level data. In our RF model, we included 'teamID' as a standard categorical variable. While this is a practical and widely accepted solution in educational data mining to capture team differences, it does not fully model the hierarchical, nested structure of the data (i.e., participants grouped within teams) in the way a mixed-effects RF would. Because of this constraint, the variable importance estimates produced by the RF model should be viewed as complementary evidence rather than absolute proof. For the main conclusions of this paper, we primarily rely on the results from the GLMM and GAMM models. These models are specifically designed to appropriately handle random effects and nested data, making them the most suitable tools for our statistical inferences.

Finally, the exclusion of 42% of the participants from the IPCS calculations because they did not receive peer evaluations imposes one more limitation. Our analysis indicates a non-random exclusion pattern, as these excluded students spent significantly less time on the platform. Consequently, our sample is subject to selection bias, skewing toward more visibly engaged students. However, because our main analysis demonstrates that 'TimeSpent' does not hold predictive power over IPCS scores, this specific bias is unlikely to act as a confounding variable that invalidates the IPCS findings. Still, future implementations should consider treating non-evaluated students as an additional distinct cluster to further investigate the peer dynamics of low-engagement participants.

Even with the above challenges in mind, this study complements recent research (Achواني et al., 2025; Adesina et al., 2023; Fleckney et al., 2025; Nurismawati et al., 2025; Zhang & Gao, 2024) because

- (a) we treat peer assessment (IPCS) as the outcome of team work and we combine it with objective engagement traces (time spent in LAMS activities) and team structure,
- (b) using GLMM/GAMM plus RF and clustering, we found that time spent on activities has negligible predictive power for peer-rated contribution, once team effects are accounted for, in this lifelong learning context and we aligned this phenomenon with peer assessment, not just test scores, and
- (c) we identified team membership and team size as dominant explanatory factors and showed that cluster-based profiles (high-time/average-IPCS, low-time/high-IPCS and low-time/low-IPCS) could exist at a deeper level that translate directly into differentiated support strategies.

Finally, by combining peer assessment and engagement within LAMS-supported group work, our article fills a sectoral gap and applies patterns reported mainly in university coursework to competency-based, practice-oriented training programs.

Future research should then employ larger, longitudinal datasets across multiple lifelong learning contexts to validate findings and establish causality, the mechanisms underlying engagement patterns and peer perception could be illuminated by qualitative interviews with trainees in each cluster profile. Also, direct measurement of interaction quality, communication frequency, role clarity, and team climate would replace latent 'teamID' effects with interpretable constructs. Most interestingly, future research could apply the investigation described here, in a formative way at a team-task level to identify the support needed by trainees and teams while teamwork is ongoing to support them to achieve better results.

CONCLUSION

This article makes a valuable contribution by quantifying peer evaluation into a ranking score (IPCS) for team-members and ultimately use it as their outcome of teamwork. We combined it with objective engagement traces (time spent in LAMS activities) and team structure. Using GLMM/GAMM plus RF and clustering, we concluded that time spent on activities has negligible predictive power for peer-ranked contribution, once team effects are accounted for, in lifelong learning context. We also aligned this phenomenon with peer assessment scores, not just test scores as accustomed to. Additionally, we identified team membership and team size as dominant explanatory factors and showed that cluster-based profiles (high-time/average-IPCS, low-time/high-IPCS, low-time/low-IPCS) could exist at deeper level and this might serve to identify differentiated support strategies. Finally, the application of peer assessment and engagement within LAMS-supported group work, a sectoral gap is filled, and our article applies patterns reported mainly in university coursework to competency-based, practice-oriented training programs.

Author contributions: **IK:** conceptualization, investigation, methodology, writing – original draft, writing – review & editing; **SP:** resources, supervision, writing – review & editing; **CP:** funding acquisition, supervision, writing – review & editing. All authors approved the final version of the article.

Funding: The publication fees of this manuscript have been financed by the Research Council of the University of Patras (approval decision 1041_23.09.2026).

Ethics declaration: Ethical review and formal committee approval were not required for this study in accordance with national legislation and institutional guidelines. Under Greek law 4957/2022 (article 279, paragraph 2b) and the regulations of the Research Ethics Committee of the University of Patras, formal ethics approval is not required for doctoral thesis research unfunded by external sources. All participants in this study were adult professionals (informatics teachers). Prior to the training, participants were verbally informed by their trainers that if they chose to sign up, their data would be used anonymously for scientific research purposes. Consequently, voluntary registration and participation in the training constituted informed consent. To ensure data privacy and confidentiality, access to the collected data within the LAMS platform was strictly restricted to the primary investigators.

