Studies in Educational Evaluation 91 (2026) 101699

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Summative self-assessment accuracy in higher education: An exploratory study of a grade incentive, individual differences, and perceived versus actual accuracy

Ernesto Panadero [a,b,c], Juan Fraile [d,*], Anastasios Karakolidis [e]

[a] Centre for Assessment Research Policy and Practice in Education (CARPE), School of Policy and Practice, Institute of Education, St. Patrick's Campus, Dublin City University, Ireland
[b] Education, Regulated Learning and Assessment (ERLA Group), Facultad de Educación y Deportes, Universidad de Deusto, Bilbao, Spain
[c] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[d] Universidad Francisco de Vitoria, Madrid, Spain
[e] Educational Research Centre, Ireland

ARTICLE INFO

Keywords:
Self-assessment
Summative self-assessment
Perceived accuracy
Self-assessment accuracy
Metacognition
Calibration
Social prediction

ABSTRACT

This exploratory quasi-experimental study examined whether an accuracy-based incentive for summative self-assessment improved alignment with instructor-assigned grades in Spanish higher education, which student characteristics were associated with accuracy, and whether perceived accuracy corresponded to actual accuracy. A total of 142 second-year undergraduates from seven classroom groups were assigned to an incentive or control condition, completing rubric-based self-assessments and questionnaires on self-regulation, self-efficacy, and prior academic achievement. No clear evidence of improved accuracy was found; Bayesian and equivalence analyses indicated the data were inconclusive rather than a null effect. Prior achievement predicted accuracy. Self-regulation interacted with condition, but the same interaction appeared on instructor-assigned grades, indicating pre-existing group differences. Perceived accuracy was unrelated to actual accuracy ($R^2 = .001$), so students' confidence cannot serve as a proxy for precision. Self-assessment accuracy was associated with final exam performance, although causal conclusions cannot be drawn. Grading conditions shaping self-assessment accuracy warrant systematic study.

Self-assessment has gained increasing recognition as a critical component of higher education because it engages students in reflective judgment and enhances their metacognitive skills (Panadero et al., 2016). When students evaluate their own work against established criteria, they become more aware of their strengths and weaknesses, which can lead to improved academic performance (Dochy et al., 1999; Yan et al., 2022). The educational value of self-assessment depends on the judgment approaching the actual characteristics of the work, whatever the purpose for which it is used (Panadero et al., 2016). The process of self-assessing can be valuable in itself (Andrade, 2019), but its benefits are not a substitute for accuracy: a self-assessment that departs markedly from the quality of the work feeds misguided decisions about what to do next, and discovering the discrepancy later carries emotional and motivational costs, particularly for students who overestimated (Pinedo et al., 2026). Large deviations are therefore a problem to be addressed through instruction, not a tolerable by-product of the process. When the self-assigned score also contributes to the grade, as in this study, inaccuracy additionally compromises the validity of that grade (Brown & Harris, 2013).

Research has shown that while self-assessment can be beneficial for learning and performance, many students struggle to produce accurate summative self-assessments (i.e., self-scores) (León et al., 2023; León, Vallejo, et al., 2021). This distinction is often conceptualized as the difference between actual accuracy -i.e., meaning the degree of alignment between students' scores and expert ratings- and perceived accuracy, or students' subjective beliefs about how precise their self-assessments are (León et al., 2023; Panadero et al., 2016; Yan & Brown, 2017). This lack of accuracy undermines one of the core objectives of self-assessment: fostering students' capacity to critically and realistically appraise their own work. Thus, while self-assessment can improve learning, its effectiveness is compromised when students misjudge their own performance. Furthermore, recent studies have

* Correspondence to: Universidad Francisco de Vitoria. Carretera Pozuelo-Majadahonda km. 1.8 – 28223 Pozuelo de Alarcón, Madrid, Spain.
E-mail address: juan.fraile@ufv.es (J. Fraile).

https://doi.org/10.1016/j.stueduc.2026.101699
Received 27 June 2026; Received in revised form 17 October 2026; Accepted 18 October 2026
Available online 26 October 2026
0191-491X/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

emphasized that motivational and cognitive biases, including calibration bias, can systematically influence the accuracy of self-assessments (León, Vallejo, et al., 2021; Sitzmann et al., 2010). Understanding how to enhance self-assessment accuracy is a crucial research priority, particularly in summative contexts where students assign themselves grades that may not align with expert evaluations.

In response to this challenge, identifying strategies that improve self-assessment accuracy should be a focus for educational researchers (Panadero et al., 2016). One potentially promising approach is to provide students with a direct incentive to align their summative self-assessment or self-assessed grades with those assigned by their instructor. In other words, rather than just an evaluation of the quality of the work alone and its corresponding teacher's score, students could receive a score-related incentive specifically tied to the accuracy of their self-assessment, regardless of the overall quality of their performance on the task. This approach aims to reduce the impact of overestimation biases and encourage more calibrated judgments (Kruger & Dunning, 1999; León, Vallejo et al., 2021). Such an incentive may encourage students to pay closer attention to the assessment criteria and make more careful judgments, ultimately improving their self-assessment accuracy. The present study aims to examine whether introducing such an incentive can lead to more accurate summative self-assessments in a higher education context.

1. Self-assessment: definition, types and summative self-assessment

Self-assessment "involves a wide variety of mechanisms and techniques through which students describe (i.e., assess) and possibly assign merit or worth to (i.e., evaluate) the qualities of their own learning processes and products" (Panadero et al., 2016, p. 804). From an instructional point of view, the idea behind self-assessment is that rather than relying solely on external judgments, learners engage metacognitively to monitor and regulate their understanding and skills. Its use has spread across educational levels and countries, and it has been found to enhance students' academic performance (Yan et al., 2022), self-regulated learning and self-efficacy (Panadero et al., 2017).

Crucially, there is considerable diversity in how self-assessment can be implemented, ranging from open-ended, qualitative reflections on the learning process to more structured, criterion-referenced evaluations that require students to systematically compare their work against defined standards (Panadero et al., 2016). While the former approach often emphasizes metacognition, self-regulation, and personal goal-setting, the latter typically provides a clearer, more objective framework for judgment, guiding learners to understand performance expectations in a more concrete manner (Ross, 2006). However, even in structured formats, research has documented substantial variability in students' ability to produce accurate self-assessments aligned with expert ratings (Brown & Harris, 2013; León et al., 2023; León, Vallejo, et al., 2021).

One particularly notable form is summative self-assessment, where learners assign themselves a grade or score based on established criteria. This is also known as self-scoring or self-grading, but here we will be using the term summative self-assessment. This practice continues to receive considerable attention in the literature as shown by Andrade (2018) who identified summative self-assessment as still the most prevalent area of research. Some scholars, however, have taken a more critical stance, questioning whether summative self-assessment truly promotes deeper engagement and mastery or merely shifts the act of grading onto students without providing formative feedback loops (Andrade, 2019). Others recognize its potential for improving students' evaluative judgment and academic confidence, even as they acknowledge that formative self-assessment may offer more powerful learning gains (Panadero et al., 2016; Sanchez et al., 2017). In essence, while summative self-assessment may enhance learners' calibration and realism about their abilities, contributing to a stronger sense of responsibility for their own academic outcomes, it may lack some of the metacognitive richness inherent in formative approaches. Nevertheless, it is not necessary to choose between formative or summative self-assessment, as both can be implemented within an instructional process, each offering distinct benefits that complement one another. Summative self-assessment can also serve as a reflective opportunity for students, encouraging them to critically evaluate their efforts after having theoretically submitted their best possible work. As recent studies have noted, its validity and precision remain partly contingent on prior training, clear criteria, and students' metacognitive skills (León et al., 2023; León, Vallejo, et al., 2021). As a result, researchers continue to examine summative self-assessment's role under the belief that it can foster realism and honesty in students' judgments, encouraging them to internalize performance criteria and become more accountable for their learning outcomes (Falchikov & Boud, 1989; Panadero et al., 2016).

2. Accuracy: the importance of precision in self-assessment

Accuracy in educational assessment broadly refers to the extent to which a given measure or judgment faithfully represents a learner's actual level of achievement or competence (Panadero et al., 2016). In the context of self-assessment, accuracy is particularly critical because it reflects how closely students' evaluations of their own work align with an established standard or expert judgment (Brown & Harris, 2013). Conceptualized at a global level, accuracy encompasses both the reliability and validity of students' evaluative judgments across various domains and tasks. It thus serves as a cornerstone for understanding the credibility and utility of self-assessment practices, especially when these practices inform instruction, feedback, and learning strategies (Brown & Harris, 2013; Panadero et al., 2016). It is also important to distinguish between actual self-assessment accuracy (i.e., objective alignment with expert ratings) and perceived accuracy, which refers to students' subjective confidence in their evaluations (Panadero et al., 2016; Yan & Brown, 2017).

Within this broad conceptual territory, "self-assessment accuracy" emerges as a more specific operationalization of accuracy, focusing directly on how well learners' self-assigned grades or quality judgments match external criteria or expert ratings (León et al., 2023; León, Vallejo, et al., 2021). By examining discrepancies between teacher and student evaluations, researchers can identify factors that help learners calibrate their judgments, thereby improving the meaningfulness of self-assessment. Calibration bias (i.e., systematic tendencies to overestimate or underestimate performance) plays a key role in these discrepancies and may be shaped by both cognitive and motivational influences (Pinedo et al., 2025; Sitzmann et al., 2010). This level of precision is especially relevant as educational systems consider shifting a portion of evaluative responsibility onto the learner, making it necessary to ensure that these self-generated judgments maintain their interpretive value and do not distort learning trajectories (Panadero et al., 2016).

Two clarifications about accuracy are needed for the present study. First, we use accuracy here in a specific sense: the degree of alignment between a student's self-assigned score and the instructor's score, both produced with the same analytic rubric, that is, scoring accuracy (León et al., 2023). This is one type of accuracy among several. Scoring accuracy says little about content accuracy, the extent to which students correctly identify the strengths and weaknesses of their work and what to do about them, which is arguably more important for learning (Panadero et al., 2016; see also Andrade, 2019); nor does it guarantee that the instructor's score is itself free of error. We return to these points in the Limitations. Second, a reported self-score is not a pure expression of what students believe about their work. When the score has consequences, because it contributes to a grade or is seen by the instructor, students may report a score that differs from the one they privately consider deserved, for instance to protect a grade, to avoid disappointing the instructor, or to make a good impression (Panadero et al., 2016;

Pinedo et al., 2025). The alignment we measure therefore reflects a mix of metacognitive judgment and social and strategic considerations, a point developed below.

However, the body of literature examining self-assessment accuracy faces notable limitations for two main reasons. First, many studies show repetitive simplistic designs, insufficient rigor in reporting, and a lack of methodological detail, all of which restrict the inferences that can be drawn about the stability and generality of findings (Panadero et al., 2017). Recent contributions (León, Vallejo, et al., 2021) have emphasized the importance of controlling key methodological factors such as consistent criteria, shared rubrics, and prior training, to obtain more valid and reliable assessments. Such methodological flaws not only constrain our understanding of the conditions under which self-assessment accuracy flourishes but also raise questions about the validity and relevance of conclusions that have become commonplace in the literature. And second, there is a tendency to focus on the effects of certain practices (e.g., rubrics use) without adequately exploring the nuanced influences of variables such as incentive structures, student motivation, and differential achievement levels. This omission is striking, given that motivational variables are known to influence not only learning behaviors but also the accuracy and calibration of self-evaluations (Sitzmann et al., 2010). This narrow scope hinders our understanding. Recent work has begun to move beyond documenting the accuracy gap, examining the reasons students give for their (in)accuracy, the individual differences involved, and the emotional consequences of discovering a discrepancy with the instructor's score (Pinedo et al., 2025, 2026).

