# Differences between self-assessment and peer assessment in oral presentations in higher education

*Diferencias entre autoevaluación y coevaluación en presentaciones orales en educación superior*


## Authors

Francisco José Rodríguez Rojas [1]
Antonio Jesús Sánchez-Oliver [2]
Moisés Grimaldi-Puyana [3]

[1] Universidad Internacional de la Rioja
[2, 3] Universidad de Sevilla, España

Corresponding author:
Antonio Jesús Sanchez-Oliver
sanchezoliver@us.es



Received: 24-03-26
Accepted: 07-05-26

## How to cite in APA

Rodríguez Rojas, F. J., Sanchez-Oliver, A. J., & Grimaldi-Puyana, M. (2026). Differences between self-assessment and peer assessment in oral presentations in higher education. *Retos, 81*, 132-143. https://doi.org/10.47197/retos.v81.119096



## Abstract

Introduction and Objective: This study examined the differences between self-assessment and peer-assessment in the evaluation of theoretical–practical oral presentations in higher education.
Methodology: A total of 7,776 assessments were collected from 108 undergraduate students who evaluated their own presentations (n = 216) and those of their peers (n = 7,560) using a validated ten-item rubric.
Results: Results showed that self-assessment scores were consistently higher than peer-assessment scores across all dimensions, with the largest discrepancies observed in delivery, activity design, content organisation, time distribution, and classroom management. Discussion: Peer-assessment displayed stronger internal coherence and clearer differentiation between items, suggesting a more consistent application of performance criteria. In contrast, self-assessment exhibited substantial inflation and greater heterogeneity, particularly in complex or multifaceted dimensions. Differences were smallest in teamwork and oral/non-verbal communication, indicating that some competencies may be more challenging to judge objectively. While the study is limited to a single instructional context, its findings highlight the value of integrating structured peer-assessment to support calibration, enhance evaluative judgement, and foster more reliable formative assessment practices.
Conclusion: Overall, the results emphasise the importance of combining multiple assessment modalities to strengthen students' understanding of quality and promote meaningful engagement with evaluation criteria.



## Keywords

Rubric; formative assessment; triadic assessment; shared assessment; university; physical education.



## Resumen

Introducción y Objetivo: Este estudio examinó las diferencias entre la autoevaluación y la coevaluación (evaluación entre pares) en la valoración de presentaciones orales teórico-prácticas en la educación superior.
Metodología: Se recopilaron un total de 7.776 evaluaciones de 108 estudiantes de grado, quienes evaluaron sus propias presentaciones (n = 216) y las de sus compañeros (n = 7.560) utilizando una rúbrica validada de diez ítems.
Resultados: Los resultados mostraron que las puntuaciones de autoevaluación fueron consistentemente más altas que las de coevaluación en todas las dimensiones, observándose las mayores discrepancias en la exposición, el diseño de actividades, la organización del contenido, la distribución del tiempo y la gestión del aula.
Discusión: La coevaluación mostró una mayor coherencia interna y una diferenciación más clara entre los ítems, lo que sugiere una aplicación más consistente de los criterios de desempeño. Por el contrario, la autoevaluación exhibió una inflación sustancial y una mayor heterogeneidad, particularmente en dimensiones complejas o polifacéticas. Las diferencias fueron menores en el trabajo en equipo y en la comunicación oral/no verbal, lo que indica que algunas competencias pueden ser más difíciles de juzgar objetivamente. Aunque el estudio se limita a un único contexto instruccional, sus hallazgos resaltan el valor de integrar la coevaluación estructurada para apoyar la calibración, mejorar el juicio evaluativo y fomentar prácticas de evaluación formativa más fiables.
Conclusión: En general, los resultados enfatizan la importancia de combinar múltiples modalidades de evaluación para fortalecer la comprensión de la calidad por parte de los estudiantes y promover un compromiso significativo con los criterios de evaluación.



## Palabras clave

Rúbrica; evaluación formativa; evaluación formativa triádica; evaluación compartida; universidad; educación física.

## Introduction

Formative assessment is widely recognised as a key mechanism for improving learning in higher education (Sánchez-Oliver et al., 2025). It supports students in identifying strengths and weaknesses, regulating their learning, and refining their performance over time (Black & Wiliam, 1998). Recent work in higher education further emphasises that formative assessment is most effective when it creates structured opportunities to act on feedback and engage actively with evaluative processes (Morris et al., 2021). As universities shift toward student-centred pedagogies, formative practices are expected to promote metacognition, critical thinking, and continuous improvement—competencies essential for academic and professional development (Parmigiani et al., 2025). Yet, many institutions still struggle to implement these practices systematically, often due to summative-dominant cultures and limited assessment literacy (Panadero et al., 2026).

Within this context, self-assessment and peer-assessment have become central strategies for involving students directly in the evaluation of learning (Sánchez-García et al., 2026). Self-assessment encourages students to judge their work against explicit criteria and generate "internal feedback," but research consistently shows that accuracy varies and overestimation is common when students lack calibration or external reference points (Alemdag & Narciss, 2025). From a formative perspective, self-assessment involves more than assigning a score to one's own performance; it requires students to actively monitor, regulate, and judge their learning in relation to explicit criteria. Contemporary research highlights the central role of metacognitive processes—such as planning, monitoring, and evaluation—in shaping the quality and accuracy of self-assessment judgements (Rickey et al., 2025). When students engage meaningfully with these processes, self-assessment can support deeper learning, autonomy, and the development of evaluative judgement.