AI statement: During the preparation of this manuscript, the authors used gemini.google.com and perplexity.ai primarily for linguistic assistance and language enhancement to improve readability, as the authors are non-native English speakers. Secondly, these tools were used as assistive search engines to locate relevant information and literature on the web and in scientific journals. No personal, sensitive, or participant data were uploaded to any AI systems. After using these tools/services, the authors verified all AI-generated suggestions and retrieved literature against primary sources, reviewed and edited the content as needed, and took full responsibility for the accuracy, originality, and content of the published article.

Declaration of interest: The authors declared no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

- Achواني, A. S., Rahman, T., & Rusyati, L. (2025). Exploring the relationship between student engagement and academic achievement in science: A study of junior high school students. *Journal of Environment and Sustainability Education*, 3(3), 323-330. <https://doi.org/10.62672/joease.v3i3.116>
- Adesina, O. O., Adesina, O. A., Adelopo, I., & Afrifa, G. A. (2023). Managing group work: The impact of peer assessment on student engagement. *Accounting Education*, 32(1), 90-113. <https://doi.org/10.1080/09639284.2022.2034023>

- Ahmadi, G., Mohammadi, A., Asadzandi, S., Shah, M., & Mojtahedzadeh, R. (2023). What are the indicators of student engagement in learning management systems? A systematized review of the literature. *The International Review of Research in Open and Distributed Learning*, 24(1), 117-136. <https://doi.org/10.19173/irrodl.v24i1.6453>
- Alizadeh, M., Masoomi, R., Mafinejad, M. K., Parmelee, D., Khalaf, R. J., & Norouzi, A. (2024). Team-based learning in health professions education: An umbrella review. *BMC Medical Education*, 24(1), Article 1131. <https://doi.org/10.1186/s12909-024-06147-x>
- Baker, D. F. (2008). Peer assessment in small groups: A comparison of methods. *Journal of Management Education*, 32(2), 183-209. <https://doi.org/10.1177/1052562907310489>
- Bergdahl, N., Bond, M., Sjöberg, J., Dougherty, M., & Oxley, E. (2024). Unpacking student engagement in higher education learning analytics: A systematic review. *International Journal of Educational Technology in Higher Education*, 21(1), Article 63. <https://doi.org/10.1186/s41239-024-00493-y>
- Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, 98(4), 891-901. <https://doi.org/10.1037/0022-0663.98.4.891>
- Cohen, S. (1997). What makes teams work: Group effectiveness research from the shop floor to the executive suite. *Journal of Management*, 23(3), 239-290. [https://doi.org/10.1016/S0149-2063\(97\)90034-9](https://doi.org/10.1016/S0149-2063(97)90034-9)
- Davison, H. K., Mishra, V., Bing, M. N., & Frink, D. D. (2014). How individual performance affects variability of peer evaluations in classroom teams. *Journal of Management Education*, 38(1), 43-85. <https://doi.org/10.1177/1052562912475286>
- Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment. *Psychological Science in the Public Interest*, 5(3), 69-106. <https://doi.org/10.1111/j.1529-1006.2004.00018.x>
- Fellenz, M. R. (2006). Toward fairness in assessing student groupwork: A protocol for peer evaluation of individual contributions. *Journal of Management Education*, 30, 570-591. <https://doi.org/10.1177/1052562906286713>
- Fete, M. G., Haight, R. C., Clapp, P., & McCollum, M. (2017). Peer evaluation instrument development, administration, and assessment in a team-based learning curriculum. *American Journal of Pharmaceutical Education*, 81(4), Article 68. <https://doi.org/10.5688/ajpe81468>
- Fleckney, P., Thompson, J., & Vaz-Serra, P. (2025). Designing effective peer assessment processes in higher education: A systematic review. *Higher Education Research & Development*, 44(2), 386-401. <https://doi.org/10.1080/07294360.2024.2407083>
- Goh, T.-T. (2025). Learning management system log analytics: The role of persistence and consistency of engagement behaviour on academic success. *Journal of Computers in Education*, 13, 283-306. <https://doi.org/10.1007/s40692-025-00358-x>
- Gunning, T. K., Conlan, X. A., Collins, P. K., Bellgrove, A., Antle, K., Cardilini, A. P. A., & Fraser, C. L. (2022). Who engaged in the team-based assessment? Leveraging EdTech for a self and intra-team peer-assessment solution to free-riding. *International Journal of Educational Technology in Higher Education*, 19(1), Article 38. <https://doi.org/10.1186/s41239-022-00340-y>
- Hernández-García, Á., Cuenca-Enrique, C., Del-Río-Carazo, L., & Iglesias-Pradas, S. (2024). Exploring the relationship between LMS interactions and academic performance: A learning cycle approach. *Computers in Human Behavior*, 155, Article 108183. <https://doi.org/10.1016/j.chb.2024.108183>
- Hundhausen, C., Conrad, P., Adesope, O., & Tariq, A. (2023). Combining GitHub, chat, and peer evaluation data to assess individual contributions to team software development projects. *ACM Transactions on Computing Education*, 23(3), Article 33. <https://doi.org/10.1145/3593592>
- Johnson, D. W., Johnson, R. T., & Smith, K. A. (1991). *Cooperative learning: Increasing college faculty instructional productivity*. School of Education and Human Development, George Washington University. <https://eric.ed.gov/?id=ED343465>
- Kozłowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, 7(3), 77-124. <https://doi.org/10.1111/j.1529-1006.2006.00030.x>
- Katsenos, I., & Pierrakeas, C. (2025). Assessing individual contributions of team members to team achievement by combining peer assessments and digital presence in an academic environment. *Education Sciences*, 15(3), Article 279. <https://doi.org/10.3390/educsci15030279>

- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121-1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
- Mathieu, J. E., Hollenbeck, J. R., van Knippenberg, D., & Ilgen, D. R. (2017). A century of work teams in the *Journal of Applied Psychology*. *Journal of Applied Psychology*, 102(3), 452-467. <https://doi.org/10.1037/apl0000128>
- Mathieu, J. E., Tannenbaum, S. I., Kukenberger, M. R., Donsbach, J. S., & Alliger, G. M. (2015). Team role experience and orientation: A measure and tests of construct validity. *Group and Organization Management*, 40(1), 6-34. <https://doi.org/10.1177/1059601114562000>
- Nurismawati, R., Rusyati, L., Husna, H. N., & Hidayanti, W. (2025). Exploring the relations between peer assessment, self-assessment, and academic achievement in school science experiment course. *Journal of Innovation in Educational and Cultural Research*, 6(3), 657-665. <https://doi.org/10.46843/jiecr.v6i3.2269>
- Roberts, C., Jorm, C., Nisbet, G., Gordon, C. J., Chen, T., Held, F., & Haq, I. (2024). Video-based peer assessment of collaborative teamwork in a large-scale interprofessional learning activity. *BMC Medical Education*, 24(1), Article 1307. <https://doi.org/10.1186/s12909-024-06124-4>
- Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a "big five" in teamwork? *Small Group Research*, 36(5), 555-599. <https://doi.org/10.1177/1046496405277134>
- Shi, W. W., Jagannadharao, A., Lee, J., & Bailey, B. P. (2021). Challenges and opportunities for data-centric peer evaluation tools for teamwork. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), Article 432. <https://doi.org/10.1145/3479576>
- Speckbacher, G., & Wiernsperger, M. (2025). Peer evaluations in diverse teams: How external validation of team performance influences ingroup favoritism. *Accounting, Organizations and Society*, 114, Article 101595. <https://doi.org/10.1016/j.aos.2025.101595>
- Stonewall, J., Dorneich, M., Rongerude, J., & Dorius, C. (2018). A review of bias in peer assessment. In *Proceedings of the CoNECD 2018 – Collaborative Network for Engineering and Computing Diversity Conference*. <https://doi.org/10.18260/1-2--29510>
- Sundstrom, E., De Meuse, K. E., & Futrell, D. (1990). Work teams: Applications and effectiveness. *American Psychologist*, 45(2), 120-133. <https://doi.org/10.1037/0003-066X.45.2.120>
- Sung, Y., Chang, K., Chang, T., & Yu, W. (2010). How many heads are better than one? The reliability and validity of teenagers' self- and peer assessments. *Journal of Adolescence*, 33(1), 135-145. <https://doi.org/10.1016/j.adolescence.2009.04.004>
- Tasi, P., Scheider, A.-C., & Osipovskaya, E. (2025). Evaluating strategies to increase students' accountability in entrepreneurship education. In *Proceedings of the New Perspectives in Science Education 2025*. https://conference.pixel-online.net/library_scheda.php?id_abs=7394&id_crs=3&id_edition=48&id_authors=8150&type=authors
- Tong, Y., Schunn, C. D., & Wang, H. (2023). Why increasing the number of raters only helps sometimes: Reliability and validity of peer assessment across tasks of different complexity. *Studies in Educational Evaluation*, 76, Article 101233. <https://doi.org/10.1016/j.stueduc.2022.101233>
- Winter, M., Mordel, J., Mendzheritskaya, J., Biedermann, D., Ciordas-Hertel, G.-P., Hahnel, C., Bengs, D., Wolter, I., Goldhammer, F., Drachsler, H., Artelt, C., & Horz, H. (2024). Behavioral trace data in an online learning environment as indicators of learning engagement in university students. *Frontiers in Psychology*, 15. <https://doi.org/10.3389/fpsyg.2024.1396881>
- Zhang, Y., & Gao, Y. (2024). Exploring the dynamics of student engagement with receiving peer feedback in L2 writing. *Assessing Writing*, 60, Article 100842. <https://doi.org/10.1016/j.asw.2024.100842>