A particularly understudied area lies in the role of external incentives designed to encourage students to be more accurate in their self-assessments. While meta-analyses and reviews (e.g., León et al., 2023; Panadero et al., 2016) highlight the complexity of improving self-assessment accuracy, none have documented attempts to manipulate incentives as a direct means of enhancing the alignment between self- and teacher-assigned grades. The consequences attached to self-assessment plausibly matter as well: when being unrealistic carries no consequences, students may have little reason to invest in a realistic, verifiable self-assessment (Panadero et al., 2016), which underscores the need for more systematic research on how incentive structures shape evaluative behavior. By providing students with tangible motivators -such as a small incentive for more accurate judgments- researchers may uncover previously untested avenues for improving the calibration of learners' self-appraisals. This gap in the literature is particularly striking, given that incentivizing careful, criterion-referenced judgment could theoretically lead to greater student engagement, deeper processing of evaluation criteria, and ultimately, stronger correspondences between self- and teacher-assigned grades. To date, no study has directly tested whether providing an accuracy-based incentive improves summative self-assessment precision, making this a genuinely untested question in the self-assessment literature.

The rationale for expecting an incentive to improve accuracy needs to be stated with some precision, because an incentive cannot act on every source of inaccuracy. Deviations between self-assigned and instructor scores have at least two sources. The first is evaluative competence: students may lack the domain knowledge, the understanding of the criteria, or the monitoring skill needed to judge their work, a limitation that is well documented (Dunning et al., 2004; León et al., 2023) and that no incentive can remove at the moment of self-assessment. The second is the reporting decision: a student who holds a reasonably accurate private judgment may still report a different score, for instance to protect a grade or to make a good impression (Panadero et al., 2016; Pinedo et al., 2025). An accuracy incentive acts on this second source. It changes the consequences of the reported score, so that students have a reason to report the score they consider most defensible rather than the most favorable one, and it may lead them to consult the criteria more carefully before deciding. In the specific sense of accuracy used in this study, alignment between the reported score and the instructor's score, a change in reporting is an improvement in accuracy, even though it is not an improvement in evaluative competence. The expected effect is therefore bounded: an incentive should reduce the part of the deviation that stems from strategic or careless reporting and leave the part that stems from limited evaluative competence largely intact. It follows that an incentive should be most useful for students who are capable of judging their work accurately but have no reason to report that judgment, and least useful where the judgment itself is poor. It also follows that an improvement in alignment would not, by itself, demonstrate improved metacognitive monitoring, a point developed in the next section.

3. Self-assessment accuracy: metacognitive monitoring or social prediction?

A critical conceptual issue that requires explicit attention is whether self-assessment accuracy, as operationalized through alignment with instructor scores, reflects metacognitive self-knowledge or a form of social prediction. When students are asked to self-assess using the same rubric that the instructor will use, they face a dual task: they must evaluate the quality of their own work (a metacognitive process) but they must also anticipate how the instructor will apply the criteria (a social-predictive process). Incentivizing alignment with the instructor's score may primarily target the latter rather than the former.

This distinction has important implications. As Rouy et al. (2022) demonstrated in the context of metacognitive training, improvements in reported metacognitive accuracy may reflect changes in reporting behavior rather than genuine gains in self-knowledge. In the self-assessment context, a student who becomes 'more accurate' might simply be better at predicting instructor preferences without developing deeper insight into their own competence. This concern is particularly relevant when accuracy is incentivized, because the incentive explicitly rewards alignment with an external judgment rather than the quality of internal self-monitoring.

However, the distinction between metacognitive monitoring and social prediction may be less absolute than it first appears, particularly in contexts where assessment criteria are explicit, shared, and extensively scaffolded. In rubric-based assessment, the instructor's score is not an opaque personal judgment but rather an operationalization of publicly available quality criteria. A student who accurately predicts this score must engage with those criteria and evaluate their work against them, a process that involves metacognitive engagement even if it also involves prediction of how the criteria will be applied. The degree to which self-assessment accuracy reflects 'pure' metacognition versus social prediction is therefore likely to vary across assessment contexts, and our study cannot disentangle these mechanisms. We return to this issue in the Discussion.

4. The present study: aim, research questions and hypotheses

Against this backdrop, this study explores whether providing students with a direct incentive for accuracy at the moment of self-assessment leads to more precise self-scoring. We frame the investigation as exploratory, given the quasi-experimental design and the novelty of the manipulation. The term incentive refers here to a grade-based mechanism intended to promote accuracy in students' summative self-assessment (see Procedure). Beyond the effect of the incentive, we examine which student characteristics are associated with higher accuracy, whether students' perceived accuracy corresponds to their actual accuracy, and whether accuracy is related to subsequent academic performance. We explored four research questions (RQ).

RQ1. Does providing an accuracy-based incentive at the point of self-assessment enhance the accuracy of students' summative self-assessments?

H1: Students in the incentive condition will demonstrate higher self-assessment accuracy than those in the control condition (Panadero et al.,

2016; León, Vallejo, et al., 2021). However, given that the incentive was introduced only at the moment of self-assessment rather than throughout the course, we note that this hypothesis tests a narrow form of incentive effect, specifically whether awareness of an accuracy incentive at the point of self-assessment shifts scoring behavior.

RQ2. Do individual characteristics -prior academic achievement, self-regulatory skills, self-efficacy, and gender- influence the impact of accuracy-based incentives on students' summative self-assessment accuracy?

H2: We hypothesized that individual characteristics would moderate the effect of the incentive on self-assessment accuracy. Specifically, based on prior research: (a) prior academic achievement would predict higher baseline accuracy (León et al., 2023; Pinedo et al., 2025; Yan & Brown, 2017); (b) students with stronger self-regulatory skills would show greater benefit from the incentive, given that self-regulation facilitates criterion engagement and metacognitive monitoring (Panadero & Jonsson, 2013) and that students with stronger study strategies and engagement tend to self-assess more accurately (León, Augusto-Landa, et al., 2021); (c) self-efficacy was not expected to moderate the incentive effect, as the relationship between self-efficacy and self-assessment accuracy is inconsistent in the literature (Brown & Harris, 2013); and (d) gender effects were expected to be non-significant based on prior findings (León et al., 2023).

RQ3. Does students' perceived accuracy (i.e., their self-reported ability to self-assess accurately and their perceived alignment with the instructor's evaluation) predict their actual summative self-assessment accuracy, and is this relationship moderated by the incentive condition?

H3: Students with higher perceived accuracy will exhibit higher actual summative self-assessment accuracy. However, this relationship will not be significantly moderated by the incentive condition (Sitzmann et al., 2010; Yan & Brown, 2017).

RQ4. Does the accuracy of students' self-assessments predict their subsequent academic performance on the final examination, and is this relationship moderated by prior academic achievement and experimental condition (incentive vs. control)?

H4: Higher self-assessment accuracy will be associated with higher final exam performance. However, this relationship will not be significantly moderated by prior academic achievement or the incentive condition (Brown & Harris, 2013; Panadero et al., 2016).

5. Method

5.1. Design

A quasi-experimental design with intact classroom groups was employed with two conditions. Experimental condition: Participants received a grade incentive (see Procedure for details) if the accuracy of their summative self-assessment matched closely to instructor's grade. And a control condition: Participants self-assessed their work but were not told about any incentives for accurate summative self-assessment. Of the seven intact classroom groups, three (n = 65) were assigned to the control condition and four (n = 77) to the incentive condition.

Classroom groups were allocated to the experimental and control conditions according to their scheduled lecture days, across the two academic years. This allocation was chosen for practical reasons, to prevent information transfer between conditions within the same week. For the same reason, the incentive was communicated verbally and in writing only during the final class session, immediately before students completed the self-assessment exercise. The implications of this non-random allocation for interpretation are addressed in the Limitations section.

The incentive was introduced at the point of self-assessment rather than at an earlier stage of the course. This choice was made to minimize the risk of students perceiving discrepancies between conditions through comparison with peers, as some might otherwise have noticed that classmates were operating under different arrangements. As explained later, the conditions were subsequently equalized with respect to the incentive so that the control group was not disadvantaged in any way.

No reminders were provided between the initial announcement and the submission of the self-assessment. Although no formal verification was conducted to confirm whether the control group became aware of the incentive, no students raised questions about it. In any case, any such awareness would likely have occurred only after students in the experimental group had already completed their self-assessment. For ethical reasons, the incentive was applied to the final grades of all students after the end of the semester, regardless of condition. However, the instructor- and self-assigned scores used in this study were collected before this adjustment and were not affected by the incentive application in the control group.

5.2. Participants

The final sample size was determined by the total enrollment across the academic years rather than an a priori power calculation (for post hoc sensitivity analysis see Results section). The potential participants were 171 undergraduate students enrolled in the second year of a Bachelor's Degree in Physical Activity and Sport Sciences at a Spanish university. These students were drawn from two academic years. Each participant took part during their respective academic year and all were enrolled in the same course taught by the same instructor during these two years. The course was organized into classroom groups, each meeting once per week (Monday, Wednesday, Thursday, or Friday) for a four-hour session: four groups in the first academic year and three in the second. Therefore, seven intact classroom groups of 15–26 students participated. 26 students were removed from the initial database: five for detection of plagiarism in the elaboration of their assignment and 21 for not having submitted the assignment in the ordinary call for the subject. An additional three students were removed because they did not participate in the self-assessment exercise. Thus, our final sample size is 142 participants.

The classroom groups included both students who had previously completed vocational training (and were generally older) and students who had followed the traditional high school and university entrance examination pathway. Both entry pathways were represented in both conditions (incentive condition: 41 vocational-entry and 36 university-entrance students; control condition: 41 and 24, respectively).

In the final sample, the control group consisted of 65 students (23.1% female), while the intervention group of 77 students (20.8% female). The overall mean age of participants was 22.05 years ($SD = 2.69$). The proportion of females and males did not differ significantly between the control and intervention groups, χ^2 (1, $N = 142$) = 0.109, $p = .741$. Among the 65 students in the control group, 40.0% were enrolled during the 2023–24 academic year, compared to 54.5% in the intervention group. This distribution of academic years also did not differ significantly between the groups, χ^2 (1, $N = 142$) = 2.988, $p = .084$.

To assess baseline equivalence between conditions, we compared the two groups on all key pre-intervention variables. No significant differences were found for prior academic achievement (Control $M = 7.47$, $SD = 0.61$; Intervention $M = 7.40$, $SD = 0.75$; $t(139) = 0.53$, $p = .594$), pre-intervention self-efficacy ($t(125) = -1.40$, $p = .164$), learning self-regulation ($t(125) = -0.67$, $p = .506$), avoidance self-regulation ($t(125) = 1.00$, $p = .321$), or deep learning strategies ($t(125) = -0.14$, $p = .891$). A significant difference was observed for age ($p = .007$), though this variable is not theoretically linked to self-assessment accuracy and was not included in the analytical models. These comparisons suggest that, despite the non-random assignment, the two conditions were largely equivalent on the variables most relevant to self-assessment accuracy. Importantly, instructor-assigned task grades were virtually identical across conditions (Control $M = 7.63$; Intervention $M = 7.60$; $t(140) = 0.10$, $p = .920$), indicating that the groups did not differ in task performance prior to the self-assessment exercise.

The research was approved by the Ethics Committee of the first author's university (Code ETK-22/20–21). The study ensured voluntary participation while seamlessly integrating self-assessment into the course design. Since self-assessment is acknowledged as an effective educational tool, it was employed as a standard practice for all students. Participation in the research component, however, was optional and required students to complete questionnaires and consent to the use of their data. Before collecting any data, students were thoroughly informed about the study through an explanatory session, an information sheet, and an informed consent form. All participants gave their consent willingly.