However, evidence consistently shows substantial variability in the accuracy of self-assessment. Meta-analytic findings indicate a systematic tendency toward overestimation, particularly in complex tasks, alongside wide individual differences (León et al., 2023). This variability has been associated with multiple factors, including prior achievement, self-efficacy, task complexity, assessment literacy, and motivational or emotional influences (Pinedo et al., 2025). Without adequate calibration and external reference points, students may struggle to apply performance criteria consistently to their own work, leading to inflated or unreliable judgements.

These findings suggest that self-assessment, while pedagogically valuable, is especially sensitive to contextual design and instructional support. Structured rubrics, exemplars, guided reflection, and systematic integration with peer or teacher assessment have been identified as key mechanisms for enhancing self-assessment accuracy and educational value (León et al., 2023; Pinedo et al., 2025; Rickey et al., 2025). Consequently, examining self-assessment alongside peer-assessment within the same authentic task offers an opportunity to better understand how students interpret criteria and how different evaluative modalities contribute to formative assessment processes in higher education.

Peer-assessment offers complementary benefits: it requires students to analyse their peers' work, discriminate quality, articulate criteria, and provide constructive feedback. Meta-analyses show that well-designed peer assessment can improve academic performance across subjects and levels (Double et al., 2020), and that assessor training is one of the strongest predictors of its success (Li et al., 2020). High-quality peer assessment typically includes clear rubrics, scaffolding tools such as worked examples and prompts, multiple review cycles, and purposeful links to self-assessment (Gielen & De Wever, 2015; Cheong et al., 2023).

Recent reviews reinforce that peer assessment is most effective when implemented as iterative formative feedback rather than as isolated grading exercises (Fleckney et al., 2024). Collaborative forms of peer review—where more than one assessor provides feedback—can reduce individual bias and improve feedback quality and engagement (Armengol-Asparó et al., 2022; Mandala et al., 2018). However, design choices matter. Research shows mixed effects for anonymity, which may increase psychological safety but does not guarantee better performance (Panadero & Alqassab, 2019). Online versus offline formats also produce small and context-dependent differences (Jongsma et al., 2023). Moreover, students' feedback literacy, and that of instructors, plays a crucial role in determining whether feedback is actually interpreted, valued, and used in subsequent learning (Carless & Boud, 2018; Carless & Winstone, 2020).

Peer-assessment also carries an important dialogic dimension. Qualitative studies show that students often value peer review as a space for reflection, comparison and shared sense-making (Ardill, 2025). Yet its effectiveness can be weakened by emotional concerns, perceived power dynamics, and limited competence in giving or interpreting feedback. These challenges highlight the need for constructive learning environments where peer assessment is accompanied by explicit guidance, dialogue, and practice.

Despite the maturity of the field, several gaps remain. Reviews note that higher-education research rarely examines formative assessment at scale or explores how students apply rubric criteria across complex performance tasks (Morris et al., 2021; Fleckney et al., 2024). Oral presentations represent one of these tasks. They require students to demonstrate multiple competencies simultaneously—content organisation, instructional design, communication skills, non-verbal language, material use, time management, classroom control, and teamwork. Because presentations combine theoretical and practical demands, they offer an ideal setting to analyse how students interpret rubric criteria and how self- and peer-assessment differ across specific dimensions. Yet few studies have examined item-level discrepancies or compared the internal structure of different assessment modalities using large datasets in authentic university settings.

Oral presentations constitute an especially suitable context for examining formative assessment processes in higher education. They are widely recognised as complex and authentic assessment tasks, as they require students to integrate and demonstrate multiple competencies simultaneously, including content organisation, instructional design, communication skills, time management, and classroom management (Boud & Falchikov, 2007; Nicol & Macfarlane-Dick, 2006). Such multidimensional tasks place high cognitive and metacognitive demands on learners and evaluators, making them particularly informative for analysing how assessment criteria are interpreted and applied in practice.

Within this type of task, the use of analytic rubrics plays a central role. Rubrics help make quality criteria explicit and observable, supporting transparency, consistency, and formative use of assessment information (Panadero & Jonsson, 2013). At the same time, prior research shows that applying rubric-based criteria to complex performances is challenging, especially for novice assessors, and may lead to variability and bias in judgements depending on the assessment modality (Jonsson & Panadero, 2017).

Accordingly, studying self-assessment and peer-assessment within the context of oral presentations provides a theoretically grounded opportunity to explore how students engage with evaluative criteria in authentic, ill-structured tasks, and how different assessment modalities shape judgement accuracy and internal coherence. This rationale complements the validation of the instrument used in the present study (Sánchez-Oliver et al., 2025) and situates oral presentations as a meaningful setting for investigating formative assessment practices in higher education.

Despite the substantial body of research on self-assessment and peer-assessment in higher education, important gaps remain. Previous studies have often examined these modalities in isolation, focused on narrow performance indicators, small samples, or single outcome scores (León et al., 2023; Pinedo et al., 2025; Rickey et al., 2025). As a result, we still know relatively little about how students apply assessment criteria across different dimensions of complex performance tasks, or how self- and peer-assessment differ not only in mean scores but also in their internal coherence and item-level structure (Rodríguez-Rojas et al., 2026; Sanchez-Oliver et al., 2026).

In particular, there is limited empirical evidence based on large datasets and authentic classroom contexts that systematically compares self-assessment and peer-assessment within the same task, using a shared rubric and analysing discrepancies across multiple performance dimensions. Moreover, few studies have explored whether differences between assessment modalities are uniform across rubric items or whether some competencies are more prone to inflation, bias, or inconsistent criteria application than others. Addressing these gaps, the present study compares self-assessment and peer-assessment in theoretical–practical oral presentations in higher education, examining (a) item-level and global differences between modalities, and (b) the internal correlation structure of each assessment method. By doing so, the study provides nuanced evidence on how students engage with evaluative criteria in complex tasks and contributes to a more precise understanding of the strengths and limitations of different formative assessment modalities.

## Method

### *Design*

This study employed a quantitative, cross-sectional, non-experimental design aimed at analysing university students' performance under two formative assessment modalities: self-assessment and peer assessment. The investigation focused on the systematic collection of evaluative data generated by students regarding both their own oral presentations and those of their classmates, using a validated rubric specifically designed for theoretical–practical presentations in higher education contexts. This approach aligns with the principles of formative, shared, and participatory assessment promoted within the European Higher Education Area.

### *Participants*

The sample consisted of 108 fourth-year students enrolled in the Primary Education degree with a specialization in Physical Education at the University of Seville (68 women and 40 men). Students were organised into 36 groups of three, each of which delivered two practical presentations, resulting in a total of 72 oral presentations during the course.

Across these presentations, a total of 7,776 assessment records were obtained. Of these, 216 corresponded to self-assessment, as each student evaluated their own performance after each of the two presentations (108 self-assessment × 2 presentations = 216). In contrast, 7,560 were peer-assessment ratings, consistent with the structure of the activity: in every presentation, the three members of the presenting group did not evaluate, whereas the remaining 105 students acted as peer evaluators (105 peer assessment × 72 presentations = 7,560).

This pronounced imbalance between modalities is inherent to the pedagogical design and reflects the collective involvement of the cohort in peer-based formative processes. Each evaluation entry was treated as an independent unit of analysis, as assessments pertained to different presentation events and evaluators.

### *Instruments*

This study employed the instrument validated by Sánchez-Oliver et al. (2025), specifically developed to appraise theoretical–practical presentations within the European Higher Education context. The rubric was co-constructed with active input from students and faculty and encompassed self-, peer-, and teacher-assessment modalities. Its design was anchored in a precise alignment between task objectives and content, the specification of observable indicators of work quality, and the definition of progressive achievement levels reflecting student performance. The rubric has shown adequate psychometric properties in its validation study. Evidence of internal consistency was obtained for both the theoretical (Cronbach's $\alpha = 0.88$) and practical scales (Cronbach's $\alpha = 0.90$). Exploratory and confirmatory factor analyses supported a unidimensional structure in each scale, with acceptable model fit indices (CFI ≥ 0.90) and explained variance ranging from 57% to 59.85%. These results support the reliability and construct validity of the instrument for formative assessment purposes in higher education contexts (Sánchez-Oliver et al., 2025).

The instrument to practical presentations comprises ten items targeting these dimensions: content, delivery, oral expression and non-verbal language, group distribution and organization, motor engagement time, selection and organization of materials, tasks/activities/games/exercises selected, time management, instructions and classroom control, and teamwork. Each item is rated on four ordered levels —4 = excellent, 3 = satisfactory, 2 = improvable, 1 = insufficient— accompanied by detailed descriptors of observable behaviours and outputs to ensure a transparent, structured evaluation process. Although numerical scores are produced, the rubric is primarily formative in intent; quantitative scoring is reserved for the end of the process.

### *Procedure*

Data collection took place within the framework of the IV Teaching Innovation Plan of the University of Seville (2023), through the project Formative Assessment: Self-Assessment, Peer Assessment, and

Teacher Assessment in University Teaching. The initiative involved faculty and students from the Faculty of Education and centred on developing students' oral communication skills as a core academic and professional competence.

Assessment occurred under two modalities: Self-assessment, each student evaluated their own performance immediately after delivering their group's presentation; and Peer assessment, all students present in the session evaluated the presenting group, except the three members of that group. This procedure ensured that 105 peers assessed each presentation, fostering a robust and shared evaluative environment.

All evaluations were administered digitally using Microsoft Forms, the institutional platform at the University of Seville. Following data collection, the dataset underwent thorough cleaning and verification to ensure accuracy and consistency. The assessment activities were fully embedded within regular teaching sessions during the 2024–2025 academic year, forming part of a broader pedagogical model grounded in triadic, shared, and bidirectional assessment. In this model, the rubric functioned simultaneously as a tool for documenting performance and as a scaffold for learning, reflection, and ongoing improvement.

Students did not receive formal or specific training focused exclusively on the use of the rubric or on structured assessment processes prior to the study. However, they had previously participated in self-assessment and peer-assessment activities in earlier courses within their degree programmes. Before data collection, the rubric was presented and explained in detail during class sessions to ensure students' understanding of the assessment criteria, as it formed part of the official evaluation system of the course. This approach reflects a typical formative assessment context embedded in regular teaching practice rather than an experimentally trained assessment setting.

### *Data Analysis*

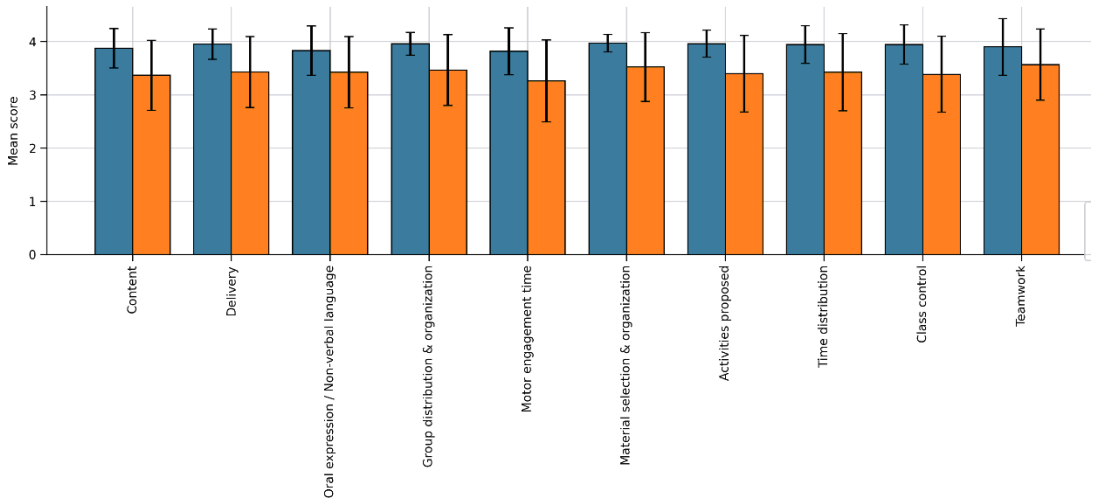
Given the ordinal nature of the rating scale (1–4), the marked ceiling effects, and the lack of pairing between self- and peer-assessments, the analysis employed two-tailed Mann–Whitney U tests to compare modalities for each rubric item and for the overall presentation score. Effect sizes were quantified using the rank-biserial correlation (r).

To evaluate the internal consistency of the instrument, both Cronbach's alpha and an ordinal approximation based on the mean Spearman inter-item correlation were calculated. All statistical analyses were conducted using Python, with appropriate libraries for non-parametric testing and reliability estimation.

## Results

A total of 7,776 assessment records were analysed, comprising 216 self-assessments and 7,560 peer assessments, derived from the 72 oral presentations delivered by the 108 participating students. Descriptive analyses showed consistently high ratings across all rubric dimensions, with a clear ceiling effect typical of formative assessment contexts. Notably, self-assessment scores exhibited markedly lower variability and were strongly concentrated in the highest performance categories, whereas peer assessment scores displayed wider dispersion and systematically lower mean values. Across the ten assessed dimensions, self-assessment scores were consistently higher than peer assessment scores. Self-ratings ranged from 3.82 to 3.97, while peer ratings varied between 3.26 and 3.57, indicating a pervasive pattern of upward bias in students' perceptions of their own performance. This trend was also reflected in the global score, with students assigning themselves a mean of 3.92, compared with 3.43 assigned by their peers. Figure 1 illustrates these trends, presenting mean scores ± standard deviations for each item.

Figure 1. Mean scores ± SD for Self-Assessment and Peer-Assessment




Given the ordinal 1–4 scale, skewed distributions, and independence of observations across modalities, differences between self- and peer assessment were examined using two tailed Mann–Whitney U tests. As shown in Table 1, all rubric dimensions —along with the overall score— showed statistically significant differences between modalities ($p < .001$ in every case). Effect sizes, calculated via the rank biserial correlation (r), ranged from –0.30 to –0.45 for individual items.

Table 1. Comparison Between Self - and Peer Assessment across Rubric Items and Overall Score

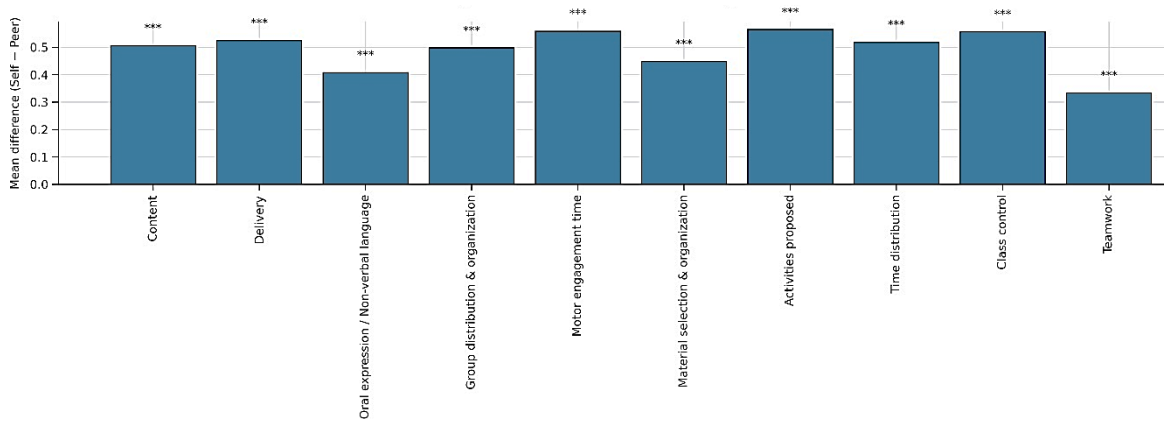
| Variable | Self-A M (SD) | Peer-A M (SD) | U | p | r |
|---|---|---|---|---|---|
| Content | 3.88 (0.37) | 3.37 (0.66) | 1,164,618 | < 0.001 | –0.43 |
| Delivery | 3.95 (0.29) | 3.43 (0.67) | 1,175,470 | < 0.001 | –0.44 |
| Oral expression / Non-verbal language | 3.83 (0.46) | 3.43 (0.67) | 1,092,390 | < 0.001 | –0.34 |
| Group distribution & organization | 3.96 (0.21) | 3.46 (0.67) | 1,151,931.5 | < 0.001 | –0.41 |
| Motor engagement time | 3.82 (0.44) | 3.26 (0.77) | 1,150,848 | < 0.001 | –0.41 |
| Material selection & organization | 3.97 (0.16) | 3.52 (0.65) | 1,117,650 | < 0.001 | –0.37 |
| Activities proposed | 3.96 (0.25) | 3.40 (0.72) | 1,182,331 | < 0.001 | –0.45 |
| Time distribution | 3.94 (0.36) | 3.43 (0.72) | 1,155,654 | < 0.001 | –0.42 |
| Class control | 3.94 (0.37) | 3.39 (0.71) | 1,187,849 | < 0.001 | –0.45 |
| Teamwork | 3.90 (0.53) | 3.57 (0.67) | 1,064,620.5 | < 0.001 | –0.30 |
| Overall score | 3.92 (0.30) | 3.43 (0.50) | 1,399,586.5 | $< 10^{-72}$ | –0.71 |

Data presenting mean scores (M) ± standard deviations (SD); Self A:self-assessments; Peer A: Peer assessment; Negative values of r indicate higher self-assessment scores. Significant differences, p< .05.

The largest divergences emerged in Delivery (r = –0.44), Activities proposed (r = –0.45), Class control (r = –0.45), Content (r = –0.43), and Time distribution (r = –0.42), revealing strong tendencies toward self-inflation in areas related to instructional organisation, clarity of presentation, and management of learning environments. By contrast, the smallest—yet still statistically significant—differences were observed in Teamwork (r = –0.30) and Oral expression and non verbal communication (r = –0.34), suggesting greater alignment between self- and peer ratings in dimensions involving interpersonal dynamics and communicative nuance. To visually summarise the magnitude and direction of these discrepancies, Figure 2 displays the mean differences (Self – Peer) for each rubric item.

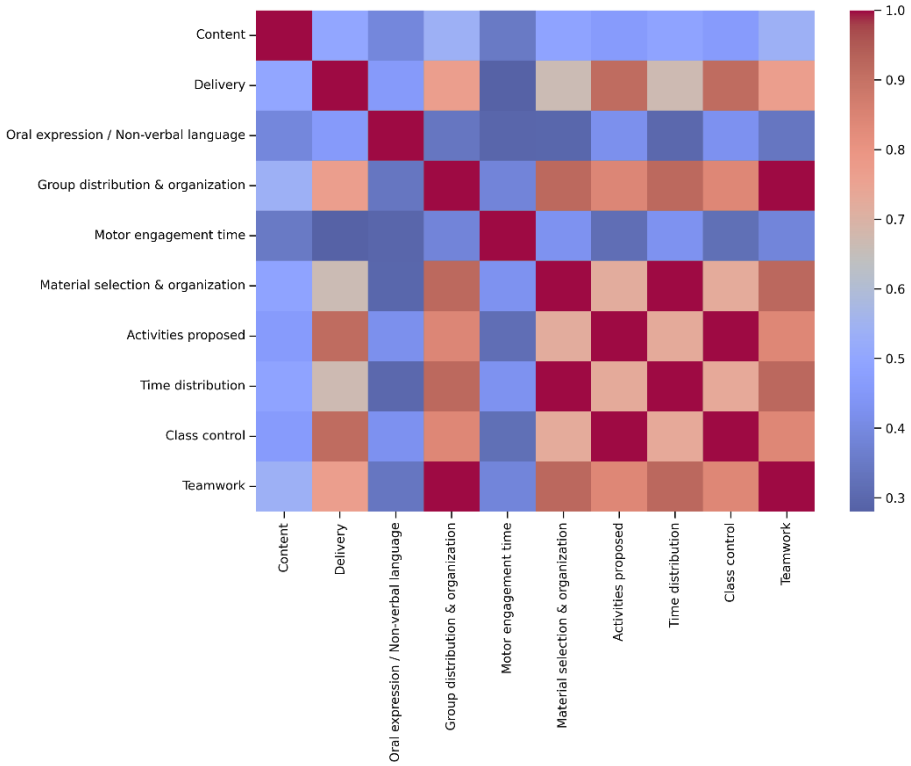
The global presentation score further reinforced the item level pattern: self-assessments produced a mean of 3.92±0.30, significantly higher than the 3.43±0.50 observed in peer assessments. The Mann–Whitney U statistic indicated a very large and highly significant difference ($U = 1{,}399{,}586.50$; $p < 10^{-72}$; $r = –0.71$), demonstrating that discrepancies accumulate across dimensions when ratings are synthesised into a single holistic indicator of performance.

Figure 2. Mean Differences (Self – Peer) with significance indicators



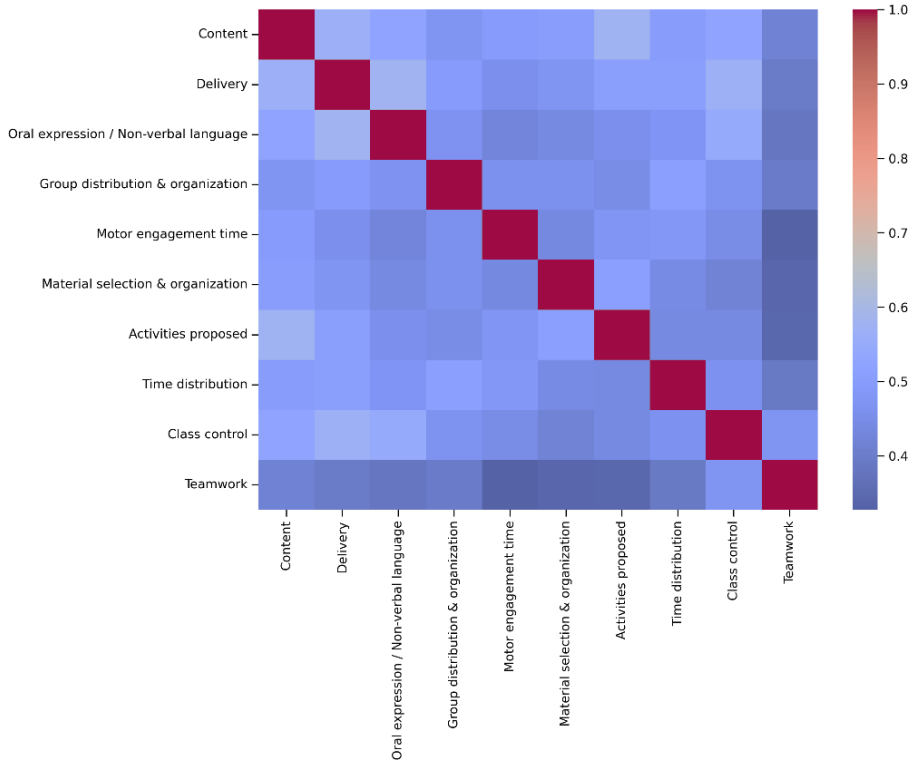

To examine the consistency and internal coherence of the rubric across modalities, Spearman correlation matrices were generated separately for self and peer assessments. As shown in Figure 3, self-assessment correlations were relatively low and heterogeneous, suggesting that students applied the rubric criteria less systematically when evaluating their own work.

Figure 3. Correlation Heatmap (Self-Assessment)




In contrast, Figure 4 illustrates the correlation structure for peer assessment, where inter item associations were stronger and more uniform. This pattern indicates a more stable and coherent interpretation of the rubric descriptors when students rated the work of others—an observation consistent with the higher internal consistency indices obtained under the peer modality.

Figure 4. Correlation Heatmap (Peer-Assessment)




## Discussion

This study examined how university students evaluate theoretical–practical oral presentations using self- and peer-assessment in authentic classroom conditions. Across all rubric dimensions, students scored themselves higher than their peers did. This pattern is consistent with prior work showing that self-judgements often lack calibration and tend toward optimism when benchmarks are absent (Morris et al., 2021; Alemdag & Narciss, 2025). These gaps may reflect limited assessment literacy, uneven interpretation of criteria, and the emotional load of judging one's own work.

The magnitude and consistency of the differences suggest the inflation of self-assessment is not limited to isolated skills. In our data, the largest gaps appeared in delivery, instructional structure, and classroom management—multidimensional areas that require fine-grained pedagogical judgement. Similar challenges are reported in the literature and point to a need for structured support when students appraise complex performances (Solis et al., 2025; Parmigiani et al., 2024). The large overall effect in our study indicates that small item-level divergences can cumulate into a meaningful global discrepancy, unless calibration is built into the process (Morris et al., 2021).

Peer-assessment, by contrast, yielded lower means but stronger internal coherence. This aligns with recent reviews: peer processes work best when implemented as formative, iterated, and anchored to clear criteria and assessor preparation (Fleckney et al., 2024; Li et al., 2020). Shared standards and basic scaffolding help student raters apply rubrics more consistently, which likely explains the tighter inter-item structure we observed for peer-ratings (Gielen & De Wever, 2015).

Design choices also matter. Collaborative peer feedback—more than one assessor per work—can reduce idiosyncratic bias and improve the quality and usefulness of comments, with knock-on benefits for engagement (Armengol-Asparó et al., 2022; Mandala et al., 2018). Dialogic feedback cultures and explicit development of feedback literacy make uptake more likely and help close the feedback loop (Carless & Boud, 2018; Carless & Winstone, 2020).

Item-level patterns fit this picture. The largest discrepancies clustered in observable dimensions (delivery, activities proposed, class control, content, time distribution), where peers can compare performances across groups; gaps were smaller for teamwork and oral/non-verbal communication. Clear standards, modelling, and worked examples tend to increase accuracy for structured criteria, while interpersonal or subtle communicative aspects remain harder to judge and benefit from guided exemplars and repeated practice (Gielen & De Wever, 2015; Li et al., 2020). These item-level differences may also be explained by the nature of the competencies assessed. Teamwork, as a collective and highly observable dimension, may be less prone to self-assessment inflation, as shared responsibility and mutual monitoring provide clearer reference points for judgement. In contrast, dimensions such as classroom management are more dynamic, individual, and context-dependent, which may increase the likelihood of overestimation in self-assessment. From this perspective, task complexity and the degree of shared accountability may help explain why some competencies show smaller discrepancies between assessment modalities.

Some implementation levers produced mixed effects in prior studies and should be used strategically. Anonymity may increase psychological safety but does not consistently improve performance or engagement; online versus offline delivery shows small, context-dependent differences (Panadero & Alqassab, 2019; Jongsma et al., 2023). These nuances suggest that format choices should be aligned with learning goals and cohort readiness rather than applied as blanket rules.

Our findings point to practical steps for courses using oral presentations. First, prepare student assessors and provide explicit scaffolds—calibration tasks, rubrics, checklists, and worked examples—to tighten the link between criteria and judgements (Li et al., 2020; Mercader et al., 2020). Second, use iterative cycles and, where feasible, collaborative peer configurations to diversify perspectives and dampen individual bias (Armengol-Asparó et al., 2022; Mandala et al., 2018). Third, integrate peer- and self-assessment purposefully (for example, peer-review before self-assessment) to enhance calibration and performance (Cheong et al., 2023; Er et al., 2021). Finally, invest in feedback literacy for students and teachers to ensure that feedback information is understood, valued, and used (Carless & Boud, 2018; Carless & Winstone, 2020).

Several limitations should be acknowledged when interpreting the findings of this study. First, although the dataset is large, the assessment observations are not fully independent. Multiple ratings were provided by the same students and were clustered around the same oral presentations, reflecting the collective and authentic nature of the formative assessment design. While non-parametric tests were used to compare assessment modalities, the absence of statistical models explicitly accounting for data dependency (e.g., multilevel approaches) suggests that the results should be interpreted primarily in terms of patterns and tendencies rather than strict inferential generalisation. Second, there is a marked imbalance between assessment modalities, with substantially more peer-assessment records than self-assessment records. This asymmetry is inherent to the instructional design, as each presentation was evaluated by the whole cohort except for the presenting group. Although this structure strengthens the robustness of the peer-assessment estimates, it may also contribute to differences in score variability and dispersion between modalities. Third, a ceiling effect was observed across most rubric dimensions, with high mean scores and reduced variability, particularly in self-assessment. This effect is common in formative assessment contexts using analytic rubrics and may limit the sensitivity of the instrument to detect finer performance differences at higher levels of achievement. Nevertheless, the consistency of the differences between modalities across all items and the magnitude of the observed effects support the robustness of the findings within the studied educational context. Moreover, the authentic setting and large number of ratings strengthen ecological validity and match current recommendations to embed structured formative practices in real courses (Solis Trujillo et al., 2025; Morris et al., 2021).

In short, self-assessment alone tends to over-estimate performance in complex oral tasks, while peer-assessment—when designed and supported—offers a more coherent external perspective. A thoughtful combination of both, wrapped in explicit standards, calibration, and opportunities for dialogue, is a practical route to fairer, more reliable, and more educationally valuable assessment in higher education (Fleckney et al., 2024; Li et al., 2020).

## Conclusions

This study shows that students systematically rate their oral presentations more positively than their peers, with the largest discrepancies emerging in delivery, activity design, content organisation, and class management. Peer-assessment displayed greater internal coherence, suggesting more consistent application of criteria, while self-assessment revealed inflation across all dimensions. Smaller differences in teamwork and oral communication indicate that some competencies are harder to judge objectively. Although limited to a single context, the findings highlight the value of combining peer and self-assessment to strengthen evaluative judgement. The results underscore the importance of integrating well-designed peer-assessment processes to complement and calibrate self-assessment, enhance evaluative judgement, and strengthen the reliability of formative assessment systems in higher education.

## Acknowledgements

We sincerely thank the teaching staff and students of the Primary Education degree programs at the University of Seville for their valuable collaboration. Their active participation was essential for the development of communicative competence and the improvement of formative assessment. We also thank the University of Seville for its support through the IV Teaching Innovation Plan, which made this project possible.

## Financing

This article derives from a project funded by the IV Teaching Innovation Plan of the University of Seville, under the 2023 call, specifically within activities for teaching staff aimed at educational innovation (L.2.2. of the IV Teaching Innovation Plan), particularly Action 221, focused on supporting teaching innovation. This article is part of the doctoral thesis of Francisco José Rodríguez Rojas.

## References

Alemdag, E., & Narciss, S. (2025). Promoting formative self-assessment through peer assessment: peer work quality matters for writing performance and internal feedback generation. *International Journal of Educational Technology in Higher Education, 22*, 22. https://doi.org/10.1186/s41239-025-00522-4

Ardill, N. (2025). Peer feedback in higher education: student perceptions of peer review and strategies for learning enhancement. *European Journal of Higher Education, 15*(4), 696–721. https://doi.org/10.1080/21568235.2025.2457466

Armengol Asparó, C., Mercader, C., & Ion, G. (2022). Making peer feedback more efficient: What conditions of its delivery make the difference? *Higher Education Research & Development, 41*(2), 226–239. https://doi.org/10.1080/07294360.2020.1840527

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: principles, policy & practice, 5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Carless, D., & Winstone, N. (2020). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*, 1–14. https://doi.org/10.1080/13562517.2020.1782372

Cheong, C. M., Luo, N., Zhu, X. H., Lu, Q., & Wei, W. (2023). Self-assessment complements peer assessment for undergraduate students in an academic writing task. *Assessment & Evaluation in Higher Education, 48*(1), 135–148. https://doi.org/10.1080/02602938.2022.2069225

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review, 32*(2), 481–509. https://doi.org/10.1007/s10648-019-09510-3

Er, E., Dimitriadis, Y., & Gašević, D. (2021). A collaborative learning approach to dialogic peer feedback: A theoretical framework. *Assessment & Evaluation in Higher Education, 46*(4), 586–600. https://doi.org/10.1080/02602938.2020.1786497

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

Gielen, M., & De Wever, B. (2015). Scripting the role of assessor and assessee in peer assessment in a wiki environment: Impact on peer feedback quality and product improvement. *Computers & Education, 88*, 370–386. https://doi.org/10.1016/j.compedu.2015.07.012

Jongsma, M. V., Scholten, D. J., van Muijlwijk Koezen, J. E., & Meeter, M. (2023). Online versus offline peer feedback in higher education: A meta-analysis. *Journal of Educational Computing Research, 61*(2), 329–354. https://doi.org/10.1177/07356331221114181

Li, H. L., Xiong, Y., Hunte, C. V., Guo, X. Y., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education, 45*(2), 193–211. https://doi.org/10.1080/02602938.2019.1620679

Mandala, M., Schunn, C., Dow, S., Goldberg, M., Pearlman, J., Clark, W., & Mena, I. (2018). Impact of collaborative team peer review on the quality of feedback in engineering design projects. *International Journal of Engineering Education, 34*(4), 1299–1313.

Mercader, C., Ion, G., & Díaz Vicario, A. (2020). Factors influencing students' peer feedback uptake: Instructional design matters. *Assessment & Evaluation in Higher Education, 45*(8), 1169–1180. https://doi.org/10.1080/02602938.2020.1726283

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education, 9*, e3292. https://doi.org/10.1002/rev3.3292

Panadero, E., & Alqassab, M. H. (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. https://doi.org/10.1080/02602938.2019.1600186

Panadero, E., Amezua-Urrutia, A., Fernández Ruiz, J., Rodríguez-Hernández, C., & Fraile, J. (2026). Assessment practices in Spanish universities: a nationwide update seven years on. *Assessment & Evaluation in Higher Education*, 1–21. https://doi.org/10.1080/02602938.2026.2621276

Parmigiani, D., Nicchia, E., Murgia, E., & Ingersoll, M. (2024). Formative assessment in higher education: An exploratory study within programs for professionals in education. *Frontiers in Education, 9*, 1366215. https://doi.org/10.3389/feduc.2024.1366215

Parmigiani, D., Nicchia, E., Murgia, E., & Ingersoll, M. (2025). Learning with peers in higher education: Exploring strengths and weaknesses of formative assessment. *Trends in Higher Education, 4*(3), 48. https://doi.org/10.3390/higheredu4030048

Pinedo, L., Panadero, E., Fernández-Ruiz, J., & García-Pérez, D. (2025). Beyond the Accuracy Gap in Self-assessment: Exploring Reasons for (In)Accuracy and the Role of Individual Differences. *Assessment in Education: Principles, Policy & Practice, 32*(4), 462–484. https://doi.org/10.1080/0969594X.2025.2565384

Rickey, N., Panadero, E. & DeLuca, C. (2025). How do students self-assess? examining the metacognitive processes of student self-assessment. *Metacognition Learning 20*, 25. https://doi.org/10.1007/s11409-025-09430-4

Rodríguez-Rojas, F., Sánchez Oliver, A. J., Muñoz Llerena, A., Angosto, S., Muñoz López, A., & Grimaldi-Puyana, M. (2026). Evaluation of Theoretical Presentations in University Students of Sport Science: Differences Between Self-Assessment, Peer Assessment, and Teacher Assessment. *SPORT TK–EuroAmerican Journal of Sport Sciences. In press*

Sánchez-García, A., Sanchez-Oliver, A. J., & Grimaldi-Puyana, M. (2026). Triadic formative assessment in higher education: interaction effects between assessment modality and evaluator sex. *Retos, 77*, 346-357. https://doi.org/10.47197/retos.v77.118382

Sánchez Oliver, A. J., Rodríguez Rojas, F. J., Feria Madueño, A., Muñoz López, A., Carnero Díaz, Ángel, Muñoz Llerena, A., Sañudo Corrales, B., Oviedo Caro, M. Ángel, Grimaldi Puyana, M., Bianchi, P., Domínguez, R., & Angosto, S. (2025). Validación de una herramienta para evaluar las presentaciones teórico-prácticas en la Educación Superior. *Retos, 67*, 903–916. https://doi.org/10.47197/retos.v67.112987

Sánchez-Oliver, A. J., Sánchez-García, A., Feria-Madueño, A., Muñoz-López, A., Carnero Díaz, Á., Muñoz-Llerena, A., Sañudo-Corrales, B., Oviedo-Caro, M. Á., Grimaldi-Puyana, M., Bianchi, P., & Domínguez, R. (2026). Evaluating oral presentations in university students: Self, peer, and hetero evaluation. *Journal of Sport and Health Research, In press*.

Solis Trujillo, B.P., Velarde-Camaqui, D., Gonzales Nuñez, C.A., Castillo Silva, E.V. & Gonzalez Said de la Oliva, M.P. (2025) The current landscape of formative assessment and feedback in graduate studies: a systematic literature review. *Frontiers in Education, 10*, 1509983. https://doi.org/10.3389/feduc.2025.1509983

## Authors' and translators' details:


| | | |
|---|---|---|
| Francisco José Rodríguez Rojas | fran.rr.frr@gmail.com | Author |
| Antonio Jesús Sánchez-Oliver | sanchezoliver@us.es | Author |
| Moisés Grimaldi-Puyana | mgrimaldi@us.es | Author |