Triadic formative assessment in higher education: interaction effects between assessment modality and evaluator sex

Evaluación formativa triádica en la educación superior: efectos de la interacción entre la modalidad de evaluación y el sexo del evaluador

Authors

Andrés Sánchez-García [1]
Antonio Jesús Sanchez-Oliver [2]
Moisés Grimaldi-Puyana [1]

[1, 2] Universidad de Sevilla, Sevilla (España)

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

Received: 17-12-25
Accepted: 27-01-26

How to cite in APA

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

Abstract

Introduction: Formative assessment has become a central pillar in higher education, as it promotes deep learning and active student engagement. Within this framework, triadic assessment—self-assessment, peer assessment, and teacher assessment—enables the incorporation of multiple perspectives. However, limited attention has been paid to how these assessment modalities interact with evaluator-related variables, such as sex, and its interaction.
Objective: This study aimed to examine differences among formative assessment modalities and to explore whether differential effects emerge according to the sex of the evaluator in a university context.
Methodology: A cross-sectional, non-experimental quantitative design was employed. The sample comprised 5,671 assessment records collected in the Primary Education degree (Physical Education specialization) at the University of Seville. A validated rubric was used to assess practical presentations through self-assessment, peer assessment, and teacher assessment. Data analysis included descriptive statistics and non-parametric tests (Kruskal–Wallis and Mann–Whitney U).
Results: The results indicated that self-assessment and peer assessment consistently yielded higher scores than teacher assessment across most items (p<.001), suggesting a systematic effect of assessment modality. Overall sex-related differences were generally small; however, interaction analyses revealed that differences according to evaluator sex emerged primarily within teacher assessment, with male evaluators tending to assign higher scores in dimensions such as content, materials, and teamwork.
Discussion and conclusions: These findings suggest that sex-related differences in assessment are not uniform across modalities but may become more salient when evaluation is conducted by an external expert, underscoring the importance of gender-aware calibration and reflexive assessment practices in higher education.

Keywords

Formative assessment; self-assessment; peer assessment; teacher assessment; Physical Education.

Resumen

Introducción: La evaluación formativa es un pilar central en la educación superior, ya que promueve la participación activa del alumnado. Dentro de este marco, la evaluación triádica —autoevaluación, evaluación entre pares y heteroevaluación— permite la incorporación de múltiples perspectivas. Sin embargo, se ha prestado poca atención a cómo estas modalidades de evaluación interactúan con variables relacionadas con el evaluador, como el sexo.
Objetivo: Este estudio tuvo como objetivo examinar las diferencias entre modalidades de evaluación formativa y explorar si surgen efectos diferenciales según el sexo del evaluador.
Metodología: Se empleó un diseño cuantitativo transversal y no experimental. La muestra comprendía 5.671 evaluaciones recogidas en el grado de Educación Primaria (especialización en Educación Física en la Universidad de Sevilla). Se utilizó una rúbrica validada para evaluar presentaciones prácticas mediante autoevaluación, evaluación entre pares y heteroevaluación. El análisis de datos incluyó estadística descriptiva y pruebas no paramétricas (Kruskal–Wallis y Mann–Whitney U).
Resultados: Los resultados indicaron que la autoevaluación y la evaluación entre pares arrojaron consistentemente puntuaciones más altas que la evaluación del profesorado en la mayoría de los ítems (p<.001), lo que sugiere un efecto sistemático de la modalidad de la evaluación. Las diferencias generales relacionadas con el sexo fueron generalmente pequeñas; sin embargo, los análisis de interacción revelaron que las diferencias según el sexo del evaluador surgieron principalmente en la evaluación del profesor, con los evaluadores masculinos tendiendo a asignar puntuaciones más altas en dimensiones como contenido, materiales y trabajo en equipo.
Discusión y conclusiones: Estos hallazgos sugieren que las diferencias relacionadas con el sexo no son uniformes entre modalidades, pero pueden volverse más relevantes cuando la evaluación la realiza un experto externo, subrayando la importancia de la calibración y las prácticas de evaluación reflexiva y conscientes del género en la educación superior.

Palabras clave

Evaluación formativa; autoevaluación; evaluación entre pares; evaluación del profesor; Educación Física.

Introduction

Formative assessment has become a key component of the teaching–learning process in higher education, as it orients instruction toward continuous improvement and meaningful feedback for students (Olague et al., 2025). A growing body of research highlights that, compared with purely summative assessment, formative assessment contributes to deeper learning and greater development of academic competencies among students (Barrientos et al., 2019; Romero-Martín et al., 2017). Barrientos et al. (2019) emphasize the importance of implementing formative strategies in higher education, demonstrating benefits in both academic performance and student motivation. Similarly, studies conducted in the Spanish university context show that both teachers and students positively value continuous and participatory assessment, as it promotes pedagogical interaction and improvement in the teaching process (Romero-Martín et al., 2017). Overall, there is broad consensus regarding the relevance of formative assessment in higher education as a means of fostering higher-quality learning and solid academic performance (Martín & Alcalá, 2022). Nevertheless, it is pertinent to proceed with the improvement of assessment strategies so as to address the existing challenges and promote training that integrates more dialogic assessment spaces (Gallardo-Fuentes et al., 2025).

The concept of shared assessment refers to the importance of students' participation in the assessment process, mainly through techniques such as self-assessment, peer assessment, and assessment shared between teachers and students (Bustamante et al., 2025). Within this framework, triadic assessment has gained relevance. This approach is understood as the combination of self-assessment, peer assessment, and teacher assessment within evaluative processes. Several authors also refer to it as formative and shared assessment, as it involves the active participation of all agents: the student, their peers, and the teacher (Pérez-Pueyo & López-Pastor, 2017; Mendoza et al., 2021). Self-assessment enables students to reflect on their own performance and take responsibility for their learning, while peer assessment—assessment among equals—promotes collaborative learning and the development of constructive critical thinking (Zubillaga-Olague et al., 2025). Teacher assessment, in turn, remains essential for guiding the process through an expert perspective. Previous research has shown that combining these three assessment modalities provides a more comprehensive view of learning and enhances students' understanding of assessment criteria (Salom & Tena, 2022). Furthermore, the implementation of triadic assessment in higher education has been associated with innovative pedagogical practices and a shift in the student's role—from a passive recipient of grades to an active agent in their own formative process. In addition, as pointed out by Estaji (2024), it is essential to strengthen knowledge about the tools and instruments that allow this type of assessment to be carried out in a more well-grounded, objective, and effective manner.

Despite these advantages, the literature suggests that students' actual participation in assessment processes remains moderate. For example, Fernández-Garcimartín et al. (2022) found that, in teacher education programs, evaluative practices continue to be dominated by traditional teacher assessment, with self-assessment and peer assessment being implemented less frequently than expected. This evidence points to the need to further promote more participatory assessment cultures in higher education (Zubillaga-Olagüe & Cañadas, 2022). Formative assessment is a means to support the development of students' competencies by linking theory and practice through reflection. Assessment processes should not only serve to grade students' results, but also help them develop competencies applicable beyond the university classroom (Barba-Martín et al., 2020).

Another aspect that remains underexplored, yet highly relevant, concerns sex-related differences in university assessment processes, particularly within formative assessment contexts. Educational research has documented certain gender gaps in overall academic performance; however, it remains unclear whether such differences influence how male and female students perceive or benefit from formative assessment. Findings in this area are inconclusive. On the one hand, some studies suggest that no significant differences exist in the perception or effectiveness of formative assessment between students of different sexes (Romero et al., 2017; Fernández-Garcimartín et al., 2022). In these studies, both male and female students report positive evaluations of continuous feedback and opportunities for participation, showing similar levels of satisfaction and average performance. On the other hand, emerging research points to potential sex-related variations in specific contexts. For instance, González-Gutiérrez

et al. (2024) found among Physical Education students that male students tended to perceive the attainment of good grades more easily than female students under continuous assessment systems, whereas female students showed slight advantages in certain qualitative aspects of the feedback received. Likewise, a recent study conducted in the Ecuadorian university context revealed differences in how male and female students value formative assessment practices, suggesting that female students may display a more favourable attitude toward feedback and continuous improvement (Esquivel Rivero et al., 2025). Using mixed-methods approaches, this study concluded that sex-based perspectives warrant attention, as certain needs or preferences related to formative assessment may differ between groups.

Despite extensive evidence supporting formative and shared assessment, little is known about how evaluator-related variables, such as sex, interact with assessment modality to shape performance judgments. Addressing this issue is essential for advancing equitable assessment practices in higher education. Although formative and shared assessment has been widely examined over recent decades, most research has focused on demonstrating its general benefits or describing implementation experiences (Barrientos-Hernán & López-Pastor, 2015). Far fewer studies have explored how student-related factors, such as sex, may interact with these evaluative practices to influence academic performance. Existing findings regarding sex-based differences are partial and sometimes contradictory, highlighting an unresolved issue in the literature. Authors such as Barba-Martín and Hortigüela-Alcalá (2022) advocate for student involvement in assessment, arguing that "if assessment is learning, students must be part of it"; however, they do not address whether such involvement has equivalent effects across all students. Similarly, Zubillaga-Olagüe and Cañadas (2022) emphasize the participation of both teachers and students in formative assessment, without examining demographic differences in that participation. Consequently, there is a need to investigate whether triadic assessment exerts differential effects according to students' sex in higher education, thereby addressing a relevant theoretical and practical gap.

Method

Research Design

This study adopted a cross-sectional, non-experimental design with a quantitative approach, aimed at analyzing university students' academic performance through different formative assessment modalities (self-assessment, peer assessment, and teacher assessment). The research focused on the systematic collection of assessments conducted by students themselves, their peers, and teaching staff, using a validated instrument designed to evaluate theoretical–practical presentations in higher education. The instrument was applied within the practical content of the course Physical Conditioning.

Participants

The sample consisted of a total of 5,671 assessment records completed by students enrolled in the Primary Education degree with a specialization in Physical Education at the University of Seville. Each assessment record was considered an independent unit of analysis, as evaluations corresponded to different presentation instances and were completed by different evaluators across assessment modalities. The assessments were collected within the course Physical Conditioning in Schools and were obtained in authentic classroom settings embedded in formative assessment processes. The variable "sex" was used in accordance with institutional records and prior literature in educational assessment, acknowledging its binary categorization as a limitation.

Regarding sex distribution, 56.4% of the assessments corresponded to female students (n = 3,202), while 43.4% were completed by male students (n = 2,462). In terms of assessment modality, peer assessment predominated (n = 4,832; 85.2%), followed by self-assessment (n = 599; 10.6%) and teacher assessment (n = 233; 4.1%). The decreasing number of assessments across modalities is explained by the smaller number of evaluators involved in teacher assessment compared with peer assessment, and in self-assessment compared with peer assessment.

Instruments

The study employed the instrument validated by Sánchez-Oliver et al. (2025), specifically designed to assess theoretical-practical presentations within the context of the European Higher Education (EHE).

In addition, the methodological framework proposed by these authors was adopted. This framework is based on a rubric that integrates clearly defined assessment criteria, differentiated levels of achievement, and a mixed (qualitative and quantitative) approach, applicable to both formative assessment processes and grading.

The rubric was developed with the active participation of both students and teaching staff and incorporated self-assessment, peer assessment, and teacher assessment modalities. Its design was grounded in the precise definition of objectives and content linked to academic tasks, identifying specific indicators to assess the quality of student work and establishing progressive levels of achievement reflecting student performance.

The instrument comprised 10 items designed to evaluate key aspects of theoretical-practical presentations: content, delivery, oral expression/non-verbal language, group distribution and organization, time of motor engagement, selection and organization of materials, tasks/activities/games/exercises selected, time management, instructions and classroom control, and teamwork. For each item, four levels of achievement were defined: excellent (4), satisfactory (3), improvable (2), and insufficient (1). Each level included detailed and precise descriptions of observable characteristics in students' outputs and behaviours, ensuring a clear and structured evaluation process. Although the rubric generated numerical values, its design prioritized formative purposes, with scoring occurring only at the end of the process.

Procedure

Data collection was conducted as part of an educational innovation project funded by the IV Teaching Innovation Plan of the University of Seville (2023), entitled Formative Assessment: Self-Assessment, Peer Assessment, and Teacher Assessment in University Teaching. The project involved both teachers and students from the Primary Education and Sport Sciences programs at the University of Seville and focused on oral communication as a transversal and fundamental competence.

Assessments were carried out using three modalities: teacher assessment (teachers evaluating the groups delivering the presentations), self-assessment (students evaluating their own performance), and peer assessment (students evaluating their classmates' presentations). All responses were collected using Microsoft Forms, provided by the University of Seville. Subsequently, the database was cleaned and refined to ensure data quality and reliability prior to statistical analysis.

Data collection focused on the development of oral communication competence, specifically through practical presentations delivered during class sessions in the 2024–2025 academic year. This approach enabled the implementation of a triadic, shared, and bidirectional assessment system, in which the rubric functioned both as a tool for recording achievement levels and as a mechanism for promoting comprehensive student learning, thus integrating assessment as a formative process rather than a purely summative one.

Data Analysis

Data analysis was conducted using the Statistical Package for the Social Sciences (SPSS, version 29.0), under the institutional license of the University of Seville. First, a descriptive analysis of the study variables was performed, reporting means and standard deviations for each dimension of formative assessment: content, delivery, oral expression, group organization, time of motor engagement, material selection, tasks/activities/games/exercises (TAGE), time management, classroom control, and teamwork.

Subsequently, the normality of the distributions was assessed using the Shapiro-Wilk test, as the sample size within each subgroup (according to sex and assessment modality) was fewer than 50 participants. The results indicated that most variables did not follow a normal distribution ($p < .05$); therefore, non-parametric tests were used for hypothesis testing. Differences between male and female students were examined using the Mann–Whitney U test, while differences among the three assessment modalities (self-assessment, peer assessment, and teacher assessment) were analysed using the Kruskal–Wallis test. When statistically significant differences were identified, post hoc pairwise comparisons with Bonferroni correction were applied.

In addition to statistical significance, effect sizes were calculated to provide an estimate of the magnitude and practical relevance of the observed differences, particularly given the large sample size. For

Mann–Whitney U tests, effect size was calculated using Rosenthal's r, obtained by dividing the standardized test statistic (Z) by the square root of the total number of observations. For Kruskal–Wallis tests, effect size was estimated using epsilon squared (ε^2), which provides an index of the proportion of variance explained by the grouping variable. Effect sizes were interpreted according to conventional benchmarks, with values around .10 considered small, .30 moderate, and .50 large.

Results

Differences Between Assessment Modalities

Table 1 presents the differences among assessment modalities (self-assessment, peer assessment, and teacher assessment) without considering sex. Overall, mean scores were moderately high, ranging from 3.33 to 3.69 out of 4, indicating a generally positive perception of the assessment process. The Kruskal–Wallis analysis revealed statistically significant differences for most items ($p < .001$), confirming that the type of assessment significantly influenced the ratings.

Table 1. Differences Between Assessment Modalities Regardless of Sex

	Self-assessment	Peer assessment	Teacher assessment	
Item	M ± SD	M ± SD	M ± SD	p-value
1	3.53 ± 0.62	3.54 ± 0.55	2.98 ± 0.73	<.001
2	3.69 ± 0.55	3.61 ± 0.55	3.10 ± 0.65	<.001
3	3.60 ± 0.55	3.60 ± 0.56	3.19 ± 0.65	<.001
4	3.53 ± 0.62	3.52 ± 0.62	3.01 ± 0.75	<.001
5	3.33 ± 0.69	3.41 ± 0.69	3.07 ± 0.77	<.001
6	3.43 ± 0.71	3.45 ± 0.70	2.99 ± 0.88	<.001
7	3.45 ± 0.71	3.40 ± 0.73	3.03 ± 0.91	<.001
8	3.42 ± 0.69	3.45 ± 0.73	3.05 ± 0.84	<.001
9	3.35 ± 0.74	3.37 ± 0.75	3.08 ± 0.74	<.001
10	3.33 ± 0.75	3.41 ± 0.74	3.02 ± 0.74	<.001

M ± SD = mean ± standard deviation. Significance level $p < .05$.

Across all items, the highest scores were consistently observed in self-assessment and peer assessment, whereas teacher assessment yielded the lowest values. For example, for the item "content is considered," mean scores were 3.53 for self-assessment, 3.54 for peer assessment, and 2.98 for teacher assessment, highlighting lower ratings when the evaluation was conducted by an external agent. Despite the presence of statistically significant differences, effect size estimates indicated small magnitudes, highlighting the importance of interpreting these findings cautiously in terms of practical impact.

Sex Differences

Table 2 presents sex differences without considering the type of assessment. Mann–Whitney U tests revealed that most items did not show statistically significant differences ($p > .05$), suggesting similar ratings between female and male evaluators. However, significant effects were identified for Items 1, 6, and 10 ($p < .05$). In these cases, female evaluators tended to assign lower scores in teacher assessment, whereas differences in self-assessment and peer assessment were minimal.

Table 2. Sex Differences Regardless of Assessment Modality

		Self-assessment	Peer assessment	Teacher assessment	
Item	Sex	M ± SD	M ± SD	M ± SD	p-value
1	Male	3.50 ± 0.66	3.56 ± 0.53	3.07 ± 0.66	.034
	Female	3.56 ± 0.59	3.52 ± 0.58	2.89 ± 0.79	
2	Male	3.72 ± 0.53	3.59 ± 0.56	3.12 ± 0.63	.093
	Female	3.66 ± 0.57	3.62 ± 0.55	3.08 ± 0.67	
3	Male	3.59 ± 0.55	3.60 ± 0.56	3.21 ± 0.66	.035
	Female	3.61 ± 0.56	3.61 ± 0.56	3.16 ± 0.64	
4	Male	3.53 ± 0.62	3.52 ± 0.62	3.02 ± 0.74	.136
	Female	3.52 ± 0.62	3.52 ± 0.62	3.00 ± 0.77	
5	Male	3.38 ± 0.66	3.41 ± 0.69	3.11 ± 0.78	.109
	Female	3.28 ± 0.72	3.40 ± 0.69	3.03 ± 0.76	
6	Male	3.47 ± 0.69	3.46 ± 0.70	3.07 ± 0.87	<.001
	Female	3.39 ± 0.73	3.43 ± 0.73	2.91 ± 0.88	

7	Male	3.40 ± 0.74	3.39 ± 0.74	2.93 ± 0.91	.447
	Female	3.49 ± 0.69	3.41 ± 0.73	3.34 ± 0.91	
8	Male	3.42 ± 0.68	3.46 ± 0.73	3.03 ± 0.83	.880
	Female	3.42 ± 0.69	3.45 ± 0.73	3.05 ± 0.84	
9	Male	3.38 ± 0.72	3.35 ± 0.73	3.07 ± 0.74	.068
	Female	3.32 ± 0.76	3.39 ± 0.77	3.09 ± 0.75	
10	Male	3.38 ± 0.75	3.44 ± 0.74	3.11 ± 0.74	<.001
	Female	3.29 ± 0.76	3.39 ± 0.74	2.94 ± 0.73	

M ± SD = mean ± standard deviation. Significance level p < .05.

For example, for the item "teamwork is considered" (Item 10), male evaluators assigned mean scores of 3.38 (self-assessment), 3.44 (peer assessment), and 3.11 (teacher assessment), compared with 3.29, 3.39, and 2.94, respectively, for female evaluators, reflecting a more pronounced gap in teacher assessment.

Differences Considering Sex and Assessment Modality

Table 3 examines the interactions between sex and assessment modality. The results confirm that, although sex-related differences are small in self-assessment and peer assessment, they become more pronounced in teacher assessment for certain items, showing significant main effects of sex as well as interaction effects. Notably, Items 1 (content), 6 (materials), 7 (tasks), and 10 (teamwork) exhibited significantly lower scores assigned by female evaluators in teacher assessment compared with male evaluators (p < .05).

Table 3. Differences According to Sex and Assessment Modality

		Self-assessment	Peer assessment	Teacher assessment	
Item	Sex	M ± SD	M ± SD	M ± SD	p-value
1	Male	3.50 ± 0.66	3.56 ± 0.53	3.07 ± 0.66	<.001
	Female	3.56 ± 0.59	3.52 ± 0.58	2.89 ± 0.79	
	Total	3.53 ± 0.62	3.54 ± 0.55	2.98 ± 0.73	
2	Male	3.72 ± 0.53	3.59 ± 0.56	3.12 ± 0.63	.091
	Female	3.66 ± 0.57	3.62 ± 0.55	3.08 ± 0.67	
	Total	3.69 ± 0.55	3.61 ± 0.55	3.10 ± 0.65	
3	Male	3.59 ± 0.55	3.60 ± 0.56	3.21 ± 0.66	.065
	Female	3.61 ± 0.56	3.61 ± 0.56	3.16 ± 0.64	
	Total	3.60 ± 0.55	3.60 ± 0.56	3.19 ± 0.65	
4	Male	3.53 ± 0.62	3.52 ± 0.62	3.02 ± 0.74	.444
	Female	3.52 ± 0.62	3.52 ± 0.62	3.00 ± 0.77	
	Total	3.53 ± 0.62	3.52 ± 0.62	3.01 ± 0.75	
5	Male	3.38 ± 0.66	3.41 ± 0.69	3.11 ± 0.78	.116
	Female	3.28 ± 0.72	3.40 ± 0.69	3.03 ± 0.76	
	Total	3.33 ± 0.69	3.41 ± 0.69	3.07 ± 0.77	
6	Male	3.47 ± 0.69	3.46 ± 0.70	3.07 ± 0.87	<.001
	Female	3.39 ± 0.73	3.43 ± 0.73	2.91 ± 0.88	
	Total	3.43 ± 0.71	3.45 ± 0.70	2.99 ± 0.88	
7	Male	3.40 ± 0.74	3.39 ± 0.74	2.93 ± 0.91	<.001
	Female	3.49 ± 0.69	3.41 ± 0.73	3.34 ± 0.91	
	Total	3.45 ± 0.71	3.40 ± 0.73	3.03 ± 0.91	
8	Male	3.42 ± 0.68	3.46 ± 0.73	3.03 ± 0.83	.076
	Female	3.42 ± 0.69	3.45 ± 0.73	3.05 ± 0.84	
	Total	3.42 ± 0.69	3.45 ± 0.73	3.05 ± 0.84	
9	Male	3.38 ± 0.72	3.35 ± 0.73	3.07 ± 0.74	.081
	Female	3.32 ± 0.76	3.39 ± 0.77	3.09 ± 0.75	
	Total	3.35 ± 0.74	3.37 ± 0.75	3.08 ± 0.74	
10	Male	3.38 ± 0.75	3.44 ± 0.74	3.11 ± 0.74	.003
	Female	3.29 ± 0.76	3.39 ± 0.74	2.94 ± 0.73	
	Total	3.33 ± 0.75	3.41 ± 0.74	3.02 ± 0.74	

M ± SD = mean ± standard deviation. Significance level p < .05. v

This pattern suggests that perceptions of academic performance are more strongly influenced when the evaluation is conducted by an external agent. For the remaining items (delivery, oral expression, group organization, time of motor engagement, time management, instructions, and classroom control), sex differences were minimal and did not reach statistical significance. The overall pattern remained consistent, with higher scores observed in self-assessment and peer assessment than in teacher assessment.

Discussion

The main objective of this study was to examine whether the application of triadic formative assessment (self-assessment, peer assessment, and teacher assessment) differentially influences university students' perceptions of academic performance according to sex. The primary findings confirm an overall positive evaluation of this participatory assessment model and demonstrate that the evaluator plays a decisive role, as ratings were consistently higher in peer-based modalities (self- and peer assessment) than in expert-based assessment (teacher assessment). In addition, the results suggest the possibility of an interaction effect whereby sex-related differences—minimal in self-assessment and peer assessment—may become more pronounced when the evaluation is conducted by teaching staff (teacher assessment).

Regarding differences among assessment modalities, students' overall perceptions were highly positive, as reflected by elevated mean scores across all dimensions, ranging from 3.33 to 3.69 out of a maximum of 4. These results are consistent with data reported in recent similar studies (Rodríguez-Rojas et al., 2025; Sánchez-Oliver et al., 2026). This pattern aligns with the broad academic consensus emphasizing the importance of formative assessment in promoting deep learning, competency development, and active student engagement in the learning process (Esquivel-Rivero et al., 2025; Romero-Martín et al., 2017). Previous research has similarly highlighted the relevance of these strategies, documenting their benefits for both academic performance and student motivation (Barrientos Hernán & López-Pastor, 2015). In line with these findings, Molina et al. (2023) reported that both students and teachers consider that the use of formative and shared assessment enhances academic performance and emphasize the importance of feedback within this assessment process for the development of professional competencies.

Nevertheless, the comparative analysis using the Kruskal–Wallis test revealed that the evaluator may play a relevant role in shaping ratings across almost all items ($p < .001$). The observed pattern was consistent: self-assessment and peer assessment systematically yielded the highest scores, whereas teacher assessment produced the lowest values. For instance, for the item "content is considered," mean scores were 3.53 for self-assessment, 3.54 for peer assessment, and 2.98 for teacher assessment. Similar trends have been reported in previous studies (Rodríguez-Rojas et al., 2025; Sánchez-Oliver et al., 2026). This gap between self-/peer perceptions and expert judgment should not necessarily be interpreted as a deficiency In student assessment accuracy, but rather as an essential calibration mechanism within the triadic assessment framework. While students and their peers may focus more strongly on intention and effort—characteristic of a formative mindset—teachers, through their "expert gaze" (Pérez-Pueyo & López-Pastor, 2017), apply more stringent standards of rigor and quality. This discrepancy may reflect the pedagogical value of confronting students' self-evaluations—which can be influenced by optimism or leniency biases—with standardized academic criteria, thereby fostering metacognitive reflection on expected levels of achievement within the European Higher Education Area (EHEA). In this regard, although Salom and Tena (2022) identified teacher assessment as the most highly valued modality in other contexts, in the present study—where active participation is explicitly encouraged—self- and peer assessment practices could indicate greater perceived involvement and shared responsibility, translating into more favourable ratings.

When examining sex-related differences, aggregated ratings according to the sex of the evaluator (male or female students conducting self- or peer assessment) revealed that most items did not show statistically significant differences ($p > .05$), as indicated by Mann–Whitney U tests. This finding suggests that, within the context of triadic formative assessment, average performance ratings provided by male and female evaluators are generally comparable. This result is consistent with previous studies reporting broadly equitable perceptions of formative assessment processes across sexes (Romero-Martín et al., 2017; Fernández-Garcimartín et al., 2022).

However, despite this overall convergence in mean scores, statistically significant effects were identified for specific items, namely Item 1 (Content), Item 6 (Selection and organization of materials), and Item 10 (Teamwork) ($p < .05$). In these cases, descriptive analyses suggest a tendency for male evaluators to assign slightly higher scores, particularly in teacher assessment. For example, for the item "teamwork is considered" (Item 10), mean scores assigned by male evaluators (Self: 3.38; Peer: 3.44; Teacher: 3.11)

exceeded those assigned by female evaluators (Self: 3.29; Peer: 3.39; Teacher: 2.94). Although subtle, this pattern may indicate that performance evaluation—even within self- and peer assessment contexts—could be partially conditioned by the evaluator's sex.

The analysis considering both sex and assessment modality, that is, the interaction between evaluator sex and type of assessment, constitutes the most critical contribution of this study. The results indicate that sex-related differences in ratings emerge primarily within teacher assessment. Specifically, significant interaction effects were detected for Item 1 (Content), Item 6 (Selection and organization of materials), Item 7 (Tasks/Activities/Games/Exercises), and Item 10 (Teamwork). In three of these items (Items 1, 6, and 10), male evaluators assigned higher scores than female evaluators, particularly in teacher assessment. These assessed dimensions—technical rigor, resource management, and group organization—are linked to competencies in which expectations of strictness or leniency may differ according to the evaluator's sex.

The small effect sizes observed across analyses reinforce the interpretation that the detected differences, while statistically robust, reflect subtle variations rather than large disparities in evaluative judgments. This distinction between statistical significance and practical relevance is particularly important in large-sample educational research and supports a nuanced reading of the interaction effects identified.

From a theoretical perspective, these findings may align with frameworks of implicit bias and expectancy effects, whereby evaluators unconsciously apply differentiated standards based on gender-related expectations. In educational settings, such gendered evaluation norms have been shown to influence judgment processes, particularly when assessment is conducted by an external authority figure, such as teaching staff. Consequently, the observed interaction between evaluator sex and assessment modality suggests the possibility that expert-based assessment contexts are more susceptible to subtle gender-related biases than participatory assessment modalities.

An exception to this pattern was observed for Item 7 (Tasks/Activities/Games/Exercises), where female evaluators assigned higher scores in teacher assessment compared with male evaluators. This item assesses didactic competence and pedagogical application, which may reflect different evaluative priorities or sensitivities related to instructional design and educational effectiveness.

Beyond potential evaluator bias, the observed differences should also be interpreted through a psychosocial lens. Previous research has suggested that women may experience higher levels of anxiety or emotional reactivity in evaluative contexts (Pichardo, 2000), even when general academic self-concept is comparable across sexes (Amezcua & Pichardo, 2000). Such dynamics could influence how evaluators of different sexes perceive effort, leadership, or the solidity of presented work, thereby shaping final ratings.

This assessment approach facilitates the identification of strengths and areas for improvement, making it possible to assess not only the knowledge acquired, but also students' ability to reflect on their own learning and that of their peers (Mendoza et al., 2021). With the implementation of formative and shared assessment, evaluation ceases to follow a single direction (from teacher to student) and begins to be conceived and contrasted from multiple perspectives (student–teacher, student–student, and self-assessment) (Barba-Martín & Hamodi-Galán, 2021). These strategies foster the development of critical thinking, metacognition, and self-regulated learning—core competencies within the framework of Higher Education and aligned with the need for lifelong learning promoted by UNESCO and other international organizations (López-Velarde, 2016). However, the findings corroborate what was reported by Vallés-Rapp et al. (2021), namely that for formative assessment to be effective, time and dedication are required from both teachers and students.

Limitations and Future Research Directions

Despite the robustness of the findings, several limitations should be acknowledged when interpreting the results of this study. First, the cross-sectional and non-experimental design does not allow causal inferences to be drawn regarding the relationship between assessment modality, sex, and perceived academic performance, nor does it permit conclusions about the long-term effects of triadic assessment

on learning trajectories. Nevertheless, the large overall sample size contributes to the stability and consistency of the observed patterns, providing a solid empirical basis for identifying interaction trends between evaluator sex and assessment modality.

Second, the study was conducted within a specific institutional and disciplinary context, involving students enrolled in the Primary Education degree with a specialization in Physical Education at a single university. While this contextual specificity limits the generalizability of the findings, it also represents a strength, as it allows for a detailed and ecologically valid examination of triadic assessment practices in a real educational setting. Replication of the study across different universities, degree programs, and academic disciplines is therefore warranted to determine whether the observed amplification of sex-related differences in teacher assessment reflects a context-specific phenomenon or a more generalized pattern within higher education.

Third, there was a notable imbalance in the frequency of assessment modalities, with peer assessment accounting for most records and teacher assessment representing a smaller proportion of observations. Although appropriate non-parametric statistical techniques were employed to address this imbalance, the limited number of teacher assessment records may have influenced the precision of the interaction effects detected. At the same time, this distribution reflects the authentic implementation of participatory assessment models in higher education, where peer assessment is often more frequent than expert assessment, thereby enhancing the ecological validity of the findings.

Based on these limitations, several avenues for future research can be identified. Longitudinal designs would allow for the examination of changes in students' perceptions, self-efficacy, and calibration between self-, peer-, and teacher assessment over time, offering deeper insight into the sustained impact of triadic assessment on learning outcomes and competency development. In addition, mixed-methods approaches combining quantitative analyses with qualitative data—such as interviews or focus groups with both students and teaching staff—would enable a more nuanced exploration of the psychosocial mechanisms underlying the observed differences. Such approaches could help clarify whether variations in teacher assessment are more closely related to unconscious biases in rubric application or to differences in academic self-concept, evaluative anxiety, or performance interpretation across sexes.

Finally, future research should consider the development of intervention studies grounded in action research methodologies. These interventions could incorporate targeted training in gender awareness, inclusive assessment practices, and rubric calibration for university teaching staff. Evaluating the impact of such initiatives would contribute to the design of fairer, more transparent, and more inclusive assessment systems, while also leveraging discrepancies between participatory and expert-based assessment as pedagogical opportunities for reflective and self-regulated learning. In this regard, further research should also examine how to improve students' engagement with self-assessment and peer assessment processes, since previous studies (Souto-Suárez et al., 2020) have shown that, although formative assessment enhances meaningful learning and the acquisition of professional competencies, students often display a certain reluctance towards self- and peer assessment when these practices are perceived as being closely linked to grading, which may ultimately hinder their participation in the teaching–learning process.

Conclusions

The findings indicate that assessment modality appears to be the most influential factor, with teacher assessment functioning as a calibration mechanism that tends to yield lower scores than those observed in participatory modalities. This contrast between self-perception and expert judgment may play an important role in the development of metacognitive reflection and in helping students to understand expected levels of academic achievement.

Regarding sex, the results suggest that differences in overall mean ratings between male and female evaluators are generally small, pointing to a largely comparable appraisal of the assessment process. However, interaction analyses indicate that sex may become a differentiating factor when evaluation is conducted by an external agent (teaching staff), with male evaluators tending to assign higher scores in certain dimensions, such as content, materials, and teamwork.

This amplification of differences according to the evaluator's sex suggests the possibility of implicit biases operating within teacher assessment contexts. Consequently, these findings highlight the importance of advancing toward more gender-sensitive assessment practices and of critically reviewing evaluation criteria and procedures. The effective implementation of triadic assessment should therefore not only prioritize student participation but also incorporate targeted teacher training aimed at reducing potential biases and using discrepancies between self-, peer-, and teacher assessment as an opportunity to foster calibrated self-regulation and reflective learning among students.

Acknowledgements

The authors gratefully acknowledge the collaboration of the faculty and students of the Degree in Primary Education (Physical Education specialisation) at the University of Seville, whose active involvement was instrumental in advancing communicative competence and enhancing formative assessment practices.

Funding

This study was developed within a project funded by the IV Own Teaching Plan of the University of Seville (2023 call), under the framework of faculty activities for educational innovation (L.2.2), specifically Action 221 devoted to supporting teaching innovation. Moreover, it forms part of the doctoral dissertation of Andrés Sánchez García.

References

Amezcua, J. A. Y Pichardo, M.C. (2000). Diferencias de género en autoconcepto en sujetos adolescentes. *Anales de Psicología, 16* (2), 207-214.

Barba-Martín, R. A. & Hamodi-Galán, C. (2021). La evaluación en educación superior: aportes de la red de evaluación formativa y compartida en educación. En C. Hamodi-Galán y R. Barba-Martín, *Evaluación formativa y compartida: nuevas propuestas de desarrollo en educación superior* (pp. 13-25). Dextra Editorial.

Barba-Martín, R. A., Bores-García, D., González-Calvo, G., & Hortigüela Alcalá, D. (2020). Evaluación formativa con los estudiantes en prácticas, para reducir la brecha teoría-práctica en la formación inicial del profesorado. *Educación Física y Deporte, 39*(1), 1-21. https://doi.org/10.17533/udea.efyd.v39n1a02

Barrientos Hernán, E., López Pastor, V. M., & Pérez-Brunicardi, D. (2019). ¿Por qué hago evaluación formativa y compartida y/o evaluación para el aprendizaje en EF? La influencia de la formación inicial y permanente del profesorado. *Retos, 36*, 37-43. https://doi.org/10.47197/retos.v36i36.66478

Bustamante Castaño, S. A., González Palacio, E. V., Chaverra Fernández, B. E., & López-Pastor, V. M. (2025). Representaciones sociales sobreevaluación en formación inicial del profesorado en Educación Física. *Retos, 65*, 673–685. https://doi.org/10.47197/retos.v65.110341

Esquivel-Rivero, Y., Rivero Rodríguez, E. M., & Sánchez Armijos, T. M. (2025). Evaluación Formativa en la Educación Superior: Prácticas Innovadoras para la Formación Docente. *Revista Científica de Salud y Desarrollo Humano, 6*(1), 648–662. https://doi.org/10.61368/r.s.d.h.v6i1.501

Estaji, M. (2024). Perceived need for a teacher education course on assessment literacy development: insights from EAP instructors. *Asian-Pacific Journal of Second and Foreign Language Education, 9*(1), 50. https://doi.org/10.1186/s40862-024-00272-2

Fernández-Garcimartín, C., Fuentes Nieto, T., Molina Soria, M., & López-Pastor, V. M. (2022). La Participación del Alumnado en la Evaluación Formativa en Formación del Profesorado. *Revista Iberoamericana De Evaluación Educativa*, *15*(1), 61-80. https://doi.org/10.15366/riee2022.15.1.004

Fraile, J., Ruiz-Bravo, P., Zamorano-Sande, D., & Orgaz-Rincón, D. (2021). Evaluación formativa, autorregulación, feedback y herramientas digitales: uso de Socrative en educación superior. *Retos, 42*, 724–734. https://doi.org/10.47197/retos.v42i0.87067

Gallardo-Fuentes, F., Carter Thuillier, B., Martínez-Angulo, C., Gallardo-Fuentes, J., & Peña Troncoso, S. (2025). Percepciones del alumnado en dos cohortes de la formación inicial en Educación Física: consistencia, compromiso y retos en la evaluación. *Retos, 70*, 39–50. https://doi.org/10.47197/retos.v70.113463

Gallardo-Fuentes, F., Carter-Thuillier, B., López-Pastor, V., Ojeda-Nahuelcura, R., & Fuentes-Nieto, T. (2022). Assessment systems in Physical Education teacher training: A case study in Chilean context. *Retos, 43*, 117-126. https://doi.org/10.47197/retos.v43i0.88570

García-Jiménez, M., & Trigo, M. E. (2025). Sesgos de género: un análisis de los factores que contribuyen a la desigualdad en la investigación psicológica. *Apuntes de Psicología*, *43*(1), 37–45. https://doi.org/10.70478/apuntes.psi.2025.43.04

González-Gutiérrez, I., López-García, S., Barcala-Furelos, M., Mecías-Calvo, M., & Navarro-Patón, R. (2024). ¿Existen diferencias en la percepción del alumnado sobre la evaluación recibida en las clases de educación física? Un estudio en función del género y etapa educativa. *Retos: Nuevas Tendencias en Educación Física, Deporte y Recreación, 59*, 632–641. https://doi.org/10.47197/retos.v59.108260

Martín, R. B., & Alcalá, D. H. (2022). Si la Evaluación es Aprendizaje, he de Formar parte de la misma. Razones que Justifican la Implicación del Alumnado. *Revista Iberoamericana de Evaluación Educativa*, *15*(1), 9-22. https://doi.org/10.15366/riee2022.15.1.001

Membrilla, J. A. A., & Martínez, M. C. P. (2000). Diferencias de género en autoconcepto en sujetos adolescentes. *Anales de Psicología, 16*(2), 272-284.

Mendoza, S. T. B., Cedeño, J. A. M., Espinales, A. N. V., & Gámez, M. R. (2021). Autoevaluación, Coevaluación y Heteroevaluación como enfoque innovador en la práctica pedagógica y su efecto en el proceso de enseñanza-aprendizaje. *Polo del Conocimiento: Revista científico-profesional*, *6*(3),828-845. https://doi.org/10.23857/pc.v6i3.2408

Molina, M., López-Pastor, V. M., Hortigüela-Alcalá, D., Pascual-Arias, C., & Fernández-Garcimartín, C. (2023). Formative and shared assessment and feedback: an example of good practice in Physical Education in Pre-service Teacher Education. *Cultura, Ciencia y Deporte, 18*(55), 157–169. https://doi.org/10.12800/ccd.v18i55.1986

Olague, M. Z., Cañadas, L., & Manso, J. (2025). Percepción del profesorado de educación básica sobre la participación de los diferentes agentes (profesorado y alumnado) en los procesos de evaluación. *Revista Fuentes*, *27*(3), 289-300. https://doi.org/10.12795/revistafuentes.2025.27856

Pérez-Pueyo, Á., & López-Pastor, V. M. (Coords.). (2017). *Evaluación formativa y compartida en educación: experiencias de éxito en todas las etapas educativas*. León, España: Universidad de León.

Pichardo Martínez, M. C. (2000). Influencia de los estilos educativos de los padres y del clima social familiar en la adolescencia temprana y media (Tesis doctoral, Universidad de Granada). http://hdl.handle.net/10481/14578

Rodríguez-Rojas, F. J., Grimaldi-Puyana, M., Bianchi, P., & Sánchez-Oliver, A. J. (2025). Autoevaluación, coevaluación y heteroevaluación como métodos de evaluación inclusivos y equitativos durante la práctica física en el grado de Educación Primaria. En M. Puig Gutiérrez, A. J. García González, A. M. García Parejo, & M. Carrillo Cabeza (Eds.), *Innovación educativa. Investigación, cambio y acción* (pp. 436–447). Dykinson.

Rojas, M. R. V., Navarrete, C. F. G., Roco, C. B. O., Villouta, C. M. Y., & Bracho, D. D. V. M. (2025). Sesgo de Género en la Educación Superior: Una Deuda Persistente y Desafíos para la Equidad. *Ciencia Latina Revista Científica Multidisciplinar, 9(3),* 6009-6033. https://doi.org/10.37811/cl_rcm.v9i3.18238

Romero-Martín, R., Castejón-Oliva, F., López-Pastor, V., & Fraile-Aranda, A. (2017). Evaluación formativa, competencias comunicativas y TIC en la formación del profesorado. *Comunicar, 52*, 73-82. https://doi.org/10.3916/C52-2017-07

Salom, M. A. C., & Tena, B. A. (2022). Complicidad entre Autoevaluación y Aprendizaje. Matices para su Implantación en la Universidad. *Revista Iberoamericana de Evaluación Educativa*, *15*(1), 23-42. https://doi.org/10.15366/riee2022.15.1.002

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.

Souto-Suárez, R., Jiménez-Jiménez, F., & Navarro-Adelantado, V. (2020). La percepción de los estudiantes sobre los sistemas de evaluación formativa aplicados en la educación superior. *Revista Iberoamericana de Evaluación Educativa, 13*(1), 11–39. https://doi.org/https://doi.org/10.15366/riee2020.13.1.001

Vallés-Rapp, C., Ureña Ortín, N., & Ruiz Lara, E. (2011). La Evaluación Formativa en Docencia Universitaria. Resultados globales de 41 estudios de caso. *REDU. Revista De Docencia Universitaria, 9*(1), 135. https://doi.org/10.4995/redu.2011.6184

Zubillaga-Olagüe, M., & Cañadas, L. (2022). Agentes participantes en los procesos de evaluación y calificación en Educación Física. *Contextos Educativos, 30*, 83–98. https://doi.org/10.18172/con.5371

Authors' and translators' details:

Andrés Sánchez-García	andressanchezgarcia9696@gmail.com	Autor
Antonio Jesús Sánchez-Oliver	sanchezoliver@us.es	Autor
Moisés Grimaldi-Puyana	mgrimaldi@us.es	Autor