5.3. Materials and instruments

Several measures were administered before and after the intervention to evaluate prior achievement, self-regulation, self-efficacy, and final performance:

5.3.1. Summative self-assessment accuracy

At the end of the course (in the final session), once the assignment had been submitted, students reviewed their work using a personal device (e.g., laptop, tablet) and the printed rubric. They then self-assessed the assignment using the rubric, producing a summative self-assessment (i.e., self-score) for each assessment criterion included in the rubric.

Consequently two indices were constructed to evaluate students' self-assessment accuracy, taking into account the weighting of each of the 17 criteria. First, the Signed Deviation Index was calculated by subtracting each student's self-assigned grade from their instructor's assigned grade. In this index, students could take scores ranging from −10 to + 10; positive values in this index indicate that students underestimated their performance, while negative values indicate they overestimated it. Second, the Absolute Deviation Index was calculated by summing the absolute difference between the student's and the instructor's score on each of the 17 criteria, weighted by criterion, and normalizing the total to the same 0–10 scale as the other performance indices. Because deviations in opposite directions do not cancel out across criteria, this index is larger than the absolute value of the difference between total scores ($M = 1.02$). Higher values indicate greater deviations and thus lower accuracy. We use the term deviation for both indices to make explicit that they quantify the distance between self-assigned and instructor-assigned scores; accuracy is inferred from smaller deviations.

The Signed Deviation Index provides insight into students' self-assessment tendencies, indicating how often they overestimate or underestimate their performance. In contrast, the Absolute Deviation Index reflects the overall magnitude of deviation regardless of direction. This measure was selected as the primary index in this study because it is less affected by the potential cancellation of positive and negative deviations when averaged across criteria or students.

This summative self-assessment accounted for 10% of the grade for this task, while the instructor's assessment accounted for 90%. The grade for this task represented 25% of the final course grade. For the purposes of the analyses in this study, only the instructor's grade was used as the task grade. The 10% contribution of the self-assessment score to the task grade applied equally to both conditions. Because this contribution was identical in both groups, it constitutes a design constant rather than a differential confound: the only systematic difference between conditions was the additional accuracy-based incentive provided to the experimental group.

5.3.2. Course assignment

One individual and long assignment was used as focal task in the intervention: the Hiking Design Assignment, which accounted for 25% of the final grade. This assignment required students to individually design a complete hiking route for a specific target group, integrating core competencies from the course. Students had to plan the route (including distance, elevation, and terrain analysis), develop safety protocols, address environmental considerations, design group management strategies appropriate for the target population, and handle logistical planning (equipment, nutrition, timing, and emergency procedures). The task was integrative, requiring the application of knowledge from multiple course modules into a single practical deliverable. This task was graded using a rubric.

5.3.3. Rubric

The assignment was graded using an analytic rubric specifically developed by a group of instructors responsible for this course. The rubric contains 17 criteria, each weighted according to its relative importance in assessing the assignment (see Appendix A). The criteria and their weights were established collaboratively by faculty members with experience in both instructional design and assessment, and the rubric has been used successfully in the course over several academic years. This iterative use has contributed to its content validity and alignment with the learning objectives of the subject.

Each criterion included four performance levels, corresponding to the following grading intervals: 0–4, 4–6, 6–8, and 8–10. This design avoids forcing the selection of a single discrete performance level and allows for more fine-grained scoring. When grading, instructors assigned a specific numeric score from 0 to 10 for each criterion. The final score was computed by applying the criterion weights to each score and then summing the weighted scores. To facilitate comparability, the resulting total was normalized to fit within a 0–10 scale.

Inter-rater reliability was examined to assess the consistency of the grading process. The two instructors responsible for evaluating the assignments were the instructor of the course from which the data were collected and a second instructor who teaches the same course but whose students were not included in the sample, in order to preserve consistency and control over the learning process. Both independently graded the assignments of 30 participants (21.1% of the final sample, balanced across conditions) using the same 17-criterion rubric. Agreement on the weighted total scores was estimated with an intraclass correlation coefficient, using a two-way random-effects model with absolute agreement: ICC(2,1) = .952, 95% CI [.799,.983], $F(29, 29) = 63.81$, $p < .001$.

5.3.4. Prior academic achievement

The students' first-year university academic record was obtained at the start of the academic year which was the second year for the participants. This information was extracted directly from the institutional academic management system to ensure data accuracy. This variable could take values between 0 and 10. In the present sample, the observed scores ranged from 5.65 to 8.84.

5.3.5. Emotion and motivation self-regulation questionnaire (EMSR-Q)

The EMSR-Q (Alonso-Tapia et al., 2014) comprises 20 items grouped theoretically into five first-order factors (Process Oriented SR, Positive Motivation SR, Performance Oriented SR, Avoidance Oriented SR, and Negative Stress SR), each with four items. Responses are given on a 5-point Likert scale ranging from 1 (Never) to 5 (Always). An example item is: "I try to control my emotions when studying."

An exploratory factor analysis (Maximum Likelihood, Promax rotation) on our sample confirmed the expected five-factor structure, with satisfactory fit indices (KMO = .828, $\chi^2 = 119.538$, $p = .089$) and explained variance of 65.88%. To explore the higher-order structure described in the instrument manual, an EFA forcing two factors was also conducted. Although this solution accounted for a smaller proportion of variance (47.4%) and presented some cross-loadings, it broadly reflected the grouping of first-order factors into two second-order dimensions (Learning Self-Regulation and Avoidance Self-Regulation), consistent with prior validation studies (see Panadero & Alonso-Tapia, 2025). However, we note that confirming this higher-order model would require confirmatory factor analysis.

For subsequent analyses, composite scores were calculated for these two higher-order dimensions, as described in the EMSR-Q manual. Reliability in this sample was appropriate: for Learning Self-Regulation, Cronbach's alpha was .797 and McDonald's omega was .798; for Avoidance Self-Regulation, Cronbach's alpha was .874 and McDonald's omega was .874.

5.3.6. *Deep learning self-regulation strategies questionnaire (DLS-Q)*

The DLS-Q (Panadero et al., 2021) assesses the use of self-regulatory and learning strategies across typical academic situations. Participants rate their agreement with 30 statements using a five-point Likert scale ranging from 1 (Never) to 5 (Always). An example item is: "I summarize what I read to better understand it." The instrument is structured into four first-order dimensions: (1) Basic Learning Self-Regulation Strategies (8 items), (2) Visual Elaboration and Summarizing Strategies (8 items, including 4 reverse-scored items), (3) Deep Information Processing Strategies (8 items), and (4) Social Learning Self-Regulation Strategies (6 items).

An exploratory factor analysis (Maximum Likelihood, Promax rotation) conducted in this study confirmed the expected multifactorial structure, with an adequate fit (KMO = .799, Bartlett's test $p < .001$). Although the initial solution extracted multiple factors, in line with the scale's complex design, we also conducted an EFA forcing a single factor to assess the viability of a composite score. This one-factor solution explained 23.9% of the variance.

Despite the relatively low variance explained by the single-factor solution, we computed a global composite score aggregating all items for three reasons. First, this approach is recommended by the original validation studies (Panadero et al., 2021) as a way to represent an overall index of deep learning self-regulation. Second, the internal consistency of the full scale was high (Cronbach's alpha = .861, McDonald's omega = .879), indicating that the items reliably measure a common construct when aggregated. Third, the low variance explained by a single factor is expected for a multidimensional instrument with 30 items spanning four conceptually distinct dimensions; it reflects the instrument's designed complexity rather than poor measurement quality. Using four separate subscale scores in our moderation analyses would introduce multicollinearity concerns and reduce statistical power given our sample size. We acknowledge that future studies with larger samples could productively examine subscale-level moderation effects.

5.3.7. *Task-specific self-efficacy questionnaire*

We designed an ad hoc self-efficacy questionnaire specifically tailored to the course content and the evaluated task. This approach is consistent with recommendations emphasizing the contextual specificity of self-efficacy measures (Bandura, 2006; Schunk & Pajares, 2009). The instrument comprised 8 items rated on a five-point Likert scale ranging from "strongly disagree" to "strongly agree". An example item is: "I believe I am capable of planning a suitable route for the group under my responsibility".

An exploratory factor analysis using Maximum Likelihood extraction and Promax rotation indicated that a single-factor solution was appropriate, explaining 50.56% of the variance. Factor loadings ranged from .464 to .949. Reliability in this sample was appropriate, with Cronbach's alpha = .859 and McDonald's omega = .860. These results support the use of a composite score representing overall self-efficacy related to this specific academic task.

5.3.8. *Perceived accuracy items*

After completing their self-assessment, students responded to two Likert-scale items (on a scale from 1 to 7). The first item measured their perceived ability to self-assess accurately, with responses ranging from "I am completely unable" to "I am completely able". The second item assessed the extent to which students believed their self-assessment aligned with the instructor's evaluation, with options from "completely different from the instructor's" to "completely identical to the instructor's". While these items provided a simple way to gauge students' perceived accuracy, we acknowledge that using only two items may limit the measurement's sensitivity and reliability. Future research could benefit from more comprehensive instruments or qualitative approaches to capture a richer understanding of students' perceptions.

5.4. *Procedure*

At the beginning of the semester, participants completed the EMSR-Q, the DLS-Q, and our ad-hoc self-efficacy questionnaire online. Their prior academic achievement was also recorded.

In the fourth week, the instructor reintroduced the assignment (as in the first session), providing a detailed explanation, distributing printed copies of the rubric, and allocating 20–30 min of that class for students to begin working individually using their laptops or tablets with keyboards, which they were asked to bring or borrow from the university library. The instructor moved around the classroom, addressing students' questions and supporting them as they began working on the assignment. The rubric was also made available on the university's Learning Management System (LMS).

Over the following weeks, the instructor continued to guide students through the task development process, dedicating approximately 20–30 min per session across two additional class sessions. These sessions focused on addressing students' questions, briefly reviewing each student's work, providing feedback, clarifying assessment criteria, and encouraging both self-assessment and peer assessment activities.

Throughout this process, students received continuous guidance and support to ensure they clearly understood the rubric and assessment expectations. Before the final class of the semester, students submitted their assignments. This preparation aimed to help them develop a solid grasp of the evaluation criteria and enable them to accurately assess their own performance in the assignment. It should be noted that this extensive instructional scaffolding, provided equally to all students, may have established a strong baseline of criterion understanding, potentially limiting the marginal effect of any additional incentive manipulation.

In the final class session, students self-assessed their submitted assignment using the rubric. The experimental group was informed verbally and in writing, immediately before the self-assessment, that they would receive a grade incentive depending on the accuracy of their summative self-assessment. No reminders were provided after this announcement, and the control group was not informed of the incentive at any point. Specifically:

- If the difference between student and instructor grades was ±0.33 points or less, the assignment's weighted contribution to the final grade was increased by 15%.
- If the difference was between ±0.34 and ±0.66 points, the weighted contribution was increased by 10%.
- If the difference was between ±0.67 and ±1.00 points, the weighted contribution was increased by 5%.

To avoid disadvantaging any participant, the same incentive was ultimately applied to the control group's final grades, although this was not communicated to them during the study period. All comparisons and analyses related to self-assessment accuracy were conducted on the grades before applying these incentives in the control group.

Each student was given as much time as needed to complete the self-assessment, which typically took at least 10 min. No individual feedback was provided between the self-assessment and the final exam. After self-assessing, students answered the two items about their perceived accuracy and also completed the self-efficacy questionnaire again (Fig. 1).

5.5. *Data analyses*

All statistical analyses were conducted using IBM SPSS (version 29),

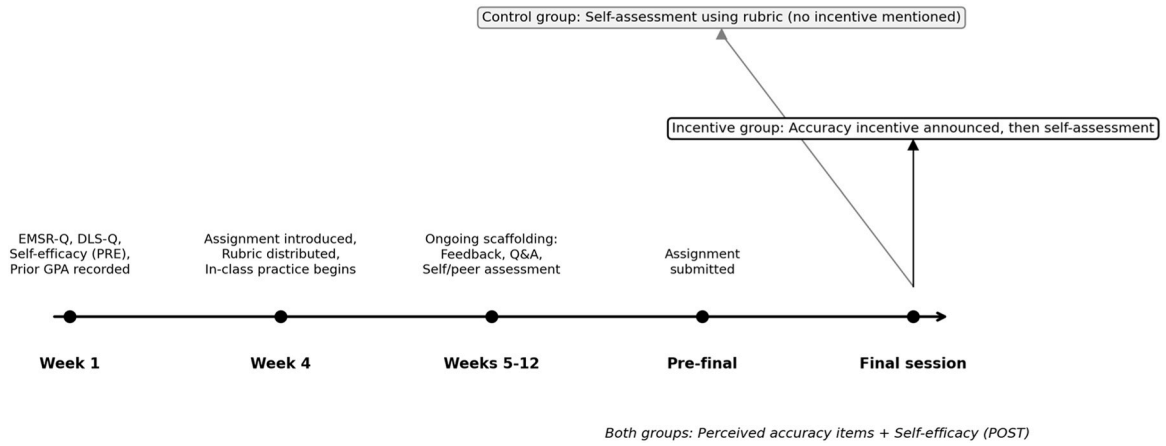


Fig. 1. *Study Procedure Timeline. Note.* The incentive was communicated verbally and in writing immediately before the self-assessment exercise in the final session. The control group completed the same self-assessment without being informed of any incentive.

except the Bayes factor, the equivalence test, and the multilevel models, which were computed in Python 3 (pingouin, SciPy, and statsmodels). As a preliminary step, exploratory factor analyses were performed on the EMSR-Q, DLS-Q, and self-efficacy questionnaire responses to assess the factor structure and item loadings in this sample. Reliability analyses were also conducted, including Cronbach's alpha and McDonald's omega coefficients, to evaluate internal consistency. McDonald's omega coefficients were calculated using Jamovi (version 2.4). Descriptive statistics and bivariate correlations were calculated for all variables.

To address RQ1, independent samples *t*-tests compared the Absolute Deviation Index between the control and experimental groups.

For RQ2, separate multiple linear regression models were estimated for each individual characteristic. The general form of these models was:

Accuracy = b0 + b1(Condition) + b2(Characteristic) + b3(Condition x Characteristic) + e

where Accuracy is the Absolute Deviation Index, Condition is coded 0 = control and 1 = incentive, and Characteristic represents the individual variable of interest (gender, prior GPA, Learning SR, Avoidance SR, DLS-Q composite, or self-efficacy).

Continuous characteristics were centered at their sample means. With this coding, b_1 estimates the difference between conditions at the mean of the characteristic, b_2 estimates the association between the characteristic and accuracy in the control condition, and b_3 estimates how much that association differs in the incentive condition. Because the Absolute Deviation Index measures the distance between the self-assigned and the instructor-assigned scores, positive coefficients indicate larger deviations, that is, lower accuracy. Self-efficacy entered these models as measured at the beginning of the semester. To help interpret the significant interactions, we conducted two post hoc analyses. First, we re-estimated the same moderation model with the instructor-assigned score as the outcome, to check whether the characteristic by condition interaction was already present in a variable that the incentive could not have affected. Second, we re-estimated the accuracy model adding the instructor-assigned score as a covariate.

For RQ3, hierarchical regression models took the form:

Step 1: Accuracy = b0 + b1(Perceived) + b2(Condition) + e

Step 2: added the interaction term Perceived x Condition

For RQ4, the moderation model was:

Final Exam = b0 + b1(Accuracy) + b2(Condition) + b3(Accuracy x Condition) + e

Assumptions of normality, homoscedasticity, and linearity were checked for all regression models. A post hoc sensitivity analysis was performed using G*Power to determine the minimum detectable effect size. Because a non-significant frequentist test cannot, on its own, distinguish a genuine absence of effect from insufficient sensitivity, we complemented the primary between-group comparison (RQ1) with two additional analyses. First, we computed a Bayes factor for the independent-samples comparison (JZS prior, default scale r = 0.707) to quantify the relative evidence for the null versus the alternative hypothesis. Second, we conducted a two one-sided tests (TOST) equivalence analysis, using a smallest effect size of interest of d = 0.20, to test whether the observed difference was statistically equivalent to a negligible effect. Together, these analyses allow us to characterize the informativeness of the null result rather than assume it. In addition, to account for the nesting of students within intact classroom groups, we estimated linear mixed-effects models with a random intercept for classroom group (k = 7; REML estimation). An unconditional model was used to quantify the intraclass correlation coefficient (ICC) for the Absolute Deviation Index, and the primary comparison for RQ1 was then re-estimated with condition as a fixed effect. Given the small number of clusters, these models complement rather than replace the primary analyses, and the moderation models (RQ2) were not re-estimated in a multilevel framework. We note that interaction effects in moderation analyses typically require substantially larger samples to detect than main effects. Given our sample size, the moderation findings should be treated as exploratory and requiring replication. All scripts and syntax files are available from the corresponding author upon reasonable request.

6. Results

A total of 142 students participated in the study (n control = 65, n experimental = 77). A post hoc sensitivity analysis using G*Power indicated that, given this final sample size, the minimum detectable effect size with 80% power and α = .05 was d = 0.48. Therefore, small effects (d < 0.48) may have remained undetected. Descriptive statistics and inferential analyses for each research question are presented below.

6.1. Descriptives: overall accuracy

In this section we used an aggregated index including data from both the incentive and control groups, and therefore does not isolate or reflect any specific effect of the intervention. We report it here solely to give a broad snapshot of how students generally performed in terms of self-assessment precision. Based on the Signed Deviation Index, students, on average, assigned themselves a total score 0.55 points higher than their instructor's assigned score (SD = 1.20). A majority of students (n = 97, 68.3%) overestimated their performance by assigning themselves a higher score than their instructor. Students who overestimated their performance (M = 7.12, SD = 1.34) had statistically significantly lower instructor-assigned scores than those who underestimated their performance (M = 8.70, SD = 0.92), t(119.808) = -8.236, p < .001. The

raw difference between instructor-assigned and self-assigned scores ranged from −1.66 to 0.13 across the different rubric criteria. The index showed a negative value for 14 out of the 17 criteria, indicating that students generally overestimated their scores across most criteria. The correlation between instructor-assigned scores and student self-assigned scores was 0.557 ($p < .001$). This correlation ranged from 0.298 to 0.666 and was positive and statistically significant at the .001 level across all 17 criteria.

It is worth noting that instructor-assigned scores were clustered in the upper performance bands, with most students scoring between 7 and 9 on the 0–10 scale. This distribution creates an asymmetry in the potential for overestimation versus underestimation: a student scored at 8.5, for example, can overestimate by at most 1.5 points but can underestimate by up to 8.5 points. This restricted range may have compressed variance in accuracy scores, particularly on the overestimation side, and should be considered when interpreting the magnitude of between-group differences.

In subsequent sections, we analyze self-assessment accuracy separately by group to accurately examine any impact of the incentive on students' precision.

6.2. RQ1. Does Providing an Accuracy-based Incentive Enhance the Accuracy of Students' Summative Self-Assessments?

As described in the Methods section, the Absolute Deviation Index provides a robust measure of self-assessment accuracy by accounting for the overall magnitude of deviations, regardless of their direction (underestimation or overestimation). This index was specifically employed to assess the impact of the incentive on students' summative self-assessment accuracy. The results showed that the Absolute Deviation Index had an overall mean of 1.74 ($SD = 0.74$). No statistically significant difference was observed between the control ($M = 1.64$, $SD = 0.75$) and the incentive condition ($M = 1.82$, $SD = 0.72$) in self-assessment accuracy, $t(140) = 1.469$, $p = .144$. Notably, the incentive condition exhibited a slightly larger overall distance from their instructor's scores (Cohen's $d = 0.25$).

The correlation between students' self-assessment scores and their instructor's assigned scores was also analyzed as an indicator of self-assessment accuracy. Consistent with the results based on the Absolute Deviation Index, the correlation was stronger for the control ($r(65) = .633$, $p < .001$) than for the incentive condition ($r(77) = .462$, $p < .001$).[1] To clarify the interpretation of this null result, we examined it with two complementary analyses. The Bayes factor indicated only anecdotal evidence in favor of the null ($BF_{01} = 2.07$), and the equivalence test was non-significant (TOST $p = .61$), indicating that an effect as large as $d = 0.20$ could not be statistically excluded. Taken together, these results indicate that the data are inconclusive rather than supportive of a true absence of effect. We therefore do not treat the non-significant comparison as evidence that the incentive is ineffective; H1 was not supported, but neither can it be confidently rejected. This interpretive caution applies throughout the discussion of RQ1.

To further account for the nested structure of the data, we re-estimated this comparison using a linear mixed-effects model with a random intercept for the seven classroom groups (REML estimation). In the unconditional model, 6.4% of the variance in the Absolute Deviation Index was attributable to differences between classroom groups (ICC = .064). In the conditional model, the effect of condition remained non-significant ($B = 0.17$, $SE = 0.19$, $t(5) = 0.87$, $p = .42$), with a larger standard error than in the single-level comparison, as expected once clustering is taken into account. Accounting for the nested structure therefore did not alter the conclusions for RQ1.

[1] The difference between the correlations for the control and the intervention group was not statistically significant, Fisher r-to-z, $z = 1.44$, $p = .151$.

6.3. RQ2. Do Individual Characteristics Influence the Impact of Accuracy-based Incentives on Students' Summative Self-Assessment Accuracy?

To examine whether the effect of accuracy-based incentives on summative self-assessment accuracy varied across individual characteristics, a series of multiple linear regression analyses were conducted. In each model, the intervention effect (control vs. intervention group) was analyzed along with the main effect of a specific individual characteristic (i.e., gender, prior academic achievement, self-regulatory skills, and self-efficacy), as well as the interaction between the intervention and that characteristic. This approach allowed for the identification of potential moderating effects, determining whether certain subgroups of students benefited more from the intervention than others. Table 1 presents the results of these six regression models.

Learning Self-Regulation (EMSR-Q) and the DLS-Q composite interacted with condition (Table 1). For Learning Self-Regulation, the coefficient in the control condition was positive ($B = 0.038$, $p = .004$). Because the outcome is a deviation index, this means that in the control condition students with higher learning-oriented self-regulation scores deviated more from the instructor's score, that is, they were less accurate ($r = .36$, $p = .005$, $n = 57$). The interaction was negative and

Table 1
Regression Models on Students' Self-Assessment Accuracy with Intervention, Student Individual Characteristics, and Relevant Interaction Terms as Predictors.

	B	*SE of B*	*β*	*p-value*
Group (Control) Intervention	0.074	0.140	0.050	.599
Gender (Male) Female	-0.146	0.216	-0.082	.500
Interaction term (Group*Gender)	0.506	0.299	0.217	.092
Intercept	1.672	0.104		< .001
R^2	4.0%			
Group (Control) Intervention	0.147	0.115	0.099	.205
Previous GPA	-0.418	0.140	-0.389	.003
Interaction term (Group*GPA)	-0.024	0.174	-0.018	.891
Intercept	1.662	0.085		< .001
R^2	17.6%			
Group (Control) Intervention	0.152	0.127	0.102	.234
Learning self-regulation style	0.038	0.013	0.369	.004
Interaction term (Group*Learning style)	-0.064	0.018	-0.453	< .001
Intercept	1.652	0.094		< .001
R^2	10.8%			
Group (Control) Intervention	0.158	0.134	0.106	.239
Avoidance self-regulation style	-0.004	0.010	-0.051	.682
Interaction term (Group*Avoidance style)	0.007	0.015	0.061	.625
Intercept	1.638	0.099		< .001
R^2	1.4%			
Group (Control) Intervention	0.158	0.130	0.106	.226
Deep learning self-regulation strategies	0.016	0.007	0.298	.018
Interaction term (Group*Deep learning)	-0.023	0.009	-0.309	.014
Intercept	1.637	0.096		< .001
R^2	6.5%			
Group (Control) Intervention	0.148	0.133	0.100	.267
Self-efficacy	0.026	0.018	0.190	.155
Interaction term (Group*Self-efficacy)	-0.044	0.025	-0.239	.074
Intercept	1.654	0.099		< .001
R^2	3.7%			

Note. Condition coded 0 = control, 1 = incentive. Continuous characteristics were centered at their means, so the Group coefficient is the difference between conditions at the mean of the characteristic, and the coefficient of the characteristic is its association with the Absolute Deviation Index in the control condition. Positive coefficients indicate larger deviations (lower accuracy). Self-efficacy is the measure taken at the beginning of the semester. $n = 142$ (gender), 141 (prior GPA), and 127 (self-regulation styles, deep learning strategies, and self-efficacy).

statistically significant ($B = -0.064$, $p < .001$), so the association reversed in the incentive condition, where higher scores were associated with smaller deviations (r = -.25, p = .034, n = 70). Pooled across conditions, Learning Self-Regulation was unrelated to accuracy (r = .04, p = .63).

The DLS-Q composite showed the same pattern: a positive coefficient in the control condition ($B = 0.016$, p = .018; r = .30, p = .022), a negative interaction ($B = -0.023$, p = .014), a non-significant negative association in the incentive condition (r = -.14, p = .26), and no association when conditions were pooled (r = .08, p = .38).

Because the conditions were composed of intact classroom groups, we examined whether these interactions could be attributed to the incentive. Two post hoc analyses suggest that these interactions cannot be confidently attributed to the incentive. First, the same interactions emerged when the instructor-assigned score, which the incentive could not have affected, was used as the outcome: higher Learning Self-Regulation was associated with lower instructor scores in the control condition (r = -.23) and with higher instructor scores in the incentive condition (r = .22), interaction $B = 0.093$, $SE = 0.036$, p = .011; for the DLS-Q composite, the corresponding correlations were r = -.24 and r = .24, interaction $B = 0.050$, $SE = 0.018$, p = .008. Since larger deviations were concentrated among students with lower instructor scores (see Descriptives), the self-regulation by condition interactions on accuracy mirror a pre-existing difference between the classroom groups in how self-regulation related to task performance. Second, when the instructor-assigned score was added as a covariate to the accuracy models, the interaction for the DLS-Q composite was no longer significant ($B = -0.005$, $SE = 0.007$, p = .476), and the interaction for Learning Self-Regulation remained significant but was reduced by half ($B = -0.031$, $SE = 0.013$, p = .015). Taken together, these results indicate that the moderation pattern reflects characteristics of the groups allocated to each condition rather than a differential response to the incentive.

In contrast, in the separate models (Table 1), the remaining individual characteristics (i.e., gender, prior academic achievement, avoidance self-regulation style, and self-efficacy) did not show statistically significant interactions with the intervention. The absence of significant findings for gender (p = .092), prior academic achievement (p = .891), avoidance self-regulation (p = .625), and self-efficacy (p = .074) suggests that the intervention did not differentially impact students based on these factors. Notably, prior academic achievement was a significant predictor of self-assessment accuracy overall ($B = -0.418$, p = .003), with students who had higher prior GPA exhibiting greater alignment between their self-assessment and instructor-assigned scores, regardless of the intervention. These findings should be interpreted with caution, as the study's sample size and sensitivity analysis indicate that smaller effects may have remained undetected.

Overall, H2 was partially supported. Prior academic achievement predicted accuracy as expected, and neither self-efficacy nor gender moderated the condition effect, also as expected. Contrary to the hypothesis concerning self-regulation, students with stronger self-regulatory skills were not more accurate overall and did not show a greater benefit from the incentive. The self-regulation by condition interactions were statistically significant, but they were also present on the instructor-assigned score and are therefore better understood as differences between the classroom groups allocated to each condition than as effects of the incentive. See Fig. 2 for the interaction pattern and Table 2 for the combined regression model.

6.4. RQ3. Does Students' Perceived Accuracy Predict Their Actual Summative Self-Assessment Accuracy, and is this Relationship Moderated by the Incentive Condition?

Student responses to the two questions regarding their perceived ability to self-assess accurately and the extent to which they believed their self-assessment aligned with the instructor's evaluation are

Table 2
Combined Multiple Regression Model Predicting the Absolute Deviation Index.

Predictor	*B*	*SE*	*t*	*p*
Intercept	1.717	0.097	17.742	< .001
Condition	-0.023	0.131	-0.173	.863
Gender	0.055	0.217	0.253	.800
Prior Academic Achievement	-0.581	0.161	-3.596	< .001
Learning Self-Regulation	0.056	0.018	3.051	.003
Avoidance Self-Regulation	-0.034	0.012	-2.809	.006
Deep Learning Strategies	0.003	0.008	0.360	.720
Self-Efficacy	0.008	0.018	0.465	.643
Condition x Gender	0.466	0.310	1.503	.136
Condition x Prior Achievement	0.062	0.202	0.309	.758
Condition x Learning SR	-0.068	0.023	-3.027	.003
Condition x Avoidance SR	0.040	0.016	2.512	.013
Condition x Deep Learning	0.001	0.011	0.131	.896
Condition x Self-Efficacy	-0.013	0.026	-0.494	.623

Note. N = 125. R^2 = .376, Adjusted R^2 = .303, F(13, 111) = 5.14, p < .001. Condition coded 0 = control, 1 = incentive. Continuous predictors were centered at their means, so the coefficients of the student characteristics are their associations with the Absolute Deviation Index in the control condition. Positive coefficients indicate larger deviations (lower accuracy).

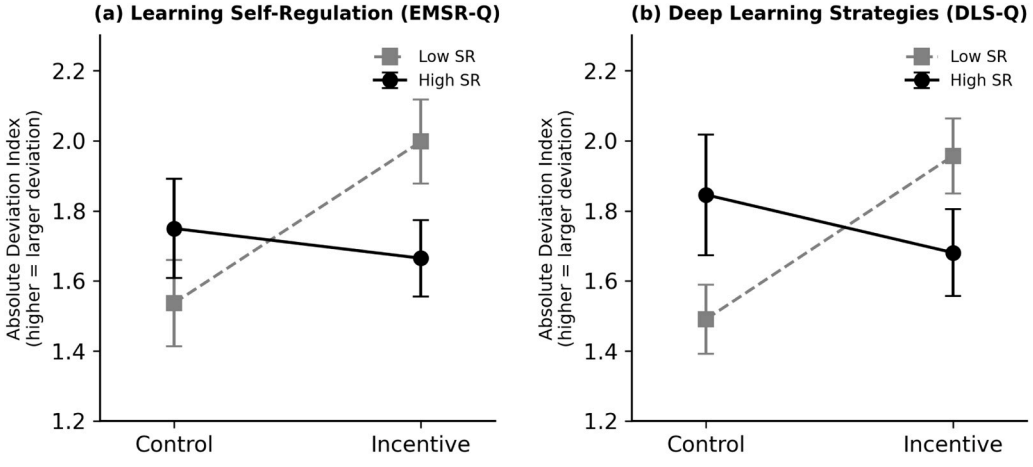


Fig. 2. *Interaction Between Self-Regulation and Condition on the Absolute Deviation Index. Note.* Higher values on the Absolute Deviation Index indicate greater deviation from instructor scores (i.e., lower accuracy). Panel (a) shows the interaction between Learning Self-Regulation (EMSR-Q; median split for visualization) and condition. Panel (b) shows the interaction between Deep Learning Strategies (DLS-Q; median split) and condition. Error bars represent standard errors. The crossover pattern indicates that the relationship between self-regulation and accuracy differed by condition.

presented in Fig. 3[2]. Overall, students reported high levels of confidence in their ability to accurately assess themselves and they tended to believe that their self-assessment aligned with their instructor's evaluation.

To examine whether condition (incentive vs. control) moderates the relationship between perceived self-assessment measures and actual self-assessment accuracy, two separate moderation analyses were conducted using hierarchical multiple regression. The first analysis assessed moderation in the relationship between perceived ability to self-assess accurately and actual self-assessment accuracy, while the second analysis examined whether condition moderated the relationship between perceived self-assessment alignment with the instructor's evaluation and actual accuracy. In both cases, perceived self-assessment scores were centered, and an interaction term was created by multiplying the centered predictor by the experimental condition (coded as 0 = control, 1 = incentive).

6.4.1. Moderation Analysis 1: Perceived Ability to Self-Assess Accurately

In the first hierarchical regression, Model 1 included perceived ability to self-assess accurately (centered) and condition (incentive vs. control) as predictors. The model was not statistically significant, $F(2, 138) = 0.921$, $p = .401$, and accounted for only 1.3% of the variance in actual accuracy, $R^2 = .013$. Neither perceived ability to self-assess accurately ($B = 0.017$, $SE\ B = 0.068$, $\beta = 0.021$, $t = 0.247$, $p = .805$) nor condition ($B = 0.166$, $SE\ B = 0.124$, $\beta = 0.113$, $t = 1.336$, $p = .184$) significantly predicted actual accuracy. In Model 2, the interaction term (Perceived Ability to Self-Assess Accurately × Condition) was added to the regression equation to test for moderation. The addition of the interaction term did not significantly improve the model, $\Delta R^2 = .015$, $F(1, 137) = 2.08$, $p = .151$. The interaction term itself was not statistically significant ($B = -0.195$, $SE\ B = 0.136$, $\beta = -0.171$, $t = -1.442$, $p = .151$), indicating that condition did not moderate the relationship between perceived ability to self-assess accurately and actual self-assessment accuracy.

6.4.2. Moderation Analysis 2: Perceived Self-Assessment Alignment with the Instructor's Evaluation

In the second hierarchical regression, Model 1 included perceived self-assessment alignment with the instructor's evaluation (centered) and condition (incentive vs. control) as predictors. The model was again not statistically significant, $F(2, 137) = 1.076$, $p = .344$, explaining only 1.5% of the variance in actual self-assessment accuracy, $R^2 = .015$. Neither perceived self-assessment alignment ($B = 0.021$, SE $B = 0.070$, $\beta = 0.025$, t = 0.298, $p = .766$) nor condition ($B = 0.180$, SE $B = 0.124$, $\beta = 0.123$, t = 1.449, $p = .150$) significantly predicted actual accuracy. In Model 2, the interaction term (Perceived Self-Assessment Alignment × Condition) was included. The model did not show a significant improvement, $\Delta R^2 = .010$, $F(1, 136) = 1.41$, $p = .238$. The interaction term was not statistically significant ($B = -0.167$, SE $B = 0.141$, $\beta = -0.151$, t = -1.185, $p = .238$), suggesting that condition did not moderate the relationship between perceived self-assessment alignment and actual self-assessment accuracy.

Therefore, both moderation analyses failed to provide evidence that the incentive condition significantly altered the relationship between perceived self-assessment (ability or alignment) and actual self-assessment accuracy. In both cases, the interaction terms were not significant ($p > .05$), and the addition of the interaction terms to the models did not significantly improve the explained variance in actual accuracy ($\Delta R^2 = .015$ and .010, respectively). These findings suggest that the relationship between perceived self-assessment measures and actual accuracy is consistent across both the incentive and control conditions, indicating that a pooled analysis may be appropriate to evaluate these relationships across all participants without separating by condition.

6.4.3. Pooled regression analysis

A pooled regression analysis was conducted to examine whether Perceived Ability to Self-Assess Accurately and Perceived Self-Assessment Alignment with the Instructor's Evaluation predicted Actual Self-Assessment Accuracy, considering all participants together without separating by condition. The overall regression model was not statistically significant, $F(2, 137) = 0.073$, $p = .929$, and accounted for only 0.1% of the variance in actual self-assessment accuracy, $R^2 = .001$. The adjusted R^2 (-.014) suggests that including these predictors does not improve the model's explanatory power. Neither Perceived Ability to Self-Assess Accurately ($B = 0.025$, SE $B = 0.081$, $\beta = 0.031$, t = 0.308, $p = .759$) nor Perceived Self-Assessment Alignment with the Instructor's Evaluation ($B = 0.002$, SE $B = 0.084$, $\beta = 0.003$, t = 0.030, $p = .976$) significantly predicted actual self-assessment accuracy.

In conclusion, H3 was not supported: perceived self-assessment accuracy did not predict actual self-assessment accuracy. As expected, the incentive condition did not moderate this relationship.

6.5. RQ4. Does the Accuracy of Students' Self-Assessments Predict Their Subsequent Academic Performance on the Final Examination?

To investigate this RQ we conducted a series of analyses. First, we computed means and standard deviations for the Final Exam Scores: Students in the control group had a mean final exam score of 6.41 ($SD = 1.39$), while students in the incentive group had a mean score of 7.13 ($SD = 1.39$). Additionally, an independent samples t-test comparing final exam performance between the two conditions revealed a statistically significant difference, $t(140) = -3.08$, $p = .002$, indicating that students in the incentive condition outperformed those in the control condition.

Next, we conducted Pearson correlation analyses separately for each condition. Control $r(65) = -0.396$, $p = .001$, and Incentive condition $r(77) = -0.353$, $p = .002$. These findings indicated that, in both conditions, students with greater self-assessment accuracy tended to achieve higher final exam grades, but the strength of the relationship was slightly stronger in the control condition.

Therefore, to determine whether the incentive condition moderates the relationship between self-assessment accuracy and final exam performance, we conducted a hierarchical multiple regression analysis. In the step 1 we entered Accuracy Index (centered) and condition (control vs. incentive) as predictors. And in the step 2 we added the interaction term (Accuracy × Condition) to test for moderation. The overall model was statistically significant, $F(3, 138) = 11.10$, $p < .001$, $R^2 = .194$, indicating that self-assessment accuracy and condition together explain approximately 19.4% of the variance in final exam performance. However, the interaction term added in step 2 was not statistically significant, $B = 0.054$, $SE = 0.298$, $\beta = 0.020$, $t = 0.181$, $p = .857$, indicating that the relationship between self-assessment accuracy and final exam performance did not differ significantly between the incentive and control conditions.

Finally, we conducted a second moderation analysis to test whether prior academic achievement moderated the relationship between self-assessment accuracy and final exam performance. In step 1 we entered Accuracy index (centered), prior academic achievement (centered), and condition as predictors. And in step 2 we added the interaction term (Accuracy index × Prior Academic Achievement) to test for moderation. The overall model was statistically significant, $F(4, 136) = 12.38$, $p < .001$, $R^2 = .267$, indicating that self-assessment accuracy, prior academic achievement, and condition together explain 26.7% of the variance in final exam performance. The Accuracy index was a significant negative predictor of final exam performance, $B = -0.466$, $SE = 0.158$, $\beta = -0.241$, t = -2.95, $p = .004$. Prior academic achievement

[2] High values in the Likert scale indicate high levels of perceived ability to self-assess accurately and perceived self-assessment alignment with the instructor's evaluation.

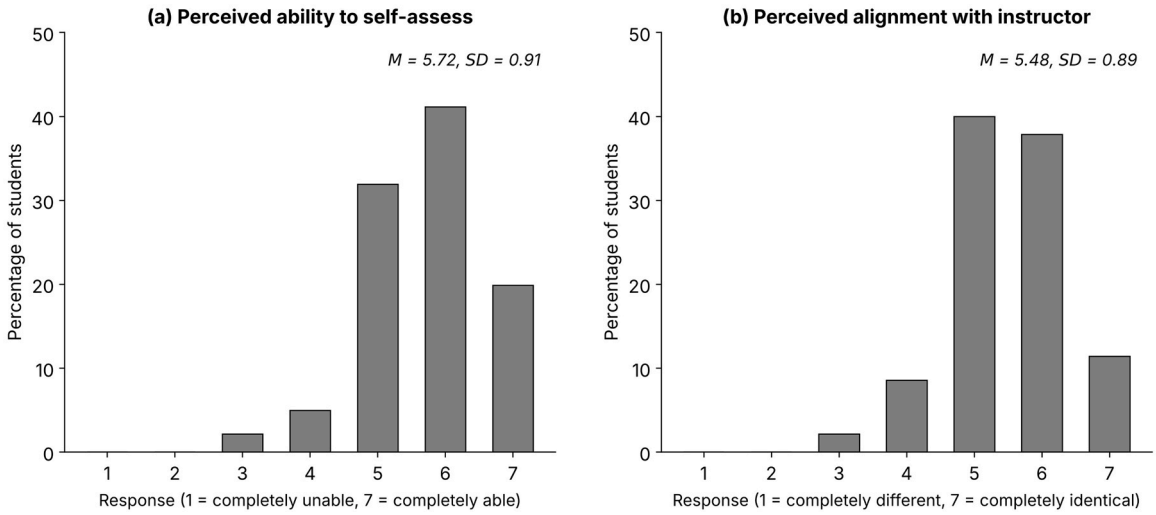


Fig. 3. *Students' Perceived Self-Assessment Accuracy. Note.* Panel (a) shows the distribution of responses to the item measuring perceived ability to self-assess accurately. Panel (b) shows the distribution of responses to the item measuring perceived alignment with the instructor's evaluation. Both items used a 7-point Likert scale. Neither perceived accuracy measure significantly predicted actual self-assessment accuracy (pooled $R^2 = .001$, $p = .929$).

was a significant positive predictor, $B = 0.606$, $SE = 0.174$, $\beta = 0.292$, $t = 3.48$, $p < .001$. The interaction term was not statistically significant, $B = 0.025$, $SE = 0.229$, $\beta = 0.008$, $t = 0.107$, $p = .915$, indicating that the association between self-assessment accuracy and final exam performance does not depend on students' prior academic achievement.

In sum, H4 was supported: self-assessment accuracy significantly predicted final exam performance, and this relationship was not significantly moderated by either prior academic achievement or the incentive condition.

7. Discussion

This study explored whether providing an accuracy-based incentive communicated at the moment of self-assessment improved students' summative self-assessment precision. Additionally, we explored the role of individual differences in moderating the effectiveness of this incentive, analyzed the relationship between students' perceived and actual self-assessment accuracy, and investigated whether self-assessment accuracy predicted students' subsequent academic performance. Three features constrain interpretation throughout: the allocation of intact classroom groups by lecture day, the communication of the incentive immediately before the self-assessment, and the absence of direct measures of the processes through which an incentive could operate, which means that the mechanisms discussed below are hypotheses rather than findings (see Limitations).

Before addressing our research questions, our initial descriptive analyses highlighted important patterns in self-assessment accuracy. Specifically, students generally tended to overestimate their grades, and those who did so to a greater extent received lower instructor-assigned scores. This pattern is consistent with the concept of calibration bias, whereby students with weaker performance tend to display greater overestimation (Kruger & Dunning, 1999; León, Vallejo, et al., 2021; Pinedo et al., 2025). These findings reinforce prior research on self-assessment accuracy biases (e.g., Brown & Harris, 2013; Panadero et al., 2016) and suggest that overestimation may be linked to lower levels of task-relevant competence, in line with the calibration literature, which holds that the knowledge required to perform well is also needed to judge one's own performance accurately (Dunning et al., 2004).

7.1. Does Providing an Accuracy-Based Incentive at the Point of Self-Assessment Enhance Accuracy? (RQ1)

Contrary to our initial expectation (H1), we found no clear statistical evidence that the incentive improved summative self-assessment accuracy. The control and incentive groups did not differ significantly in their Absolute Deviation Index, and the correlation between instructor- and self-assigned grades was, in fact, higher in the control group than in the incentive group. As this is the first study to test this type of incentive, the result is informative mainly as a first data point: under the conditions tested, the intuitive expectation that an accuracy incentive improves alignment was not borne out. As shown by the Bayesian and equivalence analyses, however, the data cannot establish that the incentive has no effect, and the explanations offered below should be read with that in mind.

Several factors may explain this null result. None of them was measured directly, so they are offered as hypotheses for future work rather than as explanations supported by our data. First and most importantly, the incentive was communicated immediately before the self-assessment, giving students no opportunity to adjust their preparation, study the rubric more carefully, or practice calibration. This timing effectively tests whether a last-minute motivational prompt shifts scoring behavior, which is a weaker manipulation than a sustained incentive integrated throughout the learning process. The null result therefore does not rule out the possibility that incentives embedded earlier in the instructional process might be effective. Indeed, research on self-assessment interventions consistently shows that sustained, scaffolded practice over time produces stronger effects than one-time activities (Panadero et al., 2017; Yan et al., 2022).

Second, all students in both conditions received extensive instructional scaffolding throughout the course, including rubric explanation, in-class practice, formative feedback, and opportunities for self- and peer assessment. This strong baseline of criterion understanding may have left limited room for an additional incentive to produce incremental gains, effectively creating a ceiling on the potential marginal effect of the manipulation. This interpretation aligns with findings showing that prior training and criterion familiarity are key determinants of self-assessment accuracy (León, Vallejo, et al., 2021; Panadero et al., 2017), and that the marginal benefit of additional interventions decreases as students become more experienced with assessment criteria (Ross, 2006). The pattern of individual differences is compatible with this reading: the only student characteristic consistently associated with accuracy was prior academic achievement, which

suggests that much of the remaining inaccuracy stemmed from limits in evaluative competence, the source on which, as argued in the introduction, an incentive cannot act.

Third, the 10% contribution of the self-assessment score to the task grade may have created a general tendency toward inflation. Because this contribution was identical in both conditions, however, it cannot by itself explain the absence of a between-group difference: the comparison still isolates the additional accuracy-based incentive as the only systematic difference between groups. It does, nonetheless, describe a grading environment in which the motivation to be accurate competed with a shared motivation to inflate, which may have constrained the overall room for the accuracy incentive to operate. This shared incentive to inflate also likely raised self-scores across the whole sample, so the general overestimation observed in the descriptive analyses should be read partly as a feature of the grading conditions rather than as evidence that students are inherently poor self-assessors. Which grading conditions support or hinder accurate self-assessment is itself an open empirical question (see Practical and Research Implications).

Finally, it is important to consider the conceptual distinction between metacognitive monitoring and social prediction discussed in the introduction. Even if the incentive had shifted students' behavior, such a shift might have reflected changes in strategic reporting -i.e., better prediction of instructor scoring- rather than improved metacognitive self-knowledge (Rouy et al., 2022). We note that neither of the candidate mechanisms through which an incentive could operate, deeper engagement with the rubric or improved prediction of the instructor's score, was measured in this study; both remain untested explanations rather than findings. Given that our data are inconclusive with respect to RQ1 (see Results), we refrain from interpreting the null as evidence that last-minute prompts cannot affect self-assessment. What can be said is more limited: under the conditions tested, and with the sensitivity our sample affords, we found no detectable benefit of the incentive, and the question of whether such prompts can override established self-assessment tendencies remains open for adequately powered, embedded designs to address.

7.2. Do Individual Characteristics Influence the Impact of Accuracy-based Incentives? (RQ2)

Self-regulatory variables did not behave as hypothesized. Students scoring higher on learning self-regulation (EMSR-Q) and deep learning strategies (DLS-Q) were not more accurate overall, and the significant interactions with condition did not reflect a differential response to the incentive. In the control groups, higher self-regulation scores were associated with larger deviations from the instructor's score; in the incentive groups, the association ran in the opposite direction. The follow-up analyses showed that the same interactions were present on the instructor-assigned score, which the incentive could not have affected, and that adjusting for that score removed the DLS-Q interaction and halved the one for learning self-regulation. The most parsimonious reading is that the classroom groups allocated to each condition differed in how self-regulation related to performance on the task, and that this difference carried over to the deviation index because students with lower scores overestimated more. These interactions therefore say little about the incentive. They do illustrate a risk inherent to designs with intact groups: a statistically significant moderation effect can arise from differences between the groups assigned to each condition rather than from the manipulation.

Two further points follow. First, the result bears on the rationale for the incentive outlined in the introduction. If an incentive acts on the reporting decision rather than on evaluative competence, its benefit should be largest for students who can judge their work accurately, and self-regulatory skills were our proxy for that capacity. The present data cannot test this account: self-reported self-regulation was unrelated to accuracy once conditions were pooled, which contrasts with the positive association between study strategies, engagement, and accuracy reported by León, Augusto-Landa, et al. (2021), and the moderator was confounded with group membership. Second, testing it properly requires random assignment, a baseline measure of self-assessment accuracy, and direct measures of evaluative competence, so that the moderation of an incentive effect can be separated from pre-existing differences between groups.

In the separate models (Table 1), no significant interactions were found involving gender, prior GPA, avoidance self-regulation style, or self-efficacy. Notably, prior academic achievement was a significant predictor of accuracy overall, consistent with prior research (León et al., 2023; Pinedo et al., 2025), but it did not moderate the incentive effect.

7.3. Does perceived accuracy predict actual accuracy? (RQ3)

Our findings indicate that students generally reported high confidence in their ability to self-assess accurately and perceived their self-assigned scores as strongly aligned with their instructor's grading. However, these subjective perceptions did not significantly predict actual summative self-assessment accuracy, contradicting our initial hypothesis (H3). When analyzing whether the incentive condition moderated this relationship, results revealed no significant interaction effects.

This disconnect between perceived and actual accuracy is a well-documented phenomenon in the calibration literature (Dunning et al., 2004; Kruger & Dunning, 1999; Yan & Brown, 2017). Our findings extend this literature by demonstrating that even when students were incentivized to be accurate, perceived confidence remained unrelated to actual precision. This underscores that subjective confidence is an unreliable indicator of self-evaluative competence, consistent with Sitzmann et al. (2010), who argued that self-assessments of knowledge are better understood as affective judgments than as cognitive evaluations (see also Pinedo et al., 2025). From the perspective of the metacognition versus social prediction distinction, this finding is particularly telling: if students were primarily engaging in social prediction (anticipating instructor scores), we might expect those who reported greater confidence in their alignment to be somewhat better at this prediction task. The absence of any such relationship suggests that perceived accuracy reflects neither metacognitive self-knowledge nor predictive competence, but rather a general dispositional confidence that is disconnected from performance. Because this finding does not depend on the between-group comparison, it is unaffected by the design limitations discussed below and is one of the most robust findings of this study; its practical implication is that self-reported confidence should not be used as a proxy for actual accuracy.

7.4. Does Self-Assessment Accuracy Predict Academic Performance? (RQ4)

Our results revealed a statistically significant association between students' summative self-assessment accuracy and their final exam performance. Students who demonstrated greater accuracy in their self-assigned grades achieved significantly higher scores on the final examination. This finding is consistent with prior research linking self-assessment accuracy to academic outcomes (Brown & Harris, 2013; Falchikov & Boud, 1989; Yan et al., 2022) and with theoretical models that position accurate self-monitoring as a component of effective self-regulated learning (Panadero et al., 2016; Zimmerman & Moylan, 2009). However, this association likely reflects the influence of general academic ability on both variables rather than a direct effect of accuracy on performance. Students who are stronger academically tend to both self-assess more accurately and perform better on exams (León et al., 2023), and our design does not allow us to establish that improving accuracy would cause improved performance.

The incentive group also scored significantly higher on the final exam. Although baseline comparisons showed no significant differences between groups on key pre-intervention variables, the quasi-

experimental design means this difference should be interpreted cautiously, as it may partly reflect unobserved group-level factors rather than the incentive itself.

7.5. Practical and research implications

Our study offers several implications for educators and researchers. From a practical perspective, our findings provide no clear evidence that an outcome-based incentive communicated at the point of self-assessment alters students' self-assessment accuracy. This does not mean that incentives are ineffective in general. Whether incentives integrated earlier in the instructional process, which would give students time to act on them, are more effective is an open question that our design did not test. Until such evidence is available, educators have better-supported options than outcome-based incentives for improving accuracy. These include incremental feedback milestones, guided self-assessment training, and rewarding the quality of reflective reasoning rather than mere alignment with instructor scores (Ross, 2006).

Furthermore, the student characteristic most consistently associated with accuracy was prior academic achievement, which is consistent with the view that judging one's work accurately draws on the same competence that produces good work (Dunning et al., 2004). This points to the value of building evaluative competence directly, through explicit reflection exercises, peer calibration activities, and structured exemplars (Panadero & Jonsson, 2013), rather than relying on students' general self-regulatory dispositions, which were not associated with accuracy in this study. Such activities may better support students in developing more precise and realistic self-assessments over time. Additionally, given the lack of correlation between perceived and actual accuracy, educators should incorporate feedback mechanisms that explicitly challenge students' over- and underestimations, fostering realistic self-evaluation through error detection exercises and cross-checking with peer or instructor assessments (Brown & Harris, 2013). Such feedback should be delivered with care, since discovering an overestimation carries emotional and self-efficacy costs that instructors should anticipate (Pinedo et al., 2026).

From a research perspective, our study highlights several priorities for future work. First, the timing of incentives deserves systematic investigation: studies should compare incentives introduced at different points in the instructional process (e.g., at the start of the course, mid-semester, and at the point of assessment) to determine whether earlier introduction allows calibration adjustment. Such studies should use random assignment at the individual or cluster level with enough classroom groups, a formal manipulation check, a self-assessment component that carries no grade weight, and a preregistered analytical plan. Second, the distinction between metacognitive self-monitoring and social prediction in self-assessment contexts requires empirical disentanglement, potentially through designs that separate students' self-evaluations from their predictions of instructor scores. Third, whether self-regulatory skills moderate the effect of an accuracy incentive remains an open question. The interactions observed here were confounded with differences between classroom groups, so answering it requires random assignment and a baseline measure of self-assessment accuracy, together with direct measures of the competence and reporting processes through which an incentive might operate. Fourth, the conditions under which self-assessment takes place, including whether the self-score carries grade weight, deserve study in their own right, as they plausibly shape both the level of accuracy and the effect of any intervention designed to improve it (Andrade, 2019). Fifth, future work should broaden the notion of accuracy beyond score alignment to include the accuracy of students' self-feedback and of the learning decisions they take after self-assessing, and should consider external assessments or multiple raters as alternative reference points to a single instructor's score.

7.6. Limitations

Several limitations should be considered when interpreting these findings.

First, and most importantly, the quasi-experimental design assigned intact classroom groups to conditions based on scheduled lecture days. This creates a potential confound between condition and group-level factors such as scheduling preferences, cohort composition, classroom climate, and peer dynamics. Although baseline comparisons revealed no significant differences between conditions on prior academic achievement, self-efficacy, self-regulation, deep learning strategies, gender, or academic year (see Participants), unobserved confounds may still exist. Between-group differences, including the final exam score difference, should therefore be interpreted with caution rather than as direct evidence of an incentive effect. To address the nested structure analytically, we re-estimated the primary comparison with a multilevel model including a random intercept for classroom group (ICC = .064; see Results), which did not alter the conclusions. However, with only seven clusters, cluster-level inference remains limited, and the moderation analyses could not be reliably re-estimated in a multilevel framework. The self-regulation by condition interactions reported for RQ2 illustrate this risk: the same interactions were present on the instructor-assigned score, which indicates that they reflect differences between the groups allocated to each condition rather than an effect of the incentive.

Second, the study did not measure the processes through which an incentive could operate. We did not assess students' engagement with the rubric, the care with which they judged their work, their understanding of the assessment criteria, or the extent to which they tried to predict the instructor's score. The study can therefore establish whether alignment changed under the incentive, but not why it would have changed. Had an effect been found, it would not have been possible to tell whether it reflected improved metacognitive monitoring, better understanding of the criteria, greater motivation, or more strategic prediction of the instructor's score. The mechanisms mentioned in the introduction and discussion are hypotheses that motivated the study, not findings, and future studies should measure them directly, for example with process measures taken during self-assessment and with separate questions asking students what score they believe the work deserves and what score they expect the instructor to give.

Third, the incentive was communicated immediately before the self-assessment, which tests only whether a last-minute announcement shifts scoring behavior. A sustained incentive integrated throughout the instructional process would constitute a fundamentally different and stronger manipulation. Our findings speak only to the narrow question of whether awareness of an incentive at the point of self-assessment changes accuracy.

Fourth, no formal manipulation check was conducted to verify that students in the incentive condition understood the incentive structure or found it credible. The absence of such verification limits our confidence that the manipulation functioned as intended.

Fifth, because the self-assessment score contributed 10% of the task grade in both conditions, the grading environment contained a shared motivation to inflate self-scores. As this contribution was constant across conditions, it does not confound the between-group comparison. It does, however, limit what the study can say about students' accuracy in absolute terms: the general overestimation we observed was produced under conditions that rewarded higher self-scores, and it would be unfair to characterize students as inaccurate self-assessors on that basis alone. A design in which the self-assessment carries no grade weight would provide both a cleaner test of the accuracy incentive and a less biased estimate of students' accuracy.

Sixth, the restricted range of instructor-assigned scores (mostly between 7 and 9 on a 0–10 scale) limited the variance available for detecting between-group differences in accuracy. The asymmetry between potential overestimation and underestimation created by this ceiling effect may have further attenuated the study's sensitivity.

Seventh, although the reliability of the main instruments was reassessed in this sample, full re-validation procedures (e.g., confirmatory factor analyses, cognitive interviews) were not conducted. The rubric used to evaluate assignments was developed collaboratively over several years but did not undergo a formal content validation or generalizability analysis, and the use of an Absolute Deviation Index may have reduced the sensitivity to detect directional biases. The measure of perceived accuracy relied on only two items, which may have underrepresented the complexity of students' beliefs. More fundamentally, our operationalization of accuracy as alignment with the instructor's rubric-based score treats that score as the reference standard. Instructor scores are themselves fallible and can vary across raters and occasions (Andrade, 2019). We estimated inter-rater agreement on a subsample and found it to be high (see Rubric), and the mean Absolute Deviation Index was 1.74 points on a 0–10 scale, so the discrepancies analyzed here are unlikely to be driven mainly by instability in the instructor's scores. Agreement between two instructors who share the same rubric and course does not, however, establish that their score is a valid standard, and alignment with external assessments or with the consensus of a larger panel of raters would provide a stronger reference. Future studies could also complement the absolute index with error measures such as the root mean square error, and extend the notion of accuracy to the quality of students' self-feedback and subsequent learning decisions.

Eighth, the moderation analyses, in particular, should be treated as exploratory given the limited statistical power for interaction effects with this sample size.

Finally, the research was conducted within a single undergraduate program in Spain, where the majority of participants were male, reflecting the typical composition of the degree. Consequently, the results may not fully translate to other cultural, institutional, or disciplinary contexts, or to more gender-balanced populations. Replication studies with larger, more diverse samples across multiple settings are necessary to assess the broader applicability of these conclusions.

8. Conclusion

Our study examined whether an accuracy-based incentive, communicated at the point of self-assessment, improves students' summative self-assessment accuracy. We found no clear statistical evidence that it did; Bayesian and equivalence analyses indicated that the data are inconclusive rather than supportive of a null effect, and the design, the timing of the manipulation, the grade weight of the self-assessment, and the absence of process measures all constrain interpretation. Students who reported higher confidence in their self-assessment were no more accurate than their peers, prior academic achievement was the student characteristic most clearly associated with accuracy, and the interactions between self-regulation and condition appeared to be partly associated with pre-existing differences between the classroom groups, limiting their interpretation as effects of the incentive. Self-assessment accuracy was associated with final examination performance, an association that likely reflects shared variance with general academic ability rather than a causal effect.

The contribution of the study is threefold. Methodologically, to our knowledge, it provides the first direct test of an accuracy-based grading incentive for summative self-assessment under controlled conditions, and it shows what a stronger test requires: an incentive embedded early in the course, a clean contrast between incentive and no-incentive conditions, and more classroom groups than the seven available here, so that moderation effects can also be estimated in a multilevel framework. Empirically, it adds two findings that do not depend on the between-group comparison: the near-zero association between perceived and actual accuracy, and the association between prior academic achievement and accuracy in both conditions. It also offers a cautionary example of how a statistically significant moderation effect can arise from differences between intact groups rather than from the manipulation. Conceptually, it makes explicit that alignment with an instructor's score under a transparent rubric combines metacognitive monitoring with social prediction, and that summative self-assessment accuracy should be studied together with the grading conditions that shape it.

CRediT authorship contribution statement

Ernesto Panadero: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization, Data curation. **Juan Fraile:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation. **Anastasios Karakolidis:** Writing – original draft, Validation, Methodology, Formal analysis.

Funding

(1) Spanish National R + D call from the Ministerio de Ciencia, Innovación y Universidades (Generación del conocimiento 2024), Reference number: PID2024–156609NB-I00. (2) Basque Government Call for Grants to support the activities of research groups of the Basque University System (2022–2025) project reference IT1624–22. (3) Universidad Francisco de Vitoria, Call for Research Educational Innovation 2025, Reference Number UFV2025–75.

Declaration of Competing Interest

The authors report there are no competing interests to declare.

Appendix A. Rubric for Assessing/Grading the Course Assignment

		Poor (rating 0–4)	Needs improvement (rating 4–6)	Satisfactory (rating 6–8)	Excellent (rating 8–10)
1: Adequacy of conditions and design of the route 10 points	**Information and value of the site 2.5 points**	The value of the site is not presented.	The cultural, historical, landscape, flora, and fauna significance of the site is poorly described.	The description of the cultural and historical value, landscape, flora and fauna of the site is adequate.	It provides comprehensive information on the site's cultural and historical significance, landscape, flora, and fauna.
	Place 1 point (This section only counts, as shown, in the case of a trip to a mountainous area).				The route is located in a mountainous area. For example: Sierra de Guadarrama National Park.

(*continued on next page*)

(*continued*)

	Route design and guidance 4 points	No information is provided about the route.	Information on the route is really scarce.	The route's distance, elevation profile, slope, and estimated completion time are analyzed.	The distance, elevation profile, IBP index, itinerary with estimated completion times (including movement and stops), water sources, participant monitoring strategies, and technical difficulty are analyzed based on collected data and participant characteristics.
	Map 1 point	No route map is included.	The included map of the route is not topographical.	A topographic map is provided, but it lacks scale and/or coordinates.	A topographic map of the route is provided, including scale and coordinates, with the route clearly marked.
	Activities 1 point	No activities are designed. The hike involves exclusively guiding the participants.		Activities are planned during the hike to enhance participants' knowledge, focusing on technical aspects beyond the site's cultural, historical, and environmental value.	There are activities planned during the outing to improve their knowledge (e.g. something technical, beyond conveying the cultural and historical value, landscape, flora and fauna of the place). In addition, some kind of material is prepared or information is sent before the outing to increase their interest.
	Meteorology 0.5 points	Weather is not taken into account.	The weather forecast for the day of the hike is given.	The weather forecast for the day of the hike is presented and taken into account in order to make and transmit to the participants some previous decisions.	Weather information from previous days and historical weather patterns are considered, along with the hourly forecast for the hike. The decisions made based on this information and their impact on participants are explained and communicated.
2: Security and risk management 3.5 points	**Risk management 2 points**	Risk assessment is not carried out, or is very incomplete.	A risk assessment is done, but some of the obvious risks are missing, or the degree of risks or preventive measures are not specified.	A proper risk assessment is conducted, but some key preventive measures are missing.	A complete prior assessment of the existing risks is made, specifying the degree of risk and the necessary preventive measures in each case.
	Method for hike information (MIDE) 0.5 points				Evaluate the hike and attach the report (screenshot), through the Hike Reporting Method (MIDE): http://mide.montanasegura.com/
	Emergency management 1 point	The emergency management protocol is ineffective in mitigating the consequences of a potential accident.	There is an emergency protocol that is too general, without taking into account the specificities of the activity to be carried out.		Emergency management protocols are designed to reduce the consequences of a potential accident by assessing different accident scenarios on a risk basis.
3: Legal aspects 1.5 points	**Permits and licenses 1.5 points**	The regulations and permits required for the hike are not taken into account.	Some necessary regulations and/or permits are listed, but some key ones are missing.		An inventory of relevant legislation governing the activity is provided, detailing all necessary permits for the hike.
4: Pre-hike information 1.5 points	**Knowledge of participants 1 point**	No prior data is collected on participants.	What prior information is collected on the participants but has not been taken into account in the organization of the hike.	It sets out what prior information is collected on participants and what actions/ decisions, if any, have been taken.	It sets out what prior information is collected about the participants, adapting the organization, setting out the anticipation of possible problems, decisions taken and contacting those persons whose particularities need to be taken into special consideration.
	Accessibility 0.5 points				At least two disabilities and the measures that would be taken to remove barriers for these (hypothetical) participants are outlined.
5: Logistics 3 points	**Equipment carried by the guide and participants 1.5 points**	Some of the essential materials for the hike are missing.		An inventory is made of almost all the material needed, although some important items are missing.	An inventory is made of the necessary equipment, including first-aid kit, to be carried by the guide and the

(*continued on next page*)

(*continued*)

					participants. The units of each material are indicated.
	Environmental impact 0.25 points	The environmental impact assessment of the hike is not taken into account.		Environmental impact prevention measures are in place, but some key ones are missing.	Key environmental impact prevention measures are outlined, along with information for participants on how these measures affect aspects such as appropriate clothing and gear.
	Transport 0.5 points	Means of transport are not taken into account.	Parking places for motor vehicles are taken into account.	Parking places for motor vehicles and other means of public transport are taken into account.	Parking availability for motor vehicles and public transport options are considered, taking into account specific conditions on the day and time of departure. Participants are informed accordingly.
	Evaluation 0.75 points	No assessment of the route planning is conducted.	There is little evaluation report: only a few errors or successes are identified.	An evaluation report is produced, highlighting successes, errors, and recommendations for improvement.	A full evaluation report is made of all sections and phases (including accidents and incidents) and taking into account the opinions collected from the participants.
6: Bibliographical references 2 points	**2 points**	Bibliographical references are not cited or those that do appear are scarce, at the end of the document and/or with multiple errors.	Only a bibliographical references section appears at the end of the dossier, in APA 7th edition style. It is cited correctly.	With some errors, references are cited throughout the text and listed at the end of the dossier in an APA 7th edition bibliography.	Correctly cited throughout the text and also listed at the end of the dossier in a bibliographical references section. APA 7th edition style.

Total: 21.5 possible points = would be a score of 10.

References

Alonso-Tapia, J., Panadero, E., & Ruiz, M. A. (2014). Development and validity of the emotion and motivation self-regulation questionnaire (EMSR-Q). *The Spanish Journal of Psychology, 17*(e55), 1–15. https://doi.org/10.1017/sjp.2014.41

Andrade, H. L. (2018). Feedback in the context of self-assessment. In A. A. Lipnevich, & J. K. Smith (Eds.), *The Cambridge Handbook of Instructional Feedback* (pp. 376–408). Cambridge University Press. https://doi.org/10.1017/9781316832134.019.

Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education, 4*, 87. https://doi.org/10.3389/feduc.2019.00087

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares, & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Information Age Publishing.

Brown, G. T. L., & Harris, L. R. (2013). Student Self-Assessment. In J. H. McMillan (Ed.), *The SAGE Handbook of Research on Classroom Assessment* (pp. 367–393). SAGE Publications, Inc.. https://doi.org/10.4135/9781452218649.n21.

Dochy, F. J. R. C., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24*(3), 331–350. https://doi.org/10.1080/03075079912331379935

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69–106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: a meta-analysis. *Review of Educational Research, 59*(4), 395–430. https://doi.org/10.3102/00346543059004395

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

León, S. P., Augusto-Landa, J. M., & García-Martínez, I. (2021). Moderating factors in university students' self-evaluation for sustainability. *Sustainability, 13*(8), 4199. https://doi.org/10.3390/su13084199

León, S. P., Panadero, E., & García-Martínez, I. (2023). How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy. *Educational Psychology Review, 35*(4), 106. https://doi.org/10.1007/s10648-023-09819-0

León, S. P., Vallejo, A. P., & Nelson, J. B. (2021). Variability in the accuracy of self-assessments among low, moderate, and high performing students in university education. *Practical Assessment, Research & Evaluation, 26*(1), 16. https://doi.org/10.7275/6q91-az58

Panadero, E., & Alonso-Tapia, J. (2025). *Manual for the use of the Emotion and Motivation Self-Regulation Questionnaire (EMSR-Q)* (Version January 2025) https://www.ernestopanadero.es/Publications/Manuals/EMSR_Q_Manual_Emotion_motivation_self_regulation_questionnaire.pdf.

Panadero, E., Alonso-Tapia, J., García-Pérez, D., Fraile, J., Galán, J. M. S., & Pardo, R. (2021). Deep learning self-regulation strategies: Validation of a situational model and its questionnaire. *Revista de Psicodidáctica (English ed.), 26*(1), 10–19. https://doi.org/10.1016/j.psicoe.2020.11.003

Panadero, E., Brown, G. T. L., & Strijbos, J. W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review, 28*, 803–830. https://doi.org/10.1007/s10648-015-9350-2

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004

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

Pinedo, L., Panadero, E., Fernández-Ruiz, J., García-Pérez, D., & Brown, G. T. L. (2026). The affective cost of miscalibration: Effects of discovering self-assessment accuracy on emotions and self-efficacy. *European Journal of Psychology of Education, 41*. https://doi.org/10.1007/s10212-026-01087-0. article 67.

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment Research & Evaluation, 11*(1), 10. https://doi.org/10.7275/9wph-vv65

Rouy, M., de Gardelle, V., Reyes, G., Sackur, J., Vergnaud, J. C., et al. (2022). Metacognitive improvement: Disentangling adaptive training from experimental confounds. *Journal of Experimental Psychology: General, 151(9)*, 2083–2091. https://doi.org/10.1037/xge0001185

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: a meta-analysis. *Journal of Educational Psychology, 109*(8), 1049-1066. https://doi.org/10.1037/edu0000190

Schunk, D. H., & Pajares, F. (2009). Self-efficacy theory. In K. R. Wentzel, & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 35–53). Routledge.

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education, 9*(2), 169–191. https://doi.org/10.5465/amle.9.2.zqr169

Yan, Z., & Brown, G. T. L. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education, 42*(8), 1247–1262. https://doi.org/10.1080/02602938.2016.1260091

Yan, Z., Lao, H., Panadero, E., Fernández-Castilla, B., Yang, L., & Yang, M. (2022). Effects of self-assessment and peer-assessment interventions on academic performance: A meta-analysis. *Educational Research Review, 37*, Article 100484. https://doi.org/10.1016/j.edurev.2022.100484

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metacognition in Education* (pp. 299–315). Routledge